Temporal localization of syntactically conditioned prosodic information

Seung-Eun Kim, Sam Tilsen (Cornell University)

This study investigates *when* in time the prosodic correlates of a syntactic contrast can be detected in acoustic and articulatory signals. Specifically, we attempt to localize information that distinguishes non-restrictive relative clauses (NRRCs) and restrictive relative clauses (RRCs), examples of which are shown in (1). On several accounts (e.g., Selkirk 2005), the two types of relative clauses differ in prosodic phrase structure, and this predicts that the utterances in (1) should differ in the vicinity of the phrase boundaries before (B1) and after (B2) the relative clause. To test this prediction, we used a neural network-based analysis procedure. The results showed that for some speakers, the syntactically conditioned prosodic information was distributed in a wide region around prosodic boundaries, while for the other speakers, the information was more concentrated at specific locations. For those speakers who showed concentrated patterns, there was variation in where prosodic information was located relative to phrase boundaries.

(1)

NRRC **[[A Mr. Hodd,]$_{ip}$ [who knows Mr. Robb,]$_{ip}$]$_{IP}$ [[often plays tennis.]$_{ip}$]$_{IP}$**

RRC **[[The Mr. Hodd who knows Mr. Robb]$_{ip}$ [often plays tennis.]$_{ip}$]$_{IP}$**

(ip: intermediate phrase, IP: intonational phrase)

Six native speakers of English (3M, 3F) participated in the experiment. Articulatory and acoustic data were collected with an NDI Wave Electromagnetic Articulograph (EMA). Speakers read target sentences at various speeds, cued by a moving visual analogue for speech rate. Blocks of NRRC or RRC sentences were alternated. For analyses, the neural network input was composed of 20 articulatory dimensions and 66 acoustic dimensions. Articulatory dimensions were the horizontal and vertical positions of the five articulator sensors (TT, TB, JAW, UL, LL) and each of their velocities. Acoustic dimensions were 33-dimensional broadband spectrogram and their first differences. The articulatory and acoustic data were extracted in 25ms steps, aligned to B1 or B2 across trials, and normalized by dimension within each speaker.

To detect the presence of information related to the prosodic contrast between relative clause types, we examined the classification accuracy of bidirectional LSTM (biLSTM) networks, using a procedure recently developed in Tilsen (2020). Each network was trained on a randomly selected half of the data, and the classification accuracy of the other half was recorded. Furthermore, to temporally localize information, the size and center of the input signal to the network were systematically varied. We started from the window center which was aligned at the segmental boundary of the end of the target name (i.e., *Hodd, Robb*). The window centers varied in 25ms steps up to 500ms before and after the boundary, which resulted in 41 centers for each B1 and B2. At each window center, different window sizes were used for network classification. The minimal window size was 25ms, and the windows increased in 25ms step up to 500ms.[1] Only windows which did not require zero-padding were used. The training and testing procedure were repeated separately for each speaker, 20 times for each analysis window, and mean network accuracy on the test data was calculated.

The results showed two qualitatively different patterns: prosodic information associated with the syntactic contrast was either distributed broadly across the 1000ms analysis region (see Figure

[1] Windows spanned both sides of the window centers. For instance, 25ms window was composed of 12.5ms on the left side and 12.5ms on the right side of the window center.

1, Speaker 1, 3, and 6), or was more concentrated at specific locations (Speaker 2, 4, and 5). For the concentrated distributions, we found differences in the location of syntactically conditioned prosodic information. At B1, Speaker 1 and 3 showed distributed patterns in that the classification accuracy was high at both pre and post-boundary regions. Speaker 6 also showed the distributed pattern; although the network showed the highest classification accuracy around the target name, over 80% of accuracy was observed at all window centers within the 1000ms analysis region. On the other hand, the rest of the speakers (Speaker 2, 4, and 5) showed a more concentrated pattern such that the high classification accuracy was found only at certain window centers. However, these speakers showed differences on where they locate critical information. The highest accuracy was found mostly at the pre-boundary region in Speaker 2, but it was found at the post-boundary region in Speaker 4. For Speaker 5, the network showed highest classification accuracy at the immediate region around the boundary. The results at B2 also showed either distributed or concentrated pattern, although speakers did not show the same pattern across B1 and B2.

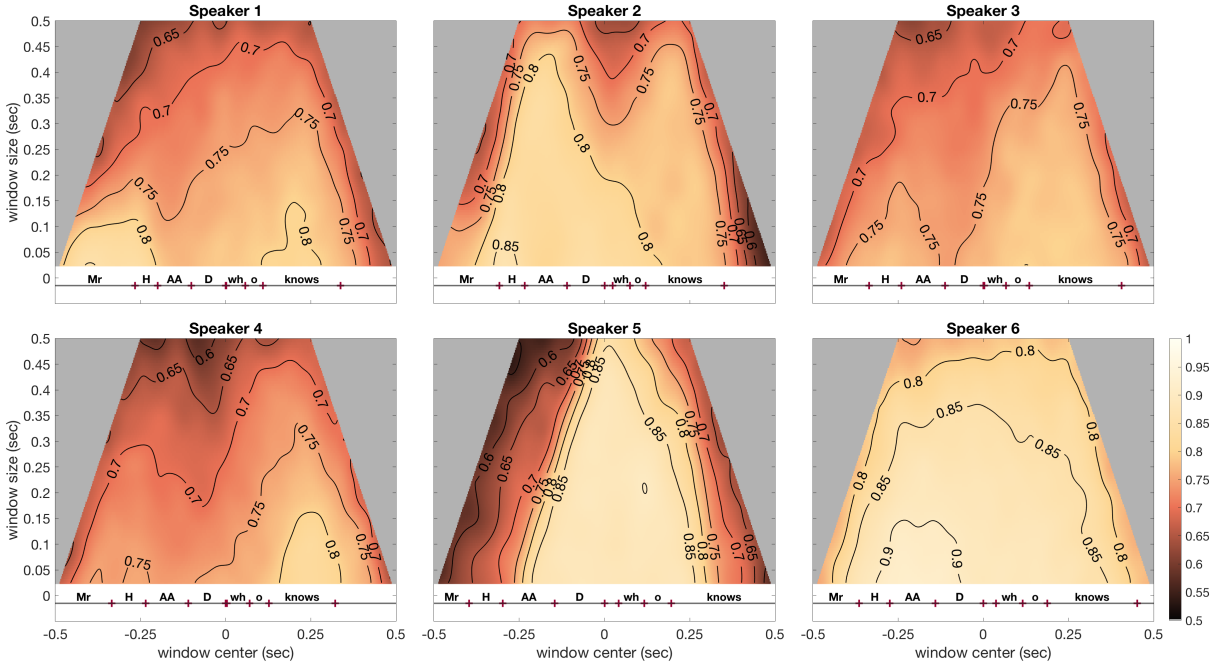

Figure 1. Classification results at B1 shown in heatmap along with the mean segment/word durations. The x-axis shows the location of the window center (0sec marks the end of the target name), and the y-axis shows the window size. The colors and the numbers represent network classification accuracy. The gray area shows that the windows at those locations could not be investigated as they include information that goes beyond the 1000ms analysis region.

In sum, this study investigated where in the utterance speakers locate prosodic information that distinguishes the two types of relative clauses. By using a neural network-based analysis method, we found two different patterns of information distribution: prosodic information may be widely distributed across phrases or may be concentrated at certain locations. This finding has important consequences for our understanding of how speakers convey syntactic contrasts through prosody.

References Selkirk, E (2005). Comments on intonational phrasing in English. *Prosodies: With special reference to Iberian languages* (pp. 11-58).

Tilsen, S (2020). Detecting anticipatory information in speech with signal chopping. *Journal of Phonetics, 82*.