Perception of Temporal Cues at Discourse Boundaries

Hsin-Yi Lin, Janice Fon
Graduate Institute of Linguistics
National Taiwan University, Taipei, Taiwan
jfon@ntu.edu.tw

Abstract

This study investigates the role of temporal cues in the perception at discourse boundaries. Target cues were penultimate lengthening, final lengthening, and pause duration. Results showed that different cues are weighted differently for different purposes. Final lengthening is more important for subjects to detect boundaries, while pause duration is more responsible in cueing the boundary sizes.

1. Introduction

Investigations of speech production have shown that a number of temporal cues in prosody are correlated with the presence of syntactic or prosodic boundary. For instance, the occurrence of final lengthening right before boundaries has been observed in languages, such as Mandarin, Japanese, and English, [1, 2, 3]. Similarly, post boundary pause is also an important cue [4, 5]. These cues not only signal the existence of phrase boundaries, but their strengths also reflect the hierarchies of boundaries. For example, generally speaking, the pauses of bigger discourse boundaries are longer than those that at smaller discourse boundaries [1, 4].

For the past thirty years, increasing numbers of studies have tried to understand how these prosodic cues of boundary in speech production are perceived by listeners for boundary detection and boundary size judgment. To examine the effect of prosody in identifying boundary location, some methods have been adopted. Among them, the earliest was the use of ambiguous algebraic expressions [e.g. “(A plus B) times C” or “A plus (B times C)”] or similar methods [6, 7, 8], where neither syntax nor semantics could provide listeners with enough clues about the intended boundaries for bracketing. Therefore, listeners needed to rely on prosody. Studies using this approach have found that duration pattern as well as other prosodic cues [7, 8] do help listeners to figure out the intended bracketing.

Fon [13] tested the effect of prosody by filtering out most of the segmental cues in the utterances. With only the lower formants left, subjects needed to rely on the cues of the F0 movement, amplitude, and temporal pattern. Results showed that without any information from syntax and semantics, subjects were still able to detect the boundaries, which further confirmed the role that prosody plays in the perception of boundary. However, since the results were based on the overall prosody, the role that each cue plays is still unknown to us.

Carson et. al. [10] used final one or two words in sentences to check whether the prosody in these fragments was enough for listeners to judge the existence/absence of the following boundary. Results showed that subjects achieved high accuracy, and their judgments were highly related with duration and pitch information of final syllable. The results of Carson’s study further pointed out that listeners rely heavily on the final part of the sentence for boundary detection.

In addition to boundary detection, it has also been found from the above studies that perceived boundary sizes are correlated with the strength of temporal cues [6, 9, 10, 11, 12]. In general, as the final lengthening or pause become longer, subjects tended to consider as the boundaries bigger.

These studies have provided us with insights on how close the relationship is between temporal cues and listeners’ perception at boundaries. However, when interpreting the effect of these cues, most of the previous studies did not fully partial out the confounding influence from other prosodic cues, such as intonation or amplitude. Therefore, the present study used a more controlled experimental approach to examine the function of these temporal cues, aiming to better understand their roles and importance for our perception at boundaries. The research questions we attempt to investigate are: Are temporal cues alone sufficient for subjects to identify boundaries? If so, would stronger cues invoke faster speed in identification and trigger percept of the larger boundaries?

2. Experiment 1: Boundary Identification

2.1. Method

Experiment 1 was conducted to examine how subjects detected boundaries when only temporal cues were available.

2.1.1. Subjects

Twenty native speakers of Taiwan Mandarin, aged from 18 to 25 years old, were recruited through advertisements on the internet. None had ever suffered from language disorders or listening disabilities. Since subjects were required to respond by pressing their right index finger on the response box button, only right-handed people were recruited.

2.1.2. Stimuli

Stimuli were utterances of 18 syllables, which were composed of three syllables, “bu”, “di”, and “ga” spoken by a native female speaker. The three syllables were selected because their concatenation would not generate any meaningful interpretation in Mandarin and Southern Min, which could prevent subjects from relying on semantic grouping when doing this boundary identification task. The duration for each syllable in the utterance was set at 170 ms, and temporal manipulations were made on boundary correlated syllables, whose parameters are displayed in Table 1. These manipulations were based on the production results of Taiwan Mandarin in [9]. The synthesized boundary was placed on the 5th, 7th, 9th, 11th, or 13th syllable within the 18-syllable utterance, and divided the utterance into two sub-stretches.

The first sub-stretch is called Sentence 1 hereafter, and the second sub-stretch Sentence 2. In order to avoid influences
from other prosodic cues, loudness for each syllable was adjusted to be auditorily similar, and pitch values were resynthesized to be 210 Hz throughout the utterances.

Table 1. Parameters for temporal manipulation.

<table>
<thead>
<tr>
<th></th>
<th>No Lengthening</th>
<th>Short Lengthening</th>
<th>Long Lengthening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penult syllable</td>
<td>170 ms</td>
<td>200 ms</td>
<td></td>
</tr>
<tr>
<td>Final syllable</td>
<td>170 ms</td>
<td>250 ms</td>
<td>290 ms</td>
</tr>
<tr>
<td>Pause</td>
<td>0 ms</td>
<td>400 ms</td>
<td>600 ms</td>
</tr>
</tbody>
</table>

The concatenation of the three syllables “bu”, “di”, and “ga” was governed by the following rules, as we wanted to avoid possible grouping effect from factors other than set manipulations: (1) Adjacent syllables must be different, which means there are no stimuli such as “ga ga” combination within the stretches; (2) There must be no ABAB patterns in the stretches, which means there are no stimuli such as “bu di bu di” combination in the stimuli; (3) The transitional probability between every two adjacent syllables was the same, so no specific combination will stand out and become a solidified unit; (4) The number of times for the three syllables “bu”, “di” and “ga” to be loaded with the boundary cues was balanced, so that the occurrence of any of the syllable would not predict the appearance of boundary. An example of stimuli is shown in Figure 1.

There were in total 80 conditions for our stimuli: Penult Syllable (2) x Final Syllable (3) x Pause Duration (3) x Sentence Length (5). Results showed three main effects: Final Lengthening: \( F(4, 2155) = 9.667, p < .05 \), and Sentence Length: \( F(4, 2155) = 39.737, p < .05 \). The second was Penult Lengthening (2) x Final Lengthening (3) x Sentence Length (5). Results showed that there are two main effects: Final Lengthening: \( F(2, 2157) = 204.031, p < .05 \), and Sentence Length: \( F(4, 2155) = 39.737, p < .05 \). The second was Penult Lengthening (2) x Final Lengthening (3) x Sentence Length (5). The results showed three main effects: Final Lengthening: \( F(1, 2155) = 39.737, p < .05 \), Pause: \( F(2, 2157) = 9.667, p < .05 \), and Sentence Length: \( F(4, 2155) = 31.414, p < .05 \). The main effect of Final Lengthening can be seen in Figure 2.

2.1.5. Judgment Test

In order to make sure that subjects’ responses in Experiment 1 were really triggered by the target boundaries we inserted, we performed a judgment test to check whether the designed cues could really be taken as boundaries.

Twenty-one subjects, aged from 18 to 25, who did not participate in Experiment 1 were recruited through an internet advertisement. None had ever suffered from language disorder or listening disability.

The 160 stimuli in Experiment 1 were organized in a random order with a 4-second inter-stimulus interval. The experimental session was preceded by ten practice trials, consisting of utterances similar to the real judgment test.

Subjects were required to listen to each stimulus very carefully, and then mark the last syllable of the first sentence on the list in the questionnaire. A short break was given every five minutes.

The answers by the subjects in the judgment test were counted as correct if the marked syllable corresponded to the pre-boundary syllable in our design. For each trial, accuracy of 70% was set as the threshold, below which the reaction time data of the trial from experiment 1 will be excluded from analysis. Results of judgment test showed that accuracy for all trials passed the threshold, which suggested that the cues manipulated in Experiment 1 indeed succeeded in marking boundaries.

2.2. Results

Analyses were done to test whether subjects could detect the boundaries synthesized by us, and if so, whether prosodic cues of bigger strength would evoke faster response by the listener. Also, we want to know which of the temporal cues is more relied on for subjects to identify discourse boundaries. Since some of the trial conditions were excluded due to their oddity to perception, as described in section 2.1.2, we had to run two sets of three-way repeated ANOVA, to check the effect of our manipulation. The first one was Penult Lengthening (2) x Final Lengthening (3) x Pause (2) x Sentence Length (5). Results showed that there are two main effects: Final Lengthening: \( F(2, 2157) = 204.031, p < .05 \), and Sentence Length: \( F(4, 2155) = 39.737, p < .05 \). The second was Penult Lengthening (2) x Final Lengthening (2) x Pause (3) x Sentence Length (5). The results showed three main effects: Final Lengthening: \( F(1, 2155) = 39.737, p < .05 \), Pause: \( F(2, 2157) = 9.667, p < .05 \), and Sentence Length: \( F(4, 2155) = 31.414, p < .05 \). The main effect of Final Lengthening can be seen in Figure 2,
where RT became shorter as final lengthening longer, and post-hoc test showed that RT for long lengthening was the shortest, and that for no lengthening was the longest ($p < .01$). In Figure 2 we can also see the interaction effect between Final Lengthening and Pause, where post-hoc test showed that Pause was only significant when the Final Lengthening was at the short-lengthening level: RT for no-pause was significantly longer than short-pause and long-pause ($p < .05$), but no significant different in RT between short pause and long pause. Figure 3 showed the main effect of Sentence Length, where the post-hoc test showed that the RT for 5-syllable utterance was the longest, and for 11 and 13-syllable were the shortest ($p < .01$). Generally speaking, longer sentence length elicited shorter RT.

Figure 2: RT for Final Lengthening x Pause.

Figure 4 shows the relationship between penult lengthening and final lengthening. It showed that the RT would be shorter when penult duration was lengthened. Post hoc test showed that the effect of penult was only significant when the final duration was at the short-lengthening level: $t (1078) = 5.049$, $p < .05$.

Figure 4: RT for Penult Lengthening x Final Lengthening.

3. Experiment 2: Boundary Strength Rating

3.1. Method

Experiment 2 was aimed to investigate how temporal cues influence subjects’ percept of the sizes of boundaries.

3.1.1. Subjects

Subjects included 10 males and 10 females, aged from 18 to 25 were recruited through an internet advertisement. None had ever suffered from language disorder or listening disability.

3.1.2. Stimuli

Stimuli were utterances of 12 syllables, which were composed of three non-sense syllables, “bu”, “di”, and “ga” as those in Experiment 1, and at the 7th syllable in each stimulus, the target boundary cues were embedded. Manipulations of the temporal cues were the same as in Experiment 1.

There were in total 17 conditions for our stimuli: Penult Syllable (2) x Final Syllable (3) x Pause Duration (3) = 18 kinds of cue combination in total, and the condition where lengthened penult accompanying neither final lengthening nor pause was considered odd and was excluded. Therefore, the there remained 17 kinds of conditions. 6 utterances were generated for each condition, so there were $17 \times 6 = 102$ stimuli in total.

3.1.4. Procedure

Stimuli were presented with E-prime 1.0 on a portable computer (ASUS A3000) with a 15-inch screen. Subjects listened to the stimuli through a pair of SONY MRD-7502 earphones and made responses by pressing the keyboard of the portable computer.
screen asking them how big they thought the boundary was, and then they would answer by pressing the corresponding key, where “0” represented no boundary, “1” small boundary, “2” a medium boundary, “3” a medium-to-big boundary, and “4” a very big boundary. Right after the first question, a second question slide would be prompted asking how confident they were with their answer for the last question, and they would rate from 0 to 4 for their degree of confidence, where bigger number indicated stronger confidence. The above questions needed to be answered within 3 sec, before the slide disappeared. In order to make sure that subjects understood the concept of boundary strength, careful accounts were given, such as boundary between paragraphs is big and between clauses is small. Two practice trials were provided. Then the real experiment begun. The whole procedure took about 15 minutes, with a short break of every 5 minutes.

3.2. Results

In Experiment 2, subjects were asked to judge on a Likert scale how big the heard boundary was between the first and second sub-stretches. The confidence rating for each trial was used as a threshold, where only those boundary ratings with confidence level over 3 were included for analyses.

Since some of the trial conditions were excluded due to their oddity to perception, we had to run two sets of three-way repeated ANOVA, to check the effect of our manipulation. The first ANOVA was run without no-pause conditions: Penult Duration (2) x Final Duration (3) x Pause (2). Results showed two main effects: [Final Duration: \( F(2, 1115) = 52.202, p < .01; \) Pause: \( F(1, 1116) = 148.347, p < .01 \)]. The second ANOVA was run without no-final-lengthening conditions: Penult Duration (2) x Final Duration (2) x Pause (3). Results showed only one main effect of Pause: \( F(2, 1115) = 961.244, p < .01 \).

For the effect of Final Duration, as in Figure 5, Post hoc test showed that the rating for no-final-lengthening was significantly smaller than short final-lengthening and long final-lengthening \( p < .05 \), but there was no difference between short and long final lengthening. The effect of Pause, also shown in Figure 4, indicated that long pause would elicit the biggest rating, while no pause the smallest \( p < .01 \). In general, the rating for bigger boundary increased as the pause duration elongated, while longer final lengthening did not seem to have such effect.

4. Discussion and Conclusions

The results of the present study suggested that temporal cues alone could mark speech boundaries, and different temporal cues played different roles. Experiment 1 showed that subjects would speed up in making responses when the final syllable gets longer, while pause only matters when it follows a short-lengthening final syllable, and when the final syllable is of long-lengthening, the information from pause is not important anymore. What is more, there are no differences in RT for short pause and long pause, suggesting that pause is not as important as final lengthening in this regard.

The results in Experiment 2 showed that, when asked to judge the size of boundary, subjects relied more on pause duration. Longer pauses would elicit higher boundary rating. Though existence of final lengthening also helped elicit higher rating, its duration had no effect.

The results of the present study helped us know better about the role that temporal cues play in boundary perception. Generally speaking, for listeners, final lengthening is the main cue signaling end of speech, and pause duration the boundary strength. In the future, we aim to expand this study to include other cues, such as pitch and amplitude, to see how these prosodic cues interact with each other.

5. References