Perceived prosodic boundaries in Taiwanese and their acoustic correlates

Grace Kuo

Department of Linguistics, UCLA, Los Angeles, CA, USA

gracekuo@humnet.ucla.edu

Abstract

This paper investigates boundary strength detection in Taiwanese and Swedish by Taiwanese listeners. Earlier production studies have suggested that the Taiwanese tone sandhi group (TSG) is an independent prosodic domain, yet no previous study has reported perceptual data to support this claim. This study presents listeners’ perceptual rating data on different prosodic domain boundaries and demonstrates correlations between acoustic measures of the stimuli and the perception ratings. The perceptual rating results show that the Taiwanese tone sandhi group is distinct from other prosodic domains. Significant correlations are found between ratings and pitch, and between ratings and acoustic voice quality measures.

Index Terms: Taiwanese tone sandhi, boundary strength, prosodic domain

1. Introduction

Tone sandhi refers to changes in lexical tone when words occur in specific contexts [1]. Taiwanese is a language with extensive tone sandhi, whose occurrence is conditioned at one level of prosodic structure, the Tone Sandhi Group. Within each tone sandhi group, the tones of all the syllables except for the last must undergo tone sandhi. (1a-b) demonstrate that the interpretation of a sentence depends heavily on the prosodic structure, but the prosodic domain boundaries do not always line up with the syntactic structure. (1a) shows how the sentence is syntactically parsed and (1b) shows how it is prosodically parsed and heard. In (1b), the sandhi tones are underlined, demonstrating that tone sandhi group boundaries are clearly delineated by means of tone realizations. (# is a prosodic boundary.)

(1) ‘He loves studying Taiwanese tone sandhi.’
  a. syntactic: [h]o[ai gen-kiu[tai-gie pen-tiao]s]
  b. prosodic: [i1 ai1 gen2-kiu2 tai2-gi5] # [e32 pen2-tiao3] #

Previous research analyzing a variety of acoustic measures (VOT and closure duration [2]; linguopalatal contact and articulatory seal duration [3]; degree of nasalization [4]; f0 velocity [5]) suggests that the Taiwanese tone sandhi group is an independent prosodic domain, and that speakers make distinctions between tone sandhi groups and other prosodic domains during speech production. However, no previous study has reported perceptual data demonstrating correlations between acoustic measures and perception ratings to support this claim. If the tone sandhi group could be identified perceptually, it would suggest that listeners are able not only to discriminate between prosodic domains, but also to exploit the relevant acoustic cues to process sentences correctly.

Carlson et al. [6] found that perception ratings could be used to identify different prosodic boundaries, and the correlations between the ratings and the acoustic measures indicate which cues listeners use to make the judgments. The authors tested whether listeners could determine boundary size in Swedish speech, and found that English listeners were able to predict the strength of upcoming boundaries as well as Swedish listeners, whether presented with a 2-second fragment or a one-word fragment. Furthermore, there were significant correlations between ratings and median f0, and f0 slope (=f0 velocity) and the presence of final creak. They concluded that listeners were using prosodic information, such as voice quality and pitch, rather than semantic information as a primary cue. However, their pilot study found that Mandarin listeners could hear different Swedish boundaries only when presented with 2-second fragments, suggesting that language background might affect the listeners’ judgments. In addition, the fact that Mandarin listeners needed more information implied that there might be a tradeoff between the use of f0 and the need for longer utterances, since f0 is lexical for tone language listeners.

In this paper, we replicate Carlson et al.’s experiment with Taiwanese listeners. The stimuli used contained both Carlson et al.’s Swedish stimuli and similar Taiwanese stimuli. The rating task and acoustic measures of the stimuli will allow us to see whether and how the Taiwanese tone sandhi group is perceived by Taiwanese listeners.

2. Experiment

For our studies, we presented spontaneous Taiwanese and Swedish utterance fragments to Taiwanese listeners, who were instructed to predict the upcoming boundary strength. Our goals were twofold: (a) to test whether listeners are able to discriminate the Taiwanese tone sandhi group boundaries from boundaries of other prosodic domains, and (b) to test whether boundary discrimination takes place in a non-native language, Swedish. The further analysis of acoustic measures will indicate potential cues that account for their judgments.

2.1. Stimuli

The Swedish normal utterances (n=60) were obtained from [6]’s experiment. The Taiwanese normal utterances (n=60) were selected and were originally extracted from an interview with a female Taiwanese speaker. The stimuli consisted of both Swedish and Taiwanese utterances, which were followed by either a word boundary (no break), a phrase/tone sandhi group boundary (weak break), or an IP boundary (strong break). All the stimuli came in two different lengths (2-second and one-word) and two different signal qualities (normal speech and filtered speech). The one-word stimuli contained only the final word in the 2-second stimuli. Recall that [6] found that listeners relied on prosodic rather than semantic information to make judgments.
This result is tested here for Taiwanese especially by the adoption of a low-pass filtered version of the same stimuli. Therefore, there were 480 utterances in the stimuli in total (20 items x 2 lengths x 3 breaks x 2 qualities x 2 languages).

2.2. Method

Eighteen native speakers of Taiwanese individually judged the upcoming boundary strength for each utterance with an onscreen slider whose position was manipulated by listeners from left ("no break") to right ("strongest break"). They listened to all stimuli by participating in two sessions, “filtered” followed by “normal”. To minimize any possible learning effect, the stimuli in each session were presented in a randomized order.

The task started with an instruction phrase, in which the author made sure that each subject fully understood the task. During the task, the subjects could choose to hear each stimulus more than once, but were encouraged to make the judgments by instinct. No feedback was given on their responses.

2.2.1. Acoustic measures

In an attempt to identify the prosodic cues that could contribute to accurate boundary strength judgment, we examined the acoustic measures from the one-word stimuli in four categories: duration, pitch, harmonic amplitude / spectral tilt, and harmonic-to-noise ratios. The targeted portions were labeled in Praat [7], and the acoustic measures for the labeled portions were obtained using VoiceSauce [8].

Durational measurements include normalized rime duration and speech rate. Previous research [9] has found that “segmental lengthening in the vicinity of prosodic boundaries is restricted to the rhyme of the syllable preceding the boundary.” Therefore, the length of the vocalic segment in the rime was measured as the rime duration. Normalized rime duration was calculated using (2), where \( d(i) \) is the rime duration of word \( i \), and \( \mu_i \) and \( \sigma_i \) are the mean and standard deviation of the duration of rimes in the same boundary type (i.e. no, weak, or strong). Normalized rime duration is expected to be longer at bigger boundaries.

\[
\hat{d}(i) = \frac{d(i) - \mu_i}{\sigma_i} \quad (2)
\]

Speech rate, calculated as the reciprocal of the rime duration, is expected to be lower at bigger boundaries. Both durational measures can be used to investigate final lengthening, which results in greater normalized duration values and lower speech rate values.

Pitch measurements include f0 velocity and f0 range measured in the vocalic rime. f0 range was calculated as the difference between the maximum and minimum f0 over the vocalic rime. f0 velocity was calculated by dividing the f0 range by the duration of the vocalic rime. [5] showed that for Taiwanese falling tones at domain-final position, the f0 range was the greatest before an IP boundary, followed by before a SYL (=syllable) boundary, then before a WRD boundary, and finally before a TS (=tone sandhi group) boundary. In addition, f0 velocity was the slowest before an IP boundary, followed by TS, then WRD, and finally SYL. Therefore, in the present study, f0 velocity is expected to be slower and f0 range to be wider as the boundaries get bigger.

The harmonic amplitude / spectral tilt measurements include H1-H2, H1-A1, H1-A2, and H1-A3. H1-H2 refers to the difference in amplitude between the first and second harmonics. A lower value of H1-H2 is often correlated with smaller glottal open quotient in a laryngealized voice, such as creaky voice, while a higher value is found in breathy voice ([8], [10, [11], [12]). A1, A2 and A3 are the amplitudes of the first, second and third formants, respectively. These measures generally pattern like H1-H2: the creakier the voice is, the lower the value of the measures.

Harmonic-to-noise ratios (HNR) were calculated for four frequency ranges (<500Hz, <1500 Hz, <2500 Hz and <3500Hz). Noise measures can indicate breathiness or aperiodic voice, both of which result in lower HNR values.

3. Results

3.1. Responses

Listeners’ perceptual judgments of boundary strength were converted into logarithmic strengths. The use of the logarithmically transformed strength values, rather than the perceived strength values reduces a wide range to a more manageable size. A within-subjects comparison found significant differences in log strength between different languages and different signal qualities. Therefore, the results reported below are separated by “language” and “quality”. Figure 1 presents the log perceived strength for the three different breaks in the stimuli.

For Taiwanese normal stimuli, a two-way repeated measures ANOVA revealed significant main effects of “break” \((F(2,34)=19.95, p<0.05)\) and “length” \((F(1,17)=32.46, p<0.05)\), as well as of their interaction \((F(2,34)=10.4, p<0.05)\). For Swedish normal stimuli, effects were also found of “break” \((F(2,34)=16.24, p<0.05)\), length \((F(1,17)=20.01, p<0.05)\) and their interaction \((F(2,34)=4.247, p<0.05)\). Tukey HSD post hoc tests of both Taiwanese and Swedish responses showed that the log strengths of the three breaks are different from one another \((p<0.01)\) and the difference came from the 2-second fragments. Thus, Taiwanese listeners, like Chinese listeners in [6], could tease three breaks apart only when presented with 2-second fragments, even in their own language. When listeners heard one-word fragments, they couldn’t differentiate between weak break and no break in either language. Also, the fact that
Taiwanese listeners could predict at least some Swedish boundaries indicates that semantic information is not a necessary cue to boundary detection.

For Taiwanese filtered speech stimuli, significant main effects of “break” \((F(2,34)=9.92, p<0.05)\) and “length” \((F(1,17)=23.62, p<0.05)\) were found. For Swedish filtered speech stimuli, significant main effects were also found of “break” \((F(2,34)=13.42, p<0.05)\) and “length” \((F(1,17)=27.7, p<0.05)\). No significant interaction was found in either Taiwanese or Swedish filtered speech. In addition, Tukey HSD post hoc tests revealed that Taiwanese listeners couldn’t distinguish the weak break from the other two breaks in Taiwanese when exposed to either length of fragment. However, they could hear three different breaks in Swedish when presented with filtered 2-second fragments, though they could hear no break difference when presented with filtered one-word fragments.

### 3.2. Acoustic differences

Table 1 shows that except for H1-A1 and H1-A3 in Taiwanese, the acoustic measures showed significant differences in terms of “break”. In order to observe which measures contribute to the three-way distinction between breaks, and thus to distinguishing TSG from both of the other boundaries in Taiwanese, Tukey HSD post hoc tests were performed. HNR15 in Taiwanese and f0 range, HNR15, HNR25 and HNR35 in Swedish were good predictors of the three-way distinction in the stimuli. In other words, for Taiwanese stimuli, the bigger boundary tended to show greater noise (HNR15). For Swedish stimuli, the presence of the bigger boundary tended to show not only greater noise (HNRs), but also a wider f0 range. The distributions of these measures are displayed in Figure 2.

![Figure 2: Boxplots of measures that are significantly different for each different break type.](image)

Overall, the negative correlations between log strength and H1-H2 across utterances with different length and quality in Taiwanese suggest that H1-H2 is a weak predictor for correct identification of break type, i.e. perceptually bigger boundaries tended to contain creakier voice. For Swedish, the positive correlation between log strength and f0 range and the negative correlations between log strength and HNRs across utterances with different length and quality suggest that these acoustic measures are also predictors, i.e. perceptually bigger boundaries tended to show wider f0 range and greater noise.

As for the durational measures, their correlations with log strength in Taiwanese one-word utterances suggest that for shorter fragments, listeners could use final lengthening (longer

### 3.3. Correlations

Since the one-word stimuli were segmented from the 2-second stimuli, a correlation of log strength in the two lengths in each language was expected. For a similar reason, a correlation between the log strength in normal and filtered speech in each language was expected as well.

For Taiwanese stimuli, we found a significant but weak positive correlation between judgments of one-word and 2-second utterances in normal speech \((r = 0.10; p < 0.05)\), but not in filtered speech. In addition, judgments of normal speech and filtered speech were positively correlated in both one-word utterances \((r = 0.52; p < 0.05)\) and 2-second utterances \((r = 0.49; p < 0.05)\).

### Table 1: ANOVA results for 11 acoustic measures. \(p<0.05\) is marked with an asterisk (*).

<table>
<thead>
<tr>
<th>Measures</th>
<th>Taiwanese</th>
<th></th>
<th>Swedish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (sd)</td>
<td>(F) (p)</td>
<td>mean (sd)</td>
<td>(F) (p)</td>
</tr>
<tr>
<td>speech rate</td>
<td>8.20 (3.9)</td>
<td>7.62 (4.2)</td>
<td>45.31 *</td>
<td></td>
</tr>
<tr>
<td>f0 velocity</td>
<td>0.35 (0.61)</td>
<td>0.44 (0.6)</td>
<td>10.46 *</td>
<td></td>
</tr>
<tr>
<td>f0 range</td>
<td>40.65 (35.0)</td>
<td>69.21 (68.0)</td>
<td>45.91 *</td>
<td></td>
</tr>
<tr>
<td>H1-H2</td>
<td>7.24 (2.9)</td>
<td>3.22 (5.6)</td>
<td>76.5 *</td>
<td></td>
</tr>
<tr>
<td>H1-A1</td>
<td>24.52 (7.9)</td>
<td>14.11 (10.8)</td>
<td>6.41 *</td>
<td></td>
</tr>
<tr>
<td>H1-A2</td>
<td>24.67 (9.2)</td>
<td>21.76 (9.7)</td>
<td>10.62</td>
<td></td>
</tr>
<tr>
<td>H1-A3</td>
<td>14.39 (8.8)</td>
<td>18.48 (12.8)</td>
<td>9.01 *</td>
<td></td>
</tr>
<tr>
<td>HNR05</td>
<td>28.27 (6.6)</td>
<td>13.43 (9.7)</td>
<td>160.1</td>
<td></td>
</tr>
<tr>
<td>HNR15</td>
<td>34.34 (7.1)</td>
<td>23.02 (10.0)</td>
<td>177.6</td>
<td></td>
</tr>
<tr>
<td>HNR25</td>
<td>38.62 (6.5)</td>
<td>28.89 (9.4)</td>
<td>156.2</td>
<td></td>
</tr>
<tr>
<td>HNR35</td>
<td>39.10 (5.9)</td>
<td>34.50 (8.9)</td>
<td>114.5 *</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Significant correlations (\(r\)) between acoustic measures and log strength. \(*=p<0.01\)

<table>
<thead>
<tr>
<th>Measures</th>
<th>Taiwanese</th>
<th></th>
<th>Swedish</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Filtered</td>
<td>Normal</td>
<td>Filtered</td>
</tr>
<tr>
<td></td>
<td>2-sec</td>
<td>1wrd</td>
<td>2-sec</td>
<td>1wrd</td>
</tr>
<tr>
<td>n.duration</td>
<td>0.10*</td>
<td>-0.08*</td>
<td>0.06</td>
<td>-0.08*</td>
</tr>
<tr>
<td>rate</td>
<td>-0.06*</td>
<td>-0.11*</td>
<td>-0.06</td>
<td>-0.14*</td>
</tr>
<tr>
<td>f0 velocity</td>
<td>-0.18*</td>
<td>-0.20*</td>
<td>-0.20*</td>
<td>-0.21*</td>
</tr>
<tr>
<td>f0 range</td>
<td>-0.14*</td>
<td>-0.18*</td>
<td>-0.20*</td>
<td>-0.21*</td>
</tr>
<tr>
<td>H1-H2</td>
<td>-0.08*</td>
<td>-0.13*</td>
<td>-0.13*</td>
<td>-0.14*</td>
</tr>
<tr>
<td>H1-A1</td>
<td>-0.08*</td>
<td>-0.06*</td>
<td>-0.06</td>
<td>-0.10*</td>
</tr>
<tr>
<td>H1-A3</td>
<td>-0.08*</td>
<td>-0.06</td>
<td>-0.06</td>
<td>-0.10*</td>
</tr>
<tr>
<td>HNR05</td>
<td>-0.08*</td>
<td>-0.12*</td>
<td>-0.18*</td>
<td>-0.21*</td>
</tr>
<tr>
<td>HNR15</td>
<td>-0.08*</td>
<td>-0.18*</td>
<td>-0.20*</td>
<td>-0.21*</td>
</tr>
<tr>
<td>HNR25</td>
<td>-0.09*</td>
<td>-0.08*</td>
<td>-0.20*</td>
<td>-0.21*</td>
</tr>
<tr>
<td>HNR35</td>
<td>-0.09*</td>
<td>-0.08*</td>
<td>-0.18*</td>
<td>-0.18*</td>
</tr>
</tbody>
</table>

\[ \text{**Figure 2**: Boxplots of measures that are significantly different for each different break type.} \]

\[ \text{**Table 2**: Significant correlations (}r\text{) between acoustic measures and log strength.} *\text{=}p<0.01 \]
duration and lower speech rate) in both normal and filtered speech to help make judgments. Furthermore, it appears that listeners used final lengthening (lower speech rate) only for longer fragments of Swedish.

The f0 measures do not correlate with log strength much in Taiwanese, which may be due to the fact that Taiwanese is a tone language whose sandhi tones and lexical tones are neutralized in terms of f0 [13]. On the other hand, there are more significant correlations between f0 measures and log strength in the Swedish stimuli; listeners apparently considered f0 as a predictor when they didn’t know if it was a tone language.

For the harmonic amplitude / spectral tilt measures, H1-H2 was a predictor in Taiwanese, but not consistently in Swedish data. The correlations between log strength and other spectral tilt measures appear to be rare, which suggests that listeners did not use them as a predictor to boundary types.

The noise measures, on the other hand, seem to be helpful when listeners were predicting boundary strengths in 2-second utterances. They had the tendency to predict a bigger boundary if they heard an aperiodic sound in a long fragment.

4. Discussion

In this paper, we examined the perceived boundary strength indicated by Taiwanese listeners presented with Taiwanese and Swedish stimuli, and the correlations between these strengths and potential acoustic cues, including durational, pitch, harmonic amplitude / spectral tilt and noise measures.

The distribution of the perceived boundary strengths shows that the Taiwanese tone sandhi group is an independent prosodic domain in that listeners make distinctions between tone sandhi groups and other prosodic domains in normal 2-second utterances.

Most of the acoustic measures also show significant differences in terms of “break”. Among these acoustic measures, HNR15 shows a three-way distinction for Taiwanese stimuli, and f0 range; HNR15, HNR25 and HNR35 show a three-way distinction for Swedish stimuli.

Correlation analysis reveals that several acoustic measures provide acoustic cues that listeners rely on. While creakiness is a strong predictor for Taiwanese stimuli, f0 range and HNRs are predictors for Swedish stimuli.

Carlson et al. [6] found that Scandinavian listeners could accurately predict upcoming boundary strength, whether presented with long (2-second) or short (one-word) stimuli, and so could American English listeners. The latter fact suggests that listeners make the judgment depending on prosodic cues, rather than semantic/lexical information. Moreover, they found that with the same Swedish stimuli, Mandarin listeners could correctly predict the upcoming boundary strength only when they heard the longer 2-second fragments, which suggests that language background might influence listeners’ judgments. In the present study, we find that Taiwanese listeners differ from Swedish listeners, but pattern similarly to Mandarin listeners, when are asked to make judgments regarding boundary strength in their native language. Taiwanese listeners could only detect a boundary difference in longer, 2-second utterances. However, the fact that the Taiwanese listeners could give correct judgments in a foreign language supports the claim that prosodic cues do play a more important role than semantic information in forming judgments. In addition, given that Taiwanese listeners perform like Mandarin but not American English listeners when asked to make judgments on the boundary strength in a foreign language, language background does make a difference for listeners. Tone language listeners apparently require a more global view of the stimuli, and can’t tease the different boundaries apart with simply local information.

Carlson et al. [6] found that the presence of final creak, median f0 and f0 velocity correlated with perceived boundary strength. In our study, we find that the prosodic cues vary with listeners’ familiarity with or the property of the tested language. Our Taiwanese listeners used H1-H2, an indicator of creakiness, to make judgments regarding Taiwanese stimuli, whereas f0 range and HNRs were the prosodic cues they used for Swedish stimuli. HNRs are indicators of breathiness or aperiodicity, so lower HNR values could result from either breathier or creakier voice quality. In any case, the change in HNR involves an alteration in voice quality. A possible explanation for the fact that Taiwanese listeners do not use f0 as a cue to boundaries in Taiwanese stimuli is that pitch variations are embedded in lexical items in Taiwanese. Being aware of that, listeners might decide to disregard f0 when judging the Taiwanese stimuli. Both Carlson et al. [6] and the current study agree that pitch and voice quality are the main predictors utilized by listeners. Furthermore, the specific prosodic cues in use vary with listeners’ familiarity with or the property of the target language.

5. Acknowledgements

The author would like to thank Rolf Carlson, Julia Hirschberg and Marc Swerts for providing the Swedish stimuli and Patricia Keating for valuable discussions.

6. References