Temporal Localization of Syntactically Conditioned Prosodic Information

Seung-Eun Kim, Sam Tilsen

Department of Linguistics, Cornell University, USA
sk2996@cornell.edu, tilsen@cornell.edu

Abstract

Many prosodic theories hold that different syntactic structures are mapped to distinct prosodic organizations; these theories predict that acoustic and articulatory correlates of these structures differ mainly at phrase boundaries, yet no studies have investigated whether such predictions are correct. This study uses a novel neural network-based analysis method for temporally localizing prosodic information that is associated with syntactic contrast in acoustic and articulatory signals. Specifically, we focus on the contrast between non-restrictive and restrictive relative clauses. Neural networks were trained on multi-dimensional acoustic and articulatory data to classify the two types of relative clauses, and the network accuracies on test data were analyzed. The results found two different patterns: syntactically conditioned prosodic information was either widely distributed around the boundaries or narrowly distributed at specific locations. The findings suggest that prosodic expression of syntactic contrasts does not occur in the uniform way or at a fixed location, but rather it is accomplished with various strategies.

Keywords: temporal localization, neural networks, multi-dimensional data, syntax-prosody mapping

1. Introduction

Many theories of the syntax-prosody interface argue that different syntactic structures are associated with distinct prosodic organizations. Specifically, the left and right edges of syntactic constituents are mapped to varying strengths of prosodic boundaries based on sentence structure. Under these theories, acoustic and articulatory correlates of syntactic structures are predicted to be observed mainly at phrase boundaries, as phonetic measures at phrase-final or initial position would reflect different prosodic boundary strengths. Previous studies have taken this prediction for granted and examined phrase edges to find acoustic and articulatory differences between a contrasting set of syntactic structures (e.g. Cooper and Sorensen 1977; Garro and Parker 1982; Kim and Tilsen 2020). However, it is important to assess the empirical evidence for such approaches; that is, we need to identify when in time the syntactically conditioned prosodic information exists rather than simply presupposing that it exists at phrase edges.

The question of how to temporally localize syntactically conditioned prosodic information has not been thoroughly addressed in the literature. Previous studies have mostly examined fixed regions in the vicinity of phrase boundaries – for example a few segments or syllables at phrase-final or initial position – and have targeted a handful of specific types of measurements that are believed to be relevant (e.g. F0 values, segmental durations, etc.). However, it is unclear exactly how far from a boundary we might identify relevant prosodic information, and it is usually unknown whether the measurements used are the most appropriate ones.

In this context, the current study explores a novel method for systematically detecting prosodic information that is associated with a syntactic contrast. Specifically, we conduct analyses in which a neural network is trained to classify syntactic categories from multi-dimensional articulatory and acoustic input data. Then, network accuracy is assessed on unseen test data. This is repeated with multiple times with randomly sampled training and test sets, for a given analysis window. To temporally localize prosodic information, we systematically vary the sizes and locations of analysis windows. This method is based on the analysis procedure presented in Tilsen (2020).

One of the important features of the novel methodology is that we use multi-dimensional data in the analyses. Typically, studies measure relevant acoustic and articulatory variables and run statistical tests to determine whether there is a significant difference. This method, however, can be problematic for several reasons. First, it is possible that an interaction between the selected measurements, not the measurements per se, better reflects a syntactic difference. Moreover, these interactions may be nonlinear, so including interaction terms in a linear model will not solve the problem. Second, there may be information in acoustic and articulatory signals that we are not cognizant of, but which in fact is relevant to a syntactic difference. For example, when investigating articulatory measures associated with a syntactic contrast, researchers typically examine measures of movement timing or amplitude derived from the horizontal and vertical coordinates of the tract variables associated with constriction gestures (e.g. lip aperture, tongue tip constriction degree, etc.); yet, it is conceivable that the position of jaw can provide crucial information on a syntactic difference. Conventional analyses tend to reduce the high-dimensional acoustic and articulatory signals that we observe to a handful of variables. Our analysis method avoids this reduction by using the complex, high dimensional acoustic and articulatory signals of speech directly.

The syntactic contrast that we focus on is between a non-restrictive relative clause (NRRC) and a restrictive relative clause (RRC), examples of which are shown in (1). NRRCs and RRCs are syntactically and semantically distinct (see Arnold 2007, for an overview). In example (1), the NRRC does not contribute to identifying the referent (Mr. Hodd), but it simply gives extra information about the referent (i.e. similar to a parenthetical). However, the RRC in (1) is essential to identifying the referent from a set of possible referents. NRRCs are often separated from the main clause by commas, but RRCs are not.

Following these syntactic differences, researchers have argued that the two types of RCs differ in their prosodic structures. For example, Selkirk (2005) represents the structural differences as in (1), where the NRRC constitutes an intermediate phrase (ip) on its own, while the RRC constitutes an ip

together with the main clause subject. On the other hand, Nespor and Vogel (1986) argued that there is a mandatory intonational phrase (IP) boundary before and after an NRRC, but not in RRC. Although theories differ on specific predictions on their prosodic organizations, they all predict that the two types of RCs will differ in the vicinity of the phrase boundaries, before and after the relative clause. The pre- and post-relative clause boundaries will be referred to as B1 and B2 respectively in the current study.

(1) *Non-restrictive relative clause* (NRRC)
Context: There is one Mr. Hodd. He knows Mr. Robb.
[[A Mr. Hodd,]$_{ip}$ [who knows Mr. Robb,]$_{ip}$]$_{IP}$
[[often plays tennis.]$_{ip}$]$_{IP}$

Restrictive relative clause (RRC)
Context: There are two Mr. Hodds. Only one knows Mr. Robb.
[[The Mr. Hodd who knows Mr. Robb]$_{ip}$
[often plays tennis.]$_{ip}$]$_{IP}$

The temporal localization method allows us to address several different questions regarding syntactically conditioned prosodic information. First, we can test whether the predictions from the theories of syntax-prosody mapping are correct such that different syntactic structures have distinct prosodic organizations and thus mainly differ at phrase boundaries. If this boundary-locality is correct, we expect to observe high network accuracy in analysis windows which are near phrase boundaries. Second, we can examine whether syntactically conditioned prosodic information is distributed similarly before and after RCs, as well as examine whether there are across-participant differences. Such variation will inform our theories of the syntax-prosody mapping.

2. Methods

2.1. Participants and task

Six native speakers of English (3M, 3F) participated in the experiment. Participants were seated facing a computer monitor in a quiet room. In each trial, participants first saw a context sentence and then a target sentence, as in example (1). The context sentence was provided to draw attention to the syntactic contrast between NRRC and RRC. Definite/indefinite determiners in the beginning of the target sentence also facilitated the relevant interpretations of the target sentences. Participants were instructed to read both sentences silently when they first appeared. After 1.5 seconds, a moving rate cue appeared. This was a red box that moved from left to right across the screen at a constant speed; the period of time it took for the rate cue to move across the screen varied in ten steps. When the cue stopped moving, participants were instructed to read the target sentence in a way that reflected the speed of the rate cue. Crucially, they were instructed to vary their speech rate to correspond with their impression of how fast or slow the cue moved. The motion-based rate cue allowed us to elicit a continuous variation of speech rate of the two types of RCs. There were six blocks of 40 trials in each experimental session. Participants produced one type of RC throughout a block, and the blocks alternated between the two types of RCs.

The target words in the experiment were the names that followed "Mr." The names started in /h/, /r/, or /l/ and ended in /b/ or /d/. All the names had the same vowel /a/. Participants were instructed not to put emphasis on any of the words in the sentence, particularly the target words.

2.2. Data collection and processing

Articulatory data were collected with an NDI Wave Electromagnetic Articulograph (EMA) with a sampling rate of 400 Hz. Articulator sensors were located mid-sagitally on the upper lip (UL), lower lip (LL), gum below the lower incisors (JAW), tongue tip (TT, approximately 1cm from the apex of the tongue), and tongue body (TB, approximately 4-5 cm posterior from the TT). Reference sensors were located on the nasion and left and right mastoid processes and were used to correct for head movement. The reference and articulator sensors were filtered at 5 and 10 Hz respectively using low-pass Butterworth filters.

Acoustic data were collected at a sampling rate of 22050 Hz. In order to locate prosodic boundaries in the acoustic and articulatory signals, acoustic segmentations were conducted. For each participant, six trials were manually labelled and used to train HMMs in the Kaldi speech recognition toolkit. A forced alignment was conducted for the remaining trials. The alignments of all trials were manually inspected and corrected when necessary. A total of 240 trials were collected for each of the six participants. Out of 1440 trials, 127 trials (8.8%) that had speech errors, disfluencies, or problems in data collection were excluded from analyses.

2.3. Data analysis

Inputs to neural network analyses were composed of 86 dimensions: 20 articulatory dimensions and 66 acoustic dimensions. Articulatory dimensions were the horizontal and vertical positions of the five articulator sensors (UL, LL, JAW, TT, TB) and each of their velocities. Acoustic dimensions were 33-dimensional broadband spectrogram and their first differences. Figure 1 shows an example of the analyses input where each dimension is represented as a horizontal line.

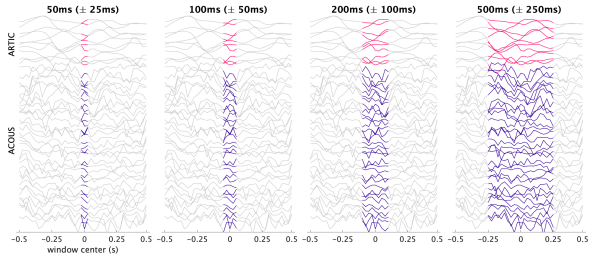

Figure 1: *An example of differently sized analysis windows. Each horizontal line represents articulatory (horizontal/vertical coordinates of five articulator sensors) or acoustic (33-dimensional broadband spectrogram) information. The first differences of the articulatory and acoustic signals were also included as an input (not shown). The information that was used for different analysis windows is represented with colored lines in each panel. In these examples, all the inputs were aligned to the end of the target name (window center: 0s) but varied in size (the title of each panel).*

For analyses at each boundary, the signals across trials were aligned to the end of the pre-boundary segment (i.e. end of the target word), which was determined from the forced alignment. This alignment point is time 0s in the examples in Figure 1. The centers and sizes of the analysis windows were then systematically varied. Window centers were defined in 25 ms steps relative to the boundary, up to ±500ms; thus, for each B1 and

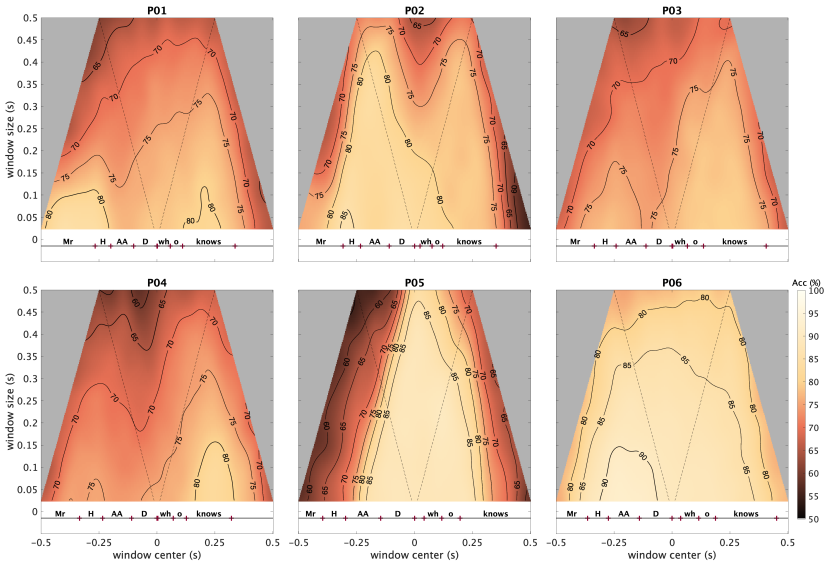

Figure 2: *Classification results at B1 (the pre-RC boundary) shown in heatmaps along with the mean segment/word durations. The x-axis shows the location of window centers, and the y-axis shows the window size. The colors and numbers represent the network classification accuracy (%). The diagonal lines indicate the left and right edges of the analysis windows at the alignment point. Gray areas are windows which are not analyzed because they would contain information outside of the time periods that we examine.*

B2, a total of 41 window centers were investigated. The size of the analysis window varied from 25ms to 500ms in a 25ms step. The analysis windows extended to both sides of the center such that the 50ms window contained half of the information (25ms) on the left side of the center and the other half (25ms) on the right side of the center (see Figure 1). At each window center, only the windows that contain information within ±500ms of the alignment point were investigated. Therefore, at center 0s (the alignment point), all 20 windows were investigated (i.e. 25:25:500ms), whereas at center 0.4s, only eight windows were investigated (i.e. 25:25:200ms). Before training/testing the networks, data in each analysis window were normalized to zero mean and unit variance by dimension within each participant.

All analyses were conducted within participant, because between-participant differences in vocal tract structure or prosodic behavior are likely to make it more difficult for the networks to learn to classify the two types of RCs. Twenty repetitions of the training-testing procedure were conducted for each analysis window. In each repetition, half of the trials were randomly assigned to a training set, and the other half were assigned as a test set. The neural network architecture we used had two bidirectional LSTM (biLSTM) layers, with 40% dropout after each layer. We analyzed the mean network accuracy on the unseen test data; thus, accuracy can be interpreted to reflect the ability of the network to learn generalizable mappings from inputs to syntactic categories. We balanced type of RC, coda of the target word, and speech rate in generating the train and test sets; for speech rate, ten different rates were divided into five rate categories. Note that the network architecture and training parameters we used were based on those in Tilsen (2020), and it is important to keep in mind that these are not necessarily optimal; therefore, the classification accuracies we obtain can only be used to infer lower bounds on the temporal extent of syntactically relevant information.

3. Results

The results showed two different distributional patterns of syntactically conditioned prosodic information: the information was either widely distributed around the boundaries or more narrowly distributed, being concentrated at specific locations. Figure 2 shows heatmaps of the classification results at B1 (the pre-RC boundary). The widely distributed pattern was observed for Participants 1, 3, and 6: classification accuracy was relatively high throughout the pre- and post-boundary regions. Notice that at the critical regions, the networks showed high accuracy even at very small window sizes. This suggests that there was sufficient amount of information that distinguishes the two types of RCs in just one (25ms window size) or two frames (50ms window size) of the data. On the other hand, the narrowly distributed pattern was observed for Participants 2, 4, and 5: high classification accuracy was found only at certain window centers. For those who showed the concentrated pattern, the region that showed the highest accuracy differed across participants. While the highest accuracy was found in the pre-boundary region in Participant 2, it was found in the post-boundary region in Participant 4 or at the immediate region around the boundary in Participant 5.

Both widely distributed and narrowly distributed patterns were also observed at B2, the post-RC boundary (see Figure 3). At B2, Participants 1, 2, and 5 showed the narrowly distributed pattern, while Participants 3, 4, and 6 showed the widely distributed pattern. Although the region that showed high accuracy was relatively small in Participants 3 and 4, as the highest accuracy was found in both pre- and post-boundary regions, we identified them to exhibit the widely distributed pattern. As in B1, participants that showed narrower distribution differed on where they locate critical information.

Comparing the results from B1 and B2, we found that not all participants used the same strategy of marking the syntactic contrast across the two boundaries. While a majority of the participants (i.e. Participants 2, 3, 5, and 6) showed the same

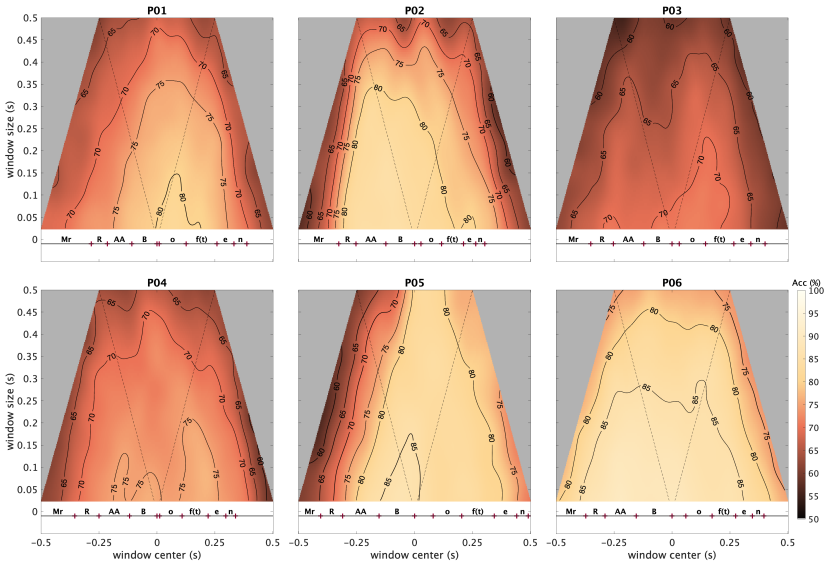

Figure 3: *Heatmaps of the classification results at B2 (the post-RC boundary). At this boundary, all participants paused after the target word, which is marked as the blank interval between "B" and "o".*

pattern, Participants 1 and 4 showed different patterns at B1 and B2. For instance, Participant 1 showed the widely distributed pattern at B1, but the information was more narrowly distributed at B2. Additionally, we could not find any common pattern across participants within B1 and B2; both wide and narrow distributions were observed at both boundaries.

4. Discussion and conclusion

In sum, this study investigated where in time relative to phrase boundaries speakers locate prosodic information that distinguishes the two types of RCs. Rather than using conventional analysis methods, we conducted a neural network-based analysis which used multi-dimensional acoustic and articulatory signals. We observed two distinct distributional patterns: syntactically conditioned prosodic information was either widely distributed around the boundaries or narrowly distributed at certain locations. Both patterns were observed at B1 and B2. Furthermore, there were participants who did not use the same strategy of marking syntactic differences across the two boundaries.

Contrary to the predictions of many syntax-prosody interface theories, our findings showed that the information that distinguishes the two RCs is not necessarily restricted to the immediate vicinities of phrases boundaries; rather, the two RCs can differ at various locations around the boundaries. Further investigations should be conducted to find out why we see various patterns; for example, the two RCs may not just differ in prosodic organization but differ in other factors such as prominence structure. Yet, our findings show that it is important to investigate a wider region around the boundaries to accurately examine how distinct syntactic structures are produced differently. Additionally, the location of syntactically conditioned prosodic information varied across participants and also across boundaries. This poses a challenge to those theories which argue for an invariant mapping between syntactic structure and prosodic organization.

The findings from the current study propose a several directions for future research. First, F0 data may provide significant information on how speakers mark syntactic contrasts in their utterance. Second, we can conduct the same network analysis but with varying inputs and find out what type of information contributes most to the distinction between the two RCs. For instance, it is possible that acoustic signals contribute significantly to network accuracy in some participants, while articulatory signals are crucial for other participants. Even within the same participant, different types of information would affect the network classification differently depending on the location in an utterance.

Overall, this study showed how speakers convey syntactic contrasts through prosody, specifically focusing on its temporal aspect. Crucially, this study demonstrated that our novel network-based analysis method is a powerful tool to localize temporal information. Although this method was used to specifically examine the syntactic contrast, it has a potential to be applied to a wide variety of contexts in phonetic research.

5. References

Arnold, Doug (2007). "Non-restrictive relatives are not orphans". In: *Journal of linguistics*, pp. 271–309.

Cooper, William E and John M Sorensen (1977). "Fundamental frequency contours at syntactic boundaries". In: *The Journal of the Acoustical Society of America* 62.3, pp. 683–692.

Garro, Luisa and Frank Parker (1982). "Some suprasegmental characteristics of relative clauses in English". In: *Journal of Phonetics* 10.2, pp. 149–161.

Kim, Seung-Eun and Sam Tilsen (2020). "Speech rate and syntactically conditioned influences on prosodic boundaries". In: *Proc. 10th International Conference on Speech Prosody 2020*, pp. 434–438.

Nespor, Marina and Irene Vogel (1986). *Prosodic phonology*. Dordrecht: Foris Publications.

Selkirk, Elisabeth (2005). "Comments on intonational phrasing in English". In: *Prosodies: With special reference to Iberian languages*. Ed. by S. Frota, M. Vigário, and M. J. Freitas. Walter de Gruyter, pp. 11–58.

Tilsen, Sam (2020). "Detecting anticipatory information in speech with signal chopping". In: *Journal of Phonetics* 82, p. 100996.