



# Perception on Pitch Reset at Discourse Boundaries

*Hsin-Yi Lin, Janice Fon*

Graduate Institute of Linguistics

National Taiwan University, Taipei, Taiwan

niceshelly46@yahoo.com.tw, jfon@ntu.edu.tw

## Abstract

This study investigates the role of pitch reset in discourse boundary perception. Previous production studies showed that pitch reset is a robust correlate of discourse boundaries. It not only signals boundary location, but also reflects boundary sizes. In this study, one aims to investigate how listeners perceive and utilize this cue for boundary detection. Results showed that listeners' perception on this cue corresponded to the patterns found in speech production. What is more, evidence showed that what listeners rely on is the amount of reset, rather than the rest pitch height.

**Index Terms:** speech perception, pitch reset, boundary

## 1. Introduction

The role of pitch cues as boundary correlates has been established in the literature of speech production, where it was found that the pitch contour within a structural unit shows a declining pattern, and when the next unit starts, the declined pitch will be reset [1, 2, 3, 5]. Evidence showed that these pitch boundary cues are not just by-products of physical aerodynamic variations, but more importantly, they are being used by speakers for linguistic purposes.

The declining pitch contour is found to be the default pitch pattern in spoken sentences, which is known as the intonation of declarative sentences [2]. Even with the occurrence of short pitch excursions with accents [6], the excursive pitch for accents will still fall back on the same track, and the overall pitch trend will keep down-trending. Declination can thus be seen as a grouping mechanism, where elements connected by the same declination contour will be recognized as belonging to the same unit [3, 7]

The close relationship between declination and structural units can also be seen from the good alignment between the two. It has been indicated that the slope of the declination contour is dependent on unit length. Longer units are with flatter slopes, while shorter units have steeper ones [8]. Such a time-dependency implies that the declination contour is a product pre-planned with the length of the unit, not just an uncontrollable physical consequence. Evidence also showed that when producing longer units [9], speakers actively adjust the onset pitch to a higher level in order to maintain the needed slope for declination. All these findings suggested that pitch declination patterns closely with linguistic units, and can therefore serve as an indicator in signaling their onset and offset.

In addition to pitch declination, pitch reset is also found to be a cue actively manipulated by speakers for linguistic purposes. Several studies have shown that pitch reset is reflective of boundary hierarchy. Bigger boundaries are generally accompanied with bigger reset [1, 5, 8]. Also, as aforementioned, the onset pitch of a declination contour will be adjusted according to the length of the unit, which means

the amount of pitch reset can be indicative of the length of the following units. This implies the possibility for listeners to use this cue to locate boundaries and judge boundary sizes.

In regard to this close relationship between boundary and pitch cues, the present study aims itself to examine whether and how listeners utilize these cues for locating boundaries in speech. Similar questions have been asked by some perception studies using different approaches. Streeter [10] tested how listeners used resynthesized prosodic information to parse structurally ambiguous sentences and found that both temporal and pitch cues are important for listeners to determine the position of boundaries. Using naturally elicited speech, Price et al. [11] had similar findings. This shows that pitch information is a cue to boundary locations. Grosjean (1983) [12] found that when presented with a single syllable, listeners could accurately determine whether the syllable is followed by a boundary, and how far away the boundary is. Such predictability was shown to have a strong relationship with the prosodic cues (pitch, temporal, amplitude cues) carried by the syllable. Some other studies approached this question by asking untrained listeners to label the size of different boundaries in spoken corpora [13, 14], and then checking the relationship between the labeled sizes and the measured boundary cues. Results of their studies showed that pitch reset indeed has a significant effect on listeners' judgment of boundary sizes. Boundaries with bigger reset tend to be considered as of bigger sizes.

To sum up from the results of these previous studies, one can find that pitch cues seem to play an important role in signaling boundaries for listeners. However, how exactly the target pitch cues were used by listeners were still not very clear, because none of the above studies examined pitch cues alone. What they have claimed about the relationship between pitch cues and boundary perception may be due to a combination of various other cues, such as temporal or semantic cues. Therefore, in order to fix the above problems and to better understand the relationship between pitch cues and perception on boundaries, the present study used a more controlled experimental method to investigate this question.

## 2. Experiment: Pitch Reset

### 2.1. Method

This experiment was conducted to examine how different degrees of pitch reset affected subjects' boundary detection.

#### 2.1.1. Subjects

Thirty-one native speakers of Taiwan Mandarin, aged from 18 to 25 years old, were recruited as subjects. None had ever suffered from language disorders or listening disabilities. However, in the following analyses, data from five subjects were excluded. Two failed to pass the practice session and

could not follow our instructions, and three showed absent-mindedness during the experiment. As these subjects might not respond to the boundary positions correctly and might produce false positive RT, their responses were not included in the analyses.

### 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. By using nonsense syllables one could prevent subjects from relying on semantic grouping when doing this boundary identification task. The synthesized boundaries were placed on the 7<sup>th</sup>, 9<sup>th</sup>, or 11<sup>th</sup> syllable within the 18-syllable utterance, which divided the utterance into two sub-stretches. The first sub-stretch is called Sentence 1 hereafter, and the second sub-stretch Sentence 2. For the sake of naturalness, both Sentence 1 and Sentence 2 were superimposed with a non-linear declining pitch trend, which was derived from the pitch contour in Mandarin declarative sentences described in [3]. The contour of non-linear decay is featured with a fast drop at the beginning, and then the dropping speed gradually slows down to an asymptotic value as sentence proceeds. The basic formula for generating the non-linear decay is given below:

$$P_i = \alpha P_{i-1} + \beta (P_1 - \alpha P_1)$$

$P_i$  is the pitch height in Hz of the  $i$ th syllable in the sentence. The coefficient  $\alpha$  is to control how fast the pitch declines to the asymptote, and  $\beta$  is to set the asymptotic value. An illustration of non-linear decay in the pitch contour is shown in Figure 1, with the settings shown in Table 1.

Stimuli were resynthesized into two sets of registers. To vary the pitch register of stimuli is beneficial to our observation on pitch reset. If the effect of pitch reset remains the same in different registers, then one can be surer that what matters is the amount of reset, not the pitch height after reset. For the lower register, the initial pitch height of Sentence 1 was set at 220 Hz, and 230 Hz for the higher one. As Sentence 1 proceeds to its last syllable, the pitch starts to decline non-linearly. The amount of declination for every Sentence 1 was 20 Hz. In other words, the lower register set ended at 200 Hz, while the higher register set ended at 210 Hz. Right after Sentence 1, without any pause in between, followed Sentence 2, and the pitch height of the initial syllable in Sentence 2 was either 10, 15, 20, 25, or 30 Hz higher than that of the last syllable of the preceding Sentence 1. Pitch contour in Sentence 2 also declines non-linearly as it proceeds to its end, and its declination ratio  $\alpha$  is set to be the same as that of the Sentence 1. The pitch contour from a real stimulus is shown in Figure 2 as an illustration. In order to avoid influences from other prosodic cues, loudness for each syllable was adjusted to be auditorily similar, and the duration for each syllable in the utterance was set at 200 ms.

To summarize, there were 3 levels of sentence length before the boundary (7-, 9-, 11-syllables long), 2 levels of pitch register of the whole utterance (lower and higher), and 5 levels of pitch reset at the boundary position (10, 15, 20, 25, 30 Hz reset), which made Length (3) × Register (2) × Reset (5) = 30 conditions in our stimuli. For each condition, 3 different utterances were generated, so there were 90 stimuli in total.

The concatenation of the three syllables “bu”, “di”, and “ga” was governed by the same rules used in Lin & Fon (2009) [4] in order to avoid possible grouping effect from factors other than the manipulations.

Table 1. Settings for pitch declination of stimuli in higher register ( $P_1 = 230$  Hz)

Sentence Length	$\alpha$	$\beta$	$P_1$ (Hz)
7-syllable	0.91	0.8	230
9-syllable	0.93	0.8	230
11-syllable	0.94	0.8	230

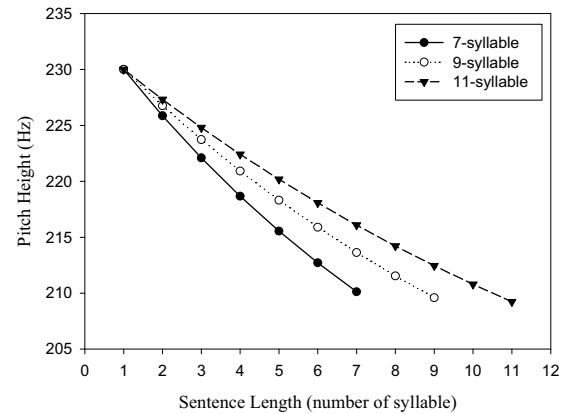


Figure 1: An illustration of non-linear decay in the stimuli at higher register.

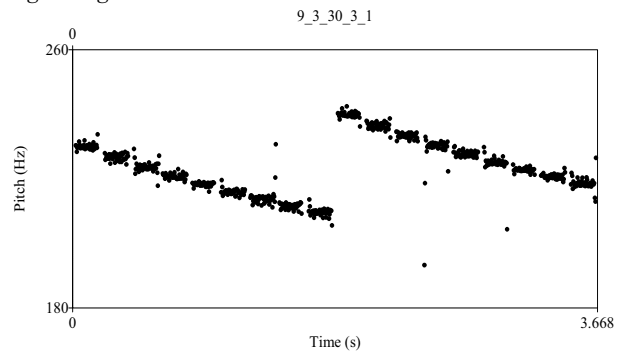


Figure 2: Pitch contour of a stimulus (9-syllable long in Sentence 1, at the higher register, with 30 Hz reset).

### 2.1.3. Equipment

E-prime 2.0 was used to present the experiment, a Serial Response Box to record the reaction time, and a pair of Sony MDR7502 earphones was used for subjects to listen to stimuli.

### 2.1.4. Procedure

Subjects were seated in front of a computer in a quiet room with earphones put on. They were told that in each trial, there were two adjacent meaningless sentences presented after a reminder bell, and they needed to press the button on the reaction box as soon as they heard Sentence 2 started. They were reminded to concentrate on the task by watching a small fixation point on the screen. A practice session of five trials was provided before the real experiment began. The stimuli trials were presented in a randomized order. A 1-second interval was inserted between every two trials. The whole experiment lasted about 20 minutes, with one short break of around 30-second inserted every five minutes. Three dummy trials were put at the beginning and the end of the experiment.

### 3. Results

Analyses were done to test whether subjects could detect the synthesized boundaries relying only on pitch cues, and if so, whether or not cues of bigger strength would evoke faster response by listeners. Reaction time (RT) was measured from the time point where pitch resets took place to the time point subjects responded.

A three-way repeated ANOVA, Pitch Reset (5) x Pitch Register (2) x Sentence Length (3), was conducted to check the effect of our manipulation. Results showed that all the three factors have main effects: Pitch Reset:  $F(4, 308) = 32.78$ ,  $p < .01$ ,  $\eta^2 = .299$ ; Pitch Register:  $F(1, 77) = 11.61$ ,  $p < .01$ ,  $\eta^2 = .131$ , and Sentence Length:  $F(2, 154) = 197.26$ ,  $p < .01$ ,  $\eta^2 = .719$ . Interaction effects were also found in the analysis: Register x Reset:  $F(4, 308) = 3.379$ ,  $p = .01$ ,  $\eta^2 = .042$ ; Sentence x Register:  $F(2, 154) = 4.198$ ,  $p < .05$ ,  $\eta^2 = .052$ ; Sentence x Reset:  $F(8, 616) = 3.48$ ,  $p < .01$ ,  $\eta^2 = .043$ . In the rest of this section, the post-hoc tests of each interaction effect will be presented.

#### 3.1.1. Reset x Register

The interaction effect between Register and Reset can be seen from Figure 3. As shown in Figure 3, the general trend indicated that as the amount of Reset becomes bigger or Register becomes higher, RT becomes shorter. Post-hoc tests with Bonferroni correction showed the effect of Register is only significant when the amount of reset is as small as 10 Hz:  $F(1, 466) = 9.038$ ,  $p < .01$ , and as the amount of reset grows, the effect of Register will disappear. As for the effect of Reset, post-hoc revealed that when the register is at the lower level, the RT for 10 Hz reset is significantly longer than the RTs for all the other resets ( $p < .01$ ), and the RT for 15 Hz reset is significantly longer than that for 25 Hz and 30 Hz ( $p < .01$ ). There is no significant difference between the RTs for 15 Hz and 20 Hz, 20 Hz and 25 Hz, nor between 25 Hz and 30 Hz. When the register is at the higher level, the RT for 10 Hz reset is significantly longer than the RTs for all the other resets but 15 Hz ( $p < .01$ ), and the RT for 15 Hz reset is significantly longer than that for 30 Hz ( $p < .01$ ). There is no significant difference between the RTs for 10 Hz and 15 Hz, between 15 Hz and 20 Hz, nor between 20 Hz, 25 Hz and 30 Hz.

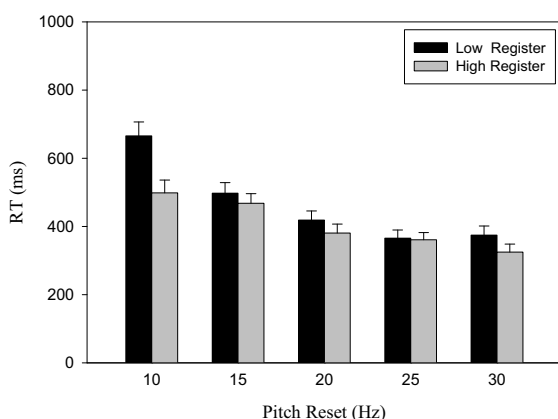


Figure 3: RT for Reset x Register.

Results from this test showed that bigger resets are not always more helpful, where resets bigger than 20 Hz do not significantly shorten subjects' response time. Also, resets between 15 Hz and 20 Hz make no difference either. This kind of pattern suggests that listeners do not respond to the amount

of pitch reset in an iconic manner, because if so, then one should have seen that bigger resets should always elicit faster responses. Rather, subjects seem to respond to reset in a categorical-like manner, and there are about three categories: RT for 10 Hz, RT for 15 to 25 Hz, and RT for 25 Hz above. This categorical patterning corresponds to what was found in production data [5], in which pitch reset for a smaller discourse boundary is about 20 Hz, and that for a bigger boundary is about 35 Hz. This difference suggests that as long as listeners can differentiate between pitch excursions with a 15 Hz difference, they can discern different boundary sizes. Therefore, to differentiate between resets with smaller differences may not be necessary, which further leads to the categorical pattern in their perception, as shown in the results. Another point worth noticing is that in general, there is no significant difference between higher and lower registers, except when the reset is small, which suggests that as long as listeners can gain enough information from reset, what they rely on for making judgment is not the absolute height of the reset pitch, but the amount of pitch reset. Otherwise, one should have seen a more robust effect of pitch register.

#### 3.1.2. Sentence Length x Register

The interaction effect between Sentence Length and Register can be seen from Figure 4. As shown in Figure 4, the general trend indicated that as the length of sentence before the boundary becomes longer, RT becomes shorter. Post-hoc tests with Bonferroni correction confirmed this observation, showing that the RTs for sentences with 7-syllable are significantly longer than that for 9-syllable long, which are also longer than that for 11-syllable long ( $p < .01$ ). This result shows that given longer time before making responses, subjects can have more preparation for their performance, and so the response time will be shorter [14]. However, for Register, its effect is only significant when the length of sentence is 7-syllable long:  $F(1, 778) = 13.629$ , ( $p < .01$ ). As the sentence becomes longer, the effect of register disappeared. The result indicates that pitch register plays only a subordinate role here. Its effect is only significant when other information is not sufficient to listeners.

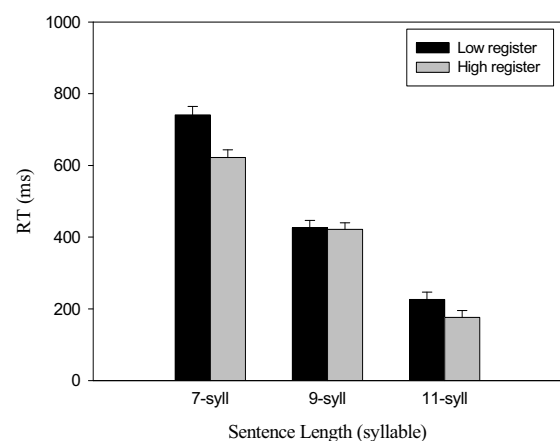


Figure 4: RT for Sentence Length x Register.

#### 3.1.3. Sentence Length x Reset

The interaction effect between Sentence Length and Reset can be seen from Figure 5. As shown in Figure 5, Sentence Length is a strong effect, where longer sentences will always elicit shorter RT no matter at which reset level ( $p < .01$ ). As for the

effect of Reset, the results showed that when the sentence length is shorter, its effect is more obvious: when sentence is 7-syllable long, RT for 10 Hz reset is significantly longer than that of 15 Hz ( $p < .05$ ), which is also significantly longer than that of 20 Hz and above ( $p < .01$ ); when Sentence 1 is 9-syllables long, results of RT showed a similar pattern as when Sentence 1 was 7-syllables long, only that there is no difference between 10 Hz and 15 Hz, nor between 15 Hz and 20 Hz. However, when the sentence is as long as 11-syllables, no effect can be found for Reset.

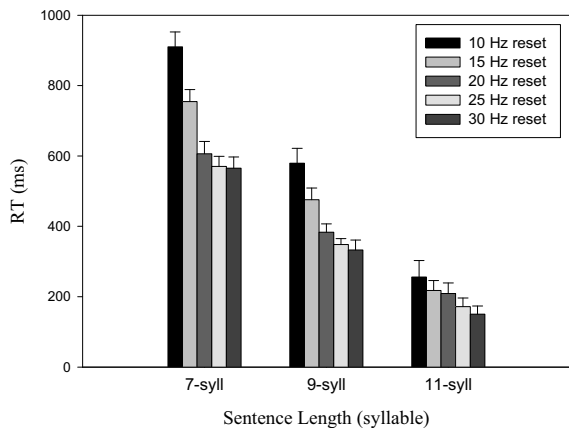


Figure 5: RT for Sentence Length  $\times$  Reset.

#### 4. Discussion & Conclusions

The present study aims to investigate the role that pitch cues play when listeners perceive boundaries. Results showed that listeners rely on pitch reset in a categorical fashion, and about three groups can be found: 10 Hz, 15-20Hz, and 25 Hz and above. As aforementioned, such a pattern may have to do with the fact that the differences of pitch reset among different boundary sizes are actually bigger than the manipulated steps in our experiment. Therefore, for the practice in real life, listeners do not need to resolve the amount of pitch reset in a fine linear fashion for detecting boundaries, which may be the reason that one observed the categorical pattern in the result shown in the present study. Another possible cause for this pattern may have to do with listeners' ability to detect difference in frequency. Previous studies showed that frequency jnds are about 3 Hz between 125-2000 Hz [16]. Since the increasing frequency step in this experiment is 5 Hz, which is quite close to the threshold, it may explain listeners' difficulties in showing fine-grained discrimination in the present study. Another cue in the experiment is pitch register, which is revealed to be a subordinate cue for the task of boundary detection. Results from the interaction effect showed that its influence is vulnerable and disappears fast when other cues get stronger. Such a result further confirms that the effect which reset has on listeners is not the absolute pitch height it resets to, but the relative amount of the reset.

The observed effect of sentence length may be caused by the fact that subjects know the possibility for a boundary to occur will increase as a sentence becomes longer. Such an expectation may lower the thresholds of information, therefore eliciting shorter RTs. Also, with longer sentences, subjects can have more time to prep their motor system to react to the imminent boundaries, which may also be an attribution for its effect [15]. Although this is more like an artifact of the experimental design, the same holds in the conversation in real

speech. Boundaries are also more likely to occur as sentences become longer.

One point worth mentioning is that about 10% of the RTs are of negative values, whose occurrence implies that subjects were hinted by some cues before pitch reset occurs. One possible hint is from the non-linear decay in pitch contour. As aforementioned, the declination speed slows down at the later part of the contour, meaning that subjects may be able to foretell the imminent boundaries due to this change. This phenomenon suggests the potential role that pitch declination plays for listeners to detect boundaries, where it acts not just as a grouping mechanism, but also as an indicator to signal the coming boundary [12]. Based on the above observation, one's next step will be to investigate the role of pitch declination in boundary perception, and to see how these pitch cues affect listeners' judgment on discourse hierarchies.

#### 5. Acknowledgements

The authors would like to thank the reviewers for their valuable suggestions. This study was supported by National Taiwan University, project number 99R90111.

#### 6. References

- [1] Ladd, D. R., "Declination reset and the hierarchical organization of utterances," *J. Acoust. Soc. Am.*, 84: 530-544, 1988.
- [2] Shen, Xiao-Nan Susan, *The Prosody of Mandarin Chinese*. University of California Press, 1990.
- [3] Shih, C., "A declination model of Mandarin Chinese," in A. Botinis [Ed.], *Intonation: Analysis, Modelling and Technology*, 243-268, Kluwer Academic Publishers, 2000.
- [4] Lin, H.-Y. & Fon, J., "Perception of temporal cues at discourse boundaries," *Proceedings in Interspeech 2009*. Brighton, United Kingdom, 2009.
- [5] Fon, J., *A Cross-Linguistic Study on Syntactic and Discourse Boundary Cues in Spontaneous Speech*. Ph.D. dissertation. The Ohio State University, 2002.
- [6] Pierrehumbert, J. B., *The phonology and phonetics of English intonation*. Doctoral dissertation. MIT, 1980.
- [7] Tseng, S.-C., "Linguistic markings of units in spontaneous Mandarin," in *ISCSLP 2006 - Lecture Notes in Artificial Intelligence 4272*, Springer Verlag, Germany, 43-54, 2006.
- [8] Swerts M., Strangert E., Heldner M., "F0 declination in spontaneous and read-aloud speech." *Proceedings of ICSLP*, Philadelphia, 3, 1501-1504, 1996.
- [9] E. J. Kutik, W. E. Cooper, and S. Boyce., "Declination of Fundamental Frequency in Speakers' Production of Parenthetical and Main Clauses." *J. Acoust. Soc. Am.*, 73(5): 1731-1738, 1983.
- [10] Streeter, L. A., "Acoustic determinants of phrase boundary perception." *J. Acoust. Soc. Am.*, 64(6): 1582-1592, 1978.
- [11] Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C., "The use of prosody in syntactic disambiguation." *J. Acoust. Soc. Am.* 90(6):2956-2970, 1991.
- [12] Grosjean, F., "How long is the sentence? Prediction and prosody in the on-line processing of language." *Linguistics*, 21: 501-529, 1983.
- [13] Swerts, M., "Prosodic features at discourse boundaries of different strength." *J. Acoust. Soc. Am.*, 101 (1):514-521, 1997.
- [14] Yang, Y. and Wang, B., "Acoustic correlates of hierarchical prosodic boundary in Mandarin." *Proceedings of the 1st International Conference of Speech Prosody (SP-2002)*. Aix-en-Provence, France, 707-710, 2002.
- [15] Fon, J., "Perception of discourse boundaries by Taiwan Mandarin Speakers." *Proceedings of the 2nd International Conference of Speech Prosody (SP-2004)*, Nara, Japan, 709-712, 2004.
- [16] Mannell, R.H. *The Perceptual and Auditory Implications of Parametric Scaling in Synthetic Speech*. Ph.D. dissertation. Macquarie University. 1994