# Processing Prosodic Boundaries in Natural and Filtered Speech

*Grace Kuo*

Department of Linguistics, Macalester College, St. Paul, USA

gkuo@macalester.edu

## Abstract

The prosody of an utterance can carry information that is critically important to understand the meaning of a sentence. Previous studies have shown that listeners are able to detect major prosodic boundaries in their native language in stimuli whose segmental information has been removed, such as low-pass filtered [1][2] and hummed speech [2][3][4][5]. The present boundary strength rating study is conducted on native and non-native speakers to Swedish, in an attempt to observe non-native speakers' accuracy in judging the upcoming boundary size in natural and filtered speech. 18 Taiwanese and 18 American English speakers were recruited for the rating task whose stimuli consisted of Swedish utterances from three prosodic boundary types (word boundary, phrase/tone sandhi group boundary, and Intonation Phrase boundary). In Experiment 1, participants rated the upcoming boundary strength on a slider for natural speech stimuli. In Experiment 2, they rated the boundary strength for filtered speech stimuli. The results show that both native and non-native speakers could accurately predict the upcoming prosodic boundary type in both natural and filtered speech. The acoustic analyses of duration, f0 range, f0 median, spectral tile, and harmonics-to-noise ratio reveal that both native and non-native speakers use these prosodic cues to make their judgment; however, they put different emphasis on different cues when they were presented with stimuli of different qualities (natural vs. filtered) and lengths.

## 1. Introduction

Previous studies of speech prosody have shown that listeners are able to predict upcoming prosodic boundaries. For example, Grosjean and Hirst [6] had the subjects listen to some part of an English sentence and asked them to predict how long the remaining sentence was. The results showed that English listeners were very accurate at predicting the amount of the rest of the sentence. However, their French listeners could only tell if a sentence ended, unable to differentiate between different amounts to come. Carlson, Hirschberg, and Swerts [7], on the other hand, found that English listeners were able to predict the strengths of the upcoming boundaries as well as Swedish listeners when the subjects were asked to express their judgment about the upcoming boundary in Swedish on a 5-point scale. This suggests that prosodic, rather than syntactic/semantic information was being used as a primary cue. In addition, the result of their follow-up study with Mandarin listeners showed that the length of the presented stimuli matters – Mandarin listeners could hear different Swedish boundaries only when presented with 2-second fragments, whereas English and Swedish listeners could differentiate the boundaries with either 2-second or one-word fragments. This finding indicates that language background affects the listeners' judgments.

Experiment 1 in the current study replicates [7]' experiment by recruiting American English and Taiwanese listeners to participate in a rating task, where half of the stimuli were from natural Swedish speech and the other half were its low-pass filtered version. It is predicted that American English listeners will accurately predict the Swedish upcoming boundaries with both 2-second and one-word fragments (as reported by [7]), and that Taiwanese listeners, speakers of another tone language, will predict the boundaries accurately only when presented with 2-second fragments (similar to [7]'s result with Mandarin speakers).

Previous study such as [1] [2] [8] showed that native listeners were able to detect major prosodic boundaries in meaningless speech materials, including re-iterant speech, low-pass filtered speech and hummed speech. Since [7] found that non-native listeners could make use of the prosodic information (in the absence of the syntactic/semantic information) to predict the upcoming boundary, Experiment 2 has listeners participate in the same rating task as in Experiment 1, but the segmental information in the stimuli was absent/reduced (low-pass filtered). It is predicted that no rating difference will be found between the natural speech stimuli and the filtered speech stimuli in that both the American English and Taiwanese listeners are able to use prosodic information to make the accurate judgments.

## 2. Experiment 1: Natural Stimuli

### 2.1. Method

#### *2.1.1. Stimuli*

The Swedish stimuli were the same stimuli [7] used in their experiment. The stimuli were obtained from a 25-minute interview with a Swedish female politician and the interview was manually annotated for perceived boundaries by three experienced transcribers. Every word was marked as being followed either by an IP boundary, a phrase boundary, or a word boundary. Sixty 2-second speech fragments followed by either of the three boundary types were chosen: 20 word boundaries (labeled as "no break" in later analysis), 20 phrase boundary ("weak break") and 20 IP boundaries ("strong break"). All stimuli came in two lengths, 2-second fragment and one-word fragment. From each 2-second fragment, the last word was extracted to be the one-word fragment.

Therefore, there were 120 utterances in total (20 items x 3 breaks x 2 fragment lengths).

#### *2.1.2. Participants*

Eighteen Taiwanese native speakers and eighteen American English native speakers participated in this experiment. None of the American English and the Taiwanese listeners had previous knowledge about or prior experience with Swedish. Neither of them had hearing or language problems according to their self-report.



10.21437/SpeechProsody.2014-185

#### 2.1.3. Procedures

The subject individually judged the upcoming boundary strength for each natural utterance with an onscreen slider, whose position was manipulated by listeners from left ("small break") to right ("big break"). During the task, the subjects could choose to hear each stimulus more than once, but were encouraged to make their judgments by instinct. To minimize any possible learning effect, the stimuli were presented in a randomized order.

#### 2.1.4. Data Analysis

The position of the slider bar was recorded by the Matlab script on a scale from 0-100. These numerical values were converted into logarithmic values to reduce the skewing in the distribution. The logarithmic strength ("log strength" hereafter) were entered into the repeated measures ANOVA, with the two within-subject factors, "Break" (no break vs. weak break vs. strong break), and "Length" (2-second vs. one-word). Since different language background might result in different rating results, English listeners' data were separated from Taiwanese listeners' data.

### 2.2. Results – Boundary Strength Ratings

#### 2.2.1. English listeners

Repeated measures ANOVA reveals significant effects of "Break" ($F(2, 2154) = 74.32$, $p < .05$) and "Length" ($F(1, 2154) = 231.4$, $p < .05$). In other words, (i) listeners tended to give higher ratings for bigger boundaries, and (ii) listeners gave higher strength ratings for 2-seond fragments than for one-word fragments. In addition, listeners were able to differentiate all three breaks in either 2-second or one-word fragments. The results are shown in Figure 1. These findings were the same as the results found in [7] – English listeners were able to accurately predict all three boundaries in a nonnative language such as Swedish when presented with natural speech stimuli.

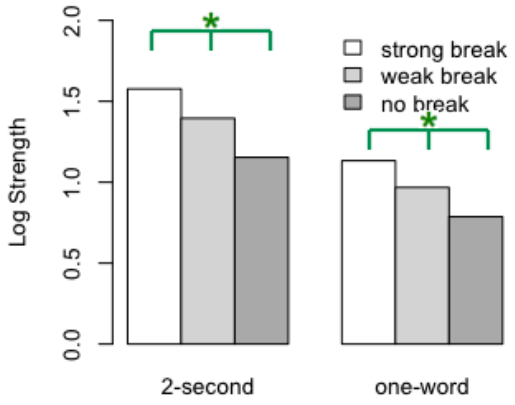


Figure 1: *English listeners' average logarithmic perceived boundary strength for Swedish Natural Stimuli. A significant difference between "Breaks" of either length is indicated with a line and asterisk above the bars.*

#### 2.2.2. Taiwanese listeners

Significant effects were found not only in "Break" ($F(2, 2154) = 61.09$, $p < .05$) and "Length" ($F(1, 2154) = 196.88$, $p < .05$), but also in their interaction ($F(2, 2154) = 3.18$, $p < .05$). The results are shown in Figure 2. These findings were similar to [7]'s study with Mandarin listeners – tone language speakers were able to accurately predict the upcoming three boundaries only when they were presented with 2-seoncd fragments.

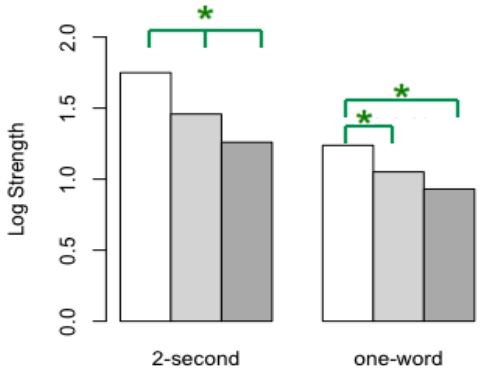


Figure 2: *Taiwanese listeners' average logarithmic perceived boundary strength for Swedish Natural Stimuli.*

## 3. Experiment 2: Filtered Stimuli

### 3.1. Method

The same 18 Taiwanese and 18 American English speakers participated in the boundary strength rating experiment. The stimuli were the low-pass filtered version of the 60 natural speech stimuli. They were generated by at a frequency cut-off of 400 Hz and 50 Hz smoothing, and the intensity was adjusted to 70 dB. The entire manipulation was done with a Praat [9] script. With low-pass filtering, most of the segmental information will be removed, yet the prosodic information, such as duration, f0, and some voice quality stay intact. Like Experiment 1, the stimuli also came in two lengths. Therefore, there were 120 utterances in total (20 items x 3 breaks x 2 fragment lengths). The experiment procedures were the same as those in Experiment 1.

In the analysis, the raw strengths are converted into the logarithmic strength, and the values were entered into the repeated measures ANOVA, which had "Break" and "Length" as the two factors.

### 3.2. Results – Boundary Strength Ratings

#### 3.2.1. English listeners

The results with the English listeners show that when they were presented with filtered Swedish stimuli, they rated he three breaks differently, ($F(2, 2154) = 21.01$, $p < .05$). In addition, they were able to correctly predict all three breaks when presented with 2-second fragments, but not with one-word fragments. The results are shown in Figure 3. It seems that filtering has prevented listeners from identifying the phrase boundaries when the stimuli is short.

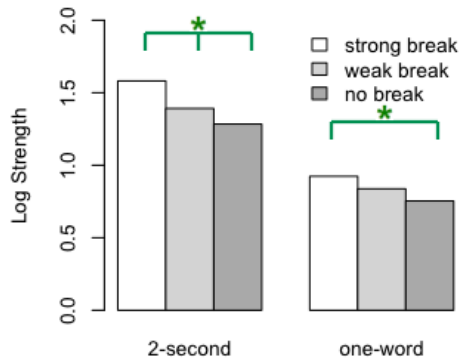


Figure 3: *English listeners' average logarithmic perceived boundary strength for Swedish Filtered Stimuli.*

### *3.2.2. Taiwanese listeners*

Significant effects were found not only in "Break" ($F(2, 2154) = 15.67$, $p < .05$) and "Length" ($F(1, 2154) = 402.2$, $p < .05$), but also in their interaction ($F(2, 2154) = 3.31$, $p < .05$). Similar to the results with the English listeners, Taiwanese listeners were still able to accurately predict the upcoming Swedish breaks in 2-second fragments when the stimuli were low-pass filtered. However, when the filtered stimuli contained only one-word, there were no difference in ratings between the three breaks.

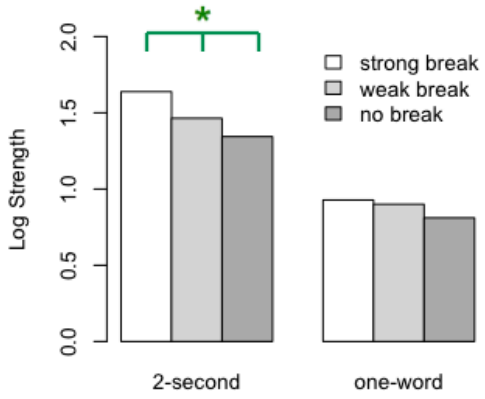


Figure 4: *Taiwanese listeners' average logarithmic perceived boundary strength for Swedish Filtered Stimuli*

## 4. Acoustic Correlates

In an attempt to identify the prosodic cues that could contribute to accurate boundary strength judgments in natural and filtered speech, we examined the acoustic measures from the last syllable of each stimulus: duration (=normalized vowel duration, and speech rate), pitch (including f0 range and f0 slope), harmonic amplitude/spectral tilt, harmonic-to-noise ratios, CPP, and Energy. The last syllable of each stimulus was labeled in Praat [9] and then the acoustic measures for the labeled portions were obtained using VoiceSauce [10]. The last syllable in Swedish contained usually part of a word.

As mentioned earlier, the filtered stimuli were a low-pass filtered version of the normal speech and the threshold was set at 400 Hz ( as indicated with the vertical line in Figure 5), thus, any information beyond 400 Hz would have been filtered out. For the filtered stimuli, the main available voice measures were the amplitude of the first harmonic (corrected H1), HNR05 (frequency range <500 Hz) and CPP.

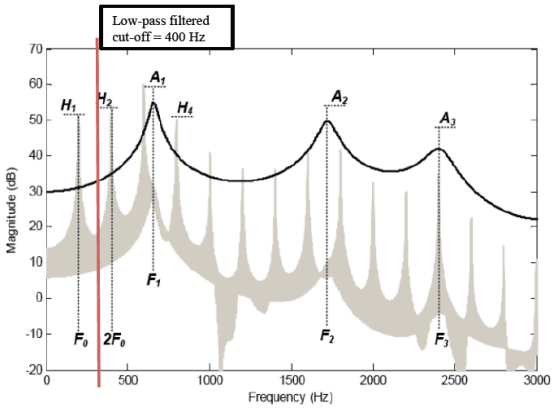


Figure 5: *An example spectrum. Adopted from [11]. The envelope in shade is the actual output from VoiceSauce [10]. The corrected values are obtained from this envelope.*

The regressions between the acoustic measures and the logarithmic boundary strength ratings are made. Multiple regression analysis was carried out for each type of stimulus (considering speech quality and the fragment length). The natural stimuli results are shown in Table 1 and the filtered stimuli results are shown in Table 2.

Table 1: *English and Taiwanese listeners listened to Swedish natural stimuli: marks show which acoustic measures contributed significantly to the regression equation.*

| | *English* | | Taiwanese | |
|---|---|---|---|---|
| | 2-sec | 1-wrd | 2-sec | 1-wrd |
| Duration | | | | |
| Rate | ✓ | ✓ | ✓ | |
| F0 range | ✓ | ✓ | ✓ | ✓ |
| F0 median | ✓ | | ✓ | ✓ |
| F0 mean | ✓ | | ✓ | |
| H1*-H2* | | ✓ | | |
| HNR05 | | | | ✓ |
| CPP | | ✓ | | ✓ |

The results show that for both English and Taiwanese listeners, when they were presented with 2-second natural stimuli, they would pay attention to durational and f0 cues. When the presented stimuli was one-word fragment, they would take voice quality into consideration.

Table 2: *English and Taiwanese listeners listened to Swedish filtered stimuli: marks show which acoustic measures contributed significantly to the regression equation.*

| | *English* | | Taiwanese | |
|---|---|---|---|---|
| | 2-sec | 1-wrd | 2-sec | 1-wrd |
| Duration | | | ✓ | |
| Rate | ✓ | | ✓ | |
| F0 range | | | ✓ | ✓ |
| F0 median | | | ✓ | ✓ |
| F0 mean | ✓ | ✓ | ✓ | |
| H1* | ✓ | ✓ | ✓ | |
| HNR05 | ✓ | ✓ | ✓ | ✓ |
| CPP | | | | ✓ |

The results reveal the significant correlations between the acoustic measures from the last syllable of the filtered utterances and the logarithmic boundary strength. It seems that English listeners rely on more voice quality cues for both lengths whereas Taiwanese listeners, who had better boundary strength when presented with 2-second fragments, tended to make use of more cues, including durational measures, to predict the upcoming boundaries. In addition, tone language speakers, such as Taiwanese, also used f0 range as a reliable cue when they were asked to predict upcoming boundary in a non-native language. Pitch is not only for lexical use for Taiwanese listeners.

## 5. Conclusion

In this study, we examined the perceived boundary strength indicated by Taiwanese and English listeners presented with Swedish natural and filtered stimuli. The distribution of the perceived boundary strengths shows that English listeners showed a three-way distinction in breaks in normal (both 2-seoncd and one-word) and filtered (only 2-second) stimuli. Taiwanese listeners also showed a three-way distinction in breaks in normal and filtered stimuli, but only when they were presented with 2-second fragments. These findings are consistent with [7]'s findings.

The acoustic analyses of normalized vowel duration, speech rate, f0 range, f0 slope, spectral tile, and harmonics-to-noise ratio reveal that non-native speakers use these prosodic cues to make their judgment; however, they put different emphasis on different cues when they were presented with stimuli of different qualities (natural vs. filtered).

## 6. Acknowledgements

The author would like to acknowledge Prof. Rolf Carlson, Julia Hirschberg and Marc Swerts for generously sharing their Swedish stimuli.

## 7. References

[1] de Rooij JJ. "Prosody and the Perception of Syntactic Boundaries", IPO ANNU Prog Rep, 10:36-39, 1975.
[2] Kreiman J. "Perception of sentence and paragraph boundaries in natural conversation", Journal of Phonetics, 10:163-175, 1982.
[3] t'Hart J., Collier, R., and Cohen, A.,"A perceptual study of intonation", Cambridge University Press, 1990.
[4] Pan, Ho-hsien, " Perceptual Tone Spaces and Taiwan Min Sandhi Rules", *personal communication*, 2011.
[5] Pannekamp, A., Toepel, U., Alter, K., Hahne, A., Friederici, AD., "Prosody-driven sentence processing: an event-related brain potential study", J. Cogn Neurosci, 17: 407-421, 2005.
[6] Grosjean, F. and Hirst, C. Using prosody to predict the end of sentences in English and French normal and brain-damaged subjects. *Language and cognitive Processes*, 11: 107-134, 1996.
[7] Carlson, R. , Hirschberg, J. and Swerts, M. Cues to upcoming Swedish prosodic boundaries – subjective judgment studies and acoustic correlates, *Speech Communication* 46: 326-333, 2005.
[8] de Rooij JJ. "Perception of prosodic boundaries", IPO Annu Prog Rep. 11: 20-24, 1976.
[9] Bowersma, P. and Weenink, D. Praat, Doing phonetics by computer (version 5.2.25) [Computer program]. Retrieved from http://www.praat.org, 2012.
[10] Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. "VoiceSauce, a program for voice analysis", *Proceedings of the 17th ICPhS*, 1846-1849. 2011.
[11] Shue, Y.-L. "The Voice Source in Speech production: Data, Analysis and Models". Doctoral dissertation, University of California, los Angeles. 2010.