The effects of syntactic and acoustic cues on the perception of prosodic boundaries

Jianjing Kuang1, May Pik Yu Chan2, Nari Rhee3

1 2 3 Department of Linguistics, University of Pennsylvania
kuangj@ling.upenn.edu, pikyu@sas.upenn.edu, nrhee@sas.upenn.edu

Abstract
This study investigates how the perception of prosodic boundaries is shaped by syntactic phrasing and acoustic cues for English and Mandarin listeners. Syntactically-parsed speech corpora were used as the stimuli for the perception experiment. The relative strength of the syntactic boundary of both the left and right sides of the constituents was extracted from the syntactic parsing annotations. A wide range of acoustic cues of both prosodic domain-final and domain-initial positions were examined. Linear-mixed-effects modeling of the likelihood of boundary perception suggests that, for both languages, prosodic boundary perception was influenced by both the strength of syntactic boundary and acoustic cues: boundary perception was heavily driven by the presence of pause; pause also modulated the contribution of other acoustic cues; and larger syntactic boundaries were generally more likely to be perceived as prosodic boundaries. However, there is also cross-linguistic variation: the effect of syntactic phrasing cues was generally stronger for English; acoustically, the effect of final lengthening and pitch reset was stronger in English, while pause was the dominant cue in Mandarin. We discuss the important implications of these findings related to the nature of prosodic hierarchy, and the nature of the prosody-syntax interface.

Index Terms: Prosodic boundaries, Boundary perception, Prosody-Syntax interface

1. Introduction
Prosodic phrasing is an important part of language processing. The perception of prosodic boundaries is conditioned by multiple factors. On the one hand, a wide range of acoustic cues contributes to the production and perception of prosodic boundaries. Domain-finally, phrase-final lengthening and pause [1, 2, 3, 4] have been widely reported to be salient cues for larger prosodic domains; domain-final boundaries are also indicated by pitch scaling cues such as final lowering [5, 6, 7, 8] as well as voice quality cues [3, 9, 10, 11]. Domain-initially, the prosodic boundaries are marked by pitch reset and other domain-initial strengthening effects on duration and voice quality [12, 13].

On the other hand, prosodic phrasing is also directly related to other linguistic processing, in particular, syntactic parsing. A number of studies have shown that listeners use prosodic boundaries to locate syntactic boundaries [14, 15, 16], and to resolve syntactic ambiguities [17, 18]. The influence of syntax on prosodic perception is less well-studied, but [19, 20] found that syntactic cues significantly influence the detection of prosodic boundaries. However, studies are still sparse in terms of how acoustic cues and syntactic cues together shape the perception of prosodic phrasing.

Moreover, even though a close link between syntactic boundaries and prosodic boundaries has been recognized in prosodic theories [21, 22, 23, 24, 25], the specific mapping mechanism between prosodic boundaries and syntactic boundaries remains to be an unsettled question. One particular challenge is that although large syntactic boundaries generally tend to map onto prosodic boundaries, mismatch between syntactic and prosodic boundaries is fairly frequent. Here we would like to discuss two major reasons for this challenge. First, this issue raises from the implicit assumption of the categoricity of prosodic and syntactic domains. Prosodic hierarchies (e.g. ip or IP) have been defined as distinct categories [6, 1], however, many acoustic cues (e.g., duration) of the different levels of prosodic boundaries are rather gradient in nature. Similarly, although syntactic phrasal structures such as XP are defined categorically, due to the recursiveness of syntactic structures, an XP can locate at a higher level of a syntactic tree or a lower level of a tree. Because of the lack of strict categorical hierarchies for both syntactic and prosodic domains, syntax and prosody are not likely to have a one-to-one mapping. Indeed, as noted by [26] and [27], almost any syntactic edge is a potential location for a prosodic phrase boundary. Second, the syntax-prosody mapping is likely to vary from language to language. At the syntax level, [23, 28] proposed that languages differ parametrically as to whether the left or right edges (or both) of syntactic constituents are aligned with the prosodic domains [23, 28]. For example, Japanese aligns the left boundary, while English aligns to the right. Empirically, [19] indeed found that English listeners are more consistent in detecting the right boundaries. At the acoustic level, it is also well-known that the acoustic encoding of prosodic boundaries is highly variable across languages. Finally, although it has not been tested empirically, it is possible that languages vary in the relative importance of syntactic cues and acoustic cues.

To provide more empirical insights on the prosody-syntax mapping, this study investigates the effects of syntactic structure and acoustic cues on the perception of prosodic boundaries by English and Mandarin listeners. Specifically, we test: 1) whether the relative strength of the syntactic boundaries has effects on the likelihood of perceived boundaries, and whether left and right edges of the syntactic constituents have the same effects. 2) Between syntactic and acoustic cues, which effects play more important roles in the perception of prosodic boundaries? 3) Among the various prosodic cues (both domain-initial and domain-final cues), which acoustic cues carry the most weight in the perception responses? 4) Finally, is there any cross-linguistic variation between English and Mandarin listeners?

This study makes several innovative methodological contributions: Syntactically-parsed spontaneous speech corpora were used as the stimuli for the perception experiment. For the reasons we discussed above, we did not assume strict categorical hierarchical boundaries in the analysis. To capture the relative-ness and gradience of syntactic boundaries, we used the number of brackets at each edge of syntactic constituents as the proxy
of the relative strength of the syntactic boundaries. Meanwhile, both domain-initial and domain-final acoustic cues were extensively examined.

2. Methods

Boundary rating experiments were conducted with native speakers of English and Mandarin respectively, to examine the role of the strength of syntactic boundaries and acoustics on the perception of boundaries. The same experimental procedure and analysis method was used to analyze responses from the two languages.

2.1. The corpora

The English corpus used in this study was a female native English speaker’s recording of reading Jane Austen’s Emma obtained from LibriVox [29], a free public domain of audiobooks. Volume II Chapter 10 of this book was chosen for analysis because the syntactic parsing is available in the 2nd edition of The Penn Parsed Corpus of Modern British English (PPCMBE2) [30]. The recording had a sampling frequency of 44.1kHz and sample depth of 32-bit. The Mandarin corpus used in this study was the Chinese Tree Bank 9.0 [31], which includes segmented, annotated and parsed news article texts. Passages were annotated with Penn Treebank-style labeled brackets. The speakers of the corpus all native Mandarin speakers who achieved Class 2 Level 1 or better on the Putonghua Shuiping Ceshi (the national standard Mandarin proficiency test). Recordings had a sample frequency of 44.1kHz and sample depth of 16-bit. In this study, we handpicked readings of news articles by 1 male and 1 female speaker. One passage per speaker was chosen to be included in the study. Across the two corpora, we chose these three talkers because their recordings had the least disfluencies, they had regular pacing throughout their reading, and they were expressive.

2.2. Participants and materials

47 native speakers of English (18-25 years; 29 female) and 29 native speakers of Mandarin (18-35 years; 18 female) were recruited from university student communities. Participants listened to sentences (English: 24; Mandarin: 22) of around 30 seconds or less each in their native languages, which were downloaded from Volume II Chapter 10 of Emma by Jane Austen on LibriVox and the Chinese Treebank 9.0 respectively. Participants identified prosodic boundaries during the listening task. Across the sentences, there were a total of 382 and 458 potential boundaries in English and Mandarin for participants to rate respectively.

2.3. Boundary Detection Task

The experiments were conducted using the online survey software Qualtrics. Boundary ratings were obtained using a boundary detection task, which was similar to the Rapid Prosody Transcription task used in [19, 32], and the boundary detection task in [20]. Selected sentences were presented auditorily one at a time, with a transcript of the sentence simultaneously displayed. The audio file played automatically once the participant entered the question, and participants were asked to select where they “think there are boundaries” within the sentence. It was not possible for participants to select either the beginning or the end of the sentence; they were allowed to replay the audio as many times as they like. Three sample questions involving the same sentence read with different prosodic focus and structures were first presented before the participants moved on to the main task. This ensured that the participants understood how to use the experimental interface. No further definition of ‘boundaries’ was given; the participants were allowed to interpret the instructions however they wish. The rate of boundary perception was computed by dividing the percentage of boundary response by the maximum number of boundary responses in a given language.

2.4. The strength of syntactic boundary

Syntactic parsings for the selected sentences in the two languages were available in the Chinese Treebank 9.0 [31] and Penn Parsed Corpus of Modern British English (PPCMBE2) [30], annotated following the Penn Treebank guidelines. The numbers of left and right edges of the constituent structures between two consecutive words were used as proxies for the depth of the syntactic structure and the additive strength of syntactic boundaries. By the design and implementation of the annotations, every word had at least one left bracket before the word, and a right bracket after. To normalize for the various sentence lengths, the numbers of left and right brackets were divided by the maximum number of left and right brackets in the sentence.

2.5. Acoustic measures of prosodic boundaries

Acoustic measures of the selected sentences were taken using Praat. The spoken corpora of Chinese Treebank 9.0 was aligned using the HMM-based Mandarin forced-aligner [33] and sentences from Librivox were forced aligned using the Penn Phonetics Lab Forced Aligner [34]. Acoustic measures taken involved the following: (i) The presence or absence of pause at the boundary junction, as defined by whether a silent portion was identified by the forced aligner; and (ii) the average syllable duration of the pre-boundary and post-boundary word. The following measurements from the pre-boundary and post-boundary word were also taken by averaging the values of the sonorous segments of the word: (iii) minimum and maximum fundamental frequency (F0) in Hz, (iv) minimum and maximum sound pressure level (SPL) – intensity measured in energy, (iv) spectral tilt as measured by the alpha ratio - the level difference between the 1-5kHz region and that of the 50-1kHz region, (v) spectral tilt as measured by L1-L0 - the level difference between the F1 region (defined as between 300 Hz and 800 Hz) and the F0 region (defined as between 0 Hz and 300 Hz), and (vi) Cepstral Peak Prominence-Smoothed (CPPS). The effectiveness of (iv-vi) as measures of voice quality are described in [35]. For (i), a binary grouping of with and without pause was used instead of a continuous variable, because the thresholds for pauses were arbitrarily defined by the forced aligners, and the effect of pause was unlikely to be linear, as most words do not have pauses in between them (86% for English, 79% for Mandarin). Apart from (i), the acoustic measures were first z-scored by speaker. To assess the effect of the acoustic variables on listeners’ boundary perception, we computed a difference in acoustic measure between the post-boundary word and the pre-boundary word, apart from average syllable duration. For measures (iii) F0 and (iv) SPL, we calculated the maximum value of the post-boundary word minus the minimum value of the pre-boundary word. This allowed us to capture effects of phrase initial strengthening, such as pitch reset. We also calculated the difference between the pre-boundary and post-boundary word for measures (iv-vi) by deducting the acoustic measurement value of the pre-boundary word from the post-boundary word.
3. Results and Discussions

3.1. Statistical model

The proportion of participants that identified a given word junction as a boundary (percentage of boundary response divided by the maximum number of boundary responses in a given language) was evaluated using a linear mixed effects model in R in the lme4 package [36]. p-values were computed via a type III ANOVA with the Satterthwaite’s degrees of freedom method using the package lmerTest (lmerTest::anova, [36]). The dependent variable was boundary rating (numeric). Main effects included the pre-boundary word’s average syllable duration (numeric), the post-boundary word’s average syllable duration (numeric), FO difference (numeric: maximum of the post-boundary word – minimum of the pre-boundary word), SPL difference (numeric: maximum of the post-boundary word – minimum of the pre-boundary word), alpha difference (numeric), L1-L0 difference (numeric), CPPS difference (numeric), proportion of left brackets (numeric), proportion of right brackets (numeric), as well as pause (0 or 1), and language (2 levels: English [reference level], Mandarin). Two-way interactions between pause/language and the remaining acoustic and syntactic measures, as well as three-way interactions between pause, language and the remaining acoustic measures were also included. Random intercepts for talker were also included.

3.2. Presence of pause

A significant main effect of pause (binary) was observed ($F(1, 799.60) = 19.63, p < 0.001$). As there were many predictors with significant interactions with pauses, figures (Figure 1 to Figure 4) are presented with the line of best fit with and without pause for each language group, allowing us to better compare the effect of pause and its interactions with other predictors in the two languages. The x-axis shows the value of the given variable, which would be a z-scored difference value for acoustic measurements, or a percentage of syntactic brackets scaled to the maximum number of brackets in a given sentence. The y-axis shows the proportion of boundary responses in a given language. There was a significant interaction effect of language and pause ($F(1, 799.60) = 19.24, p < 0.001$), suggesting that the Mandarin listeners were especially consistent in boundary detection when there was a pause.

3.3. Syntactic boundaries

Overall, results showed that both English and Mandarin listeners were sensitive to syntax, and significant main effects were found for both left ($F(1, 799.15) = 14.36, p < 0.001$) and right ($F(1, 746.41) = 148.90, p < 0.001$) brackets, but the effect of right brackets was stronger. Moreover, there were significant two-way interactions between language and left ($F(1, 799.15) = 12.30, p < 0.001$) and right ($F(1, 746.41) = 33.20, p < 0.001$) brackets, as well as between pause and right brackets ($F(1, 799.27) = 19.41, p < 0.001$). Lastly, significant three-way interaction effects were also found between pause, left bracket and language ($F(1, 799.28) = 6.51, p < 0.05$). As illustrated in Fig.1, these interactions suggest that English and Mandarin listeners weighed syntax cues differently in boundary perception.

In the case of English, the effect of right brackets on boundary perception appeared to be greater than the effect of left brackets. In comparison, boundary perception by Mandarin listeners showed more symmetric and weaker effects of left and right brackets. Moreover, for English listeners, syntactic cues had independent effects on boundary perception, whereas for Mandarin listeners, the effect of syntactic cues largely co-varied with the presence of pause.

3.4. Average Syllable Duration

Figure 2 shows the effect of the pre-boundary (Prev.AvgSylDur) and post-boundary (Next.AvgSylDur) average syllable duration on boundary perception.

Modeling results found a significant main effect of the pre-boundary average syllable duration ($F(1, 799.08) = 7.72, p < 0.01$) and the post-boundary average syllable duration ($F(1, 799.76) = 21.51, p < 0.001$), as well as significant two-way interaction effects between pause and the post-boundary average syllable duration ($F(1, 799.61) = 6.85, p < 0.01$), and between the pre-boundary average syllable duration and language ($F(1, 799.08) = 32.81, p < 0.001$). A significant three-way interaction effect of pause, the pre-boundary average syllable duration and language was also found ($F(1, 799.00) = 22.81, p < 0.001$).

Overall, duration was found to have an effect on the rate of perceived boundary in both languages, but the effects were further modulated by pause and language. Specifically, when there was no pause, the likelihood of perceived boundary was positively correlated with pre-boundary syllable duration (i.e. domain-final lengthening effect) for both languages; however,
when there was a pause, the correlation between the likelihood of perceived boundary and domain-final duration was positive for English listeners, but negative for Mandarin. This suggests that for Mandarin listeners, when pause is present, listeners no longer expect domain-final lengthening. For both languages, pre-boundary (domain-final) duration effects were stronger than post-boundary (domain-initial) duration effects. The main effect of domain-initial duration was largely driven by the presence of pause.

3.5. Reset effect of pitch, intensity and voice quality

Prosodic boundaries also often reflect pitch scaling, intensity and voice quality cues. To capture the reset effect, we calculated the maximum difference between post-boundary and pre-boundary syllables. Greater positive values indicate a greater domain-initial reset effect. The lmer model suggests that there was a significant main effect of pitch reset on the boundary detection rate ($F(1, 799.68) = 49.58, p < 0.001$) and a significant two-way interaction of pause and F0 ($F(1, 798.47) = 18.60, p < 0.001$). Furthermore, significant three-way interaction effects for pause, language and both F0 ($F(1, 799.47) = 7.02, p < 0.01$) and SPL ($F(1, 781.91) = 6.24, p < 0.05$) were found.

As illustrated in Figure 3, a general pitch reset effect on boundary perception was found for both languages. But again, this effect was modulated by pause and language. In both languages, pitch reset effect was stronger when there was a pause, but this interaction was particularly strong for English listeners, where there was a strong correlation between the size of pitch reset and the boundary detection rate. By contrast, for Mandarin listeners, the effect size of pitch reset was largely conditioned by the presence of pause; and the pitch reset effect was slightly more salient when there was no pause. Moreover, the effect of SPL difference may generally be conditioned by pause. In Mandarin, the effect of SPL generally co-varied with F0, while SPL and F0 were slightly more independent in English.

Figure 4 illustrates the effect of voice quality difference across the boundaries. A voice quality reset effect was found for both language groups. As suggested by the significant main effect of CPPs ($F(1, 799.02) = 8.13, p < 0.01$), listeners were more likely to perceive a boundary when there was a greater periodicity difference between pre-boundary and post-boundary syllables. This finding is consistent with the phrase-final creak and domain-initial strengthening effect of voice quality. The significant two-way interaction between CPPs and pause ($F(1, 799.08) = 9.65, p < 0.01$) further suggests that the effect of voice quality was more salient when there was a pause. As for the spectral slope cues, there was a significant three-way interaction between pause, language and L1-L0 ($F(1, 799.13) = 4.50, p < 0.05$). Consistent with CPPs, greater L1-L0 difference was a useful cue for boundary perception.

4. Conclusions

This study investigates how the perception of prosodic boundaries is shaped by syntactic phrasing and acoustic cues for English and Mandarin listeners. Overall, we found that the boundary detection rate was influenced by a wide range of syntactic and acoustic cues. The presence of pause was a particularly salient cue for boundary detection, and it also modulated the contribution of other acoustic and syntactic cues. Moreover, syntactic and acoustic cues appeared to be weighed differently for English and Mandarin listeners. In general, syntactic phrasing had an independent effect on boundary perception for English listeners, while it mostly co-varied with pause for Mandarin listeners. English listeners were more affected by the right-edge cues, while Mandarin listeners had less bias between left and right edges. Furthermore, both domain-initial and domain-final cues played significant roles in boundary detection, and the effects of acoustic cues were often modulated by pause and language. For example, final lengthening cues were more reliable for English listeners when there was a pause, but more reliable for Mandarin listeners when there was no pause. Taken together, our study effectively demonstrated that syntax-prosody mapping is quite gradient (with some cues being more categorical), and the mapping varies across languages.

5. Acknowledgements

We would like to thank Mark Liberman and Hongwei Ding for generously sharing the speech corpora with us. We also thank Xin Gao for her help and input in earlier stages of the project. We are also grateful for feedback from Sunghye Cho, Aini Li, Hassan Munshi and other members of the Penn phonetics lab.
6. References


