UNIVERSITY OF CALIFORNIA

Los Angeles

Perception and Acoustic Correlates of the Taiwanese Tone Sandhi Group

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Chen-Hsiu Kuo

2013

ABSTRACT OF THE DISSERTATION

Perception and Acoustic Correlates of Taiwanese Tone Sandhi Group

by

Chen-Hsiu Kuo

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2013

Professor Patricia A. Keating, Chair

This dissertation investigates how the Taiwanese Tone Sandhi Groups are perceived, and the acoustic/phonetics correlates of listeners' judgments. A series of perception experiments have been conducted to scrutinize the following topics – Taiwanese tone neutralization, Tone Sandhi Group (TSG) as a prosodic domain, perceived boundary strength in Taiwanese and Taiwanese sentence disambiguation.

Taiwanese tone neutralization was examined with a checked tone recognition experiment and a corpus study. In the tone recognition gating experiment, the difference between the accuracy rate on citation and sandhi tones suggested that the two tones maybe be incompletely neutralized but as members of the same tonal category. The corpus study found that duration, f0 range and voice quality in domain-final position preserve distinctions between all the tone pairs. More specifically, the citation tones are longer in duration, wider in F0 range and creakier in voice quality.

The prosodic domain identification experiment revealed that participants were able to identify TSG as well as the other two prosodic levels, Word and Intonation Phrase even when the

speech signal is low-pass filtered (i.e. only the prosodic cues – duration, f0, and voice quality – were retained).

The perceived boundary strength rating experiment further showed that participants assigned different ratings for the three prosodic levels in spontaneous speech stimuli, and the acoustic analyses revealed that duration, f0 and voice quality cues in the last syllable in each stimulus were useful to the listeners. However, multiple linear regression indicated that these acoustic cues were far from sufficient to account for listeners' perception. In other words, to discriminate one prosodic domain from the others, listeners apparently need acoustic/prosodic cues not only from the syllables at the prosodic boundary, but also the syllables before the boundary within the same prosodic domain.

In the sentence disambiguation experiment, participants were able to interpret ambiguous sentences accurately. The results suggest that prosodic cues such as final lengthening and pitch reset, found at the disambiguation points, provide a strong basis for sentence identification.

The dissertation of Chen-Hsiu Kuo is approved

Sun-Ah Jun

Megha Sundara

Jody Kreiman

Patricia A. Keating, Committee Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

Chapter 1 Introduction …… 1

1.1 Taiwanese Tone Sandhi ……. 1

1.2 Prosodic Domains in Taiwanese ……..…….…… 5

1.3 Prosodic Boundary Cues …… 9

1.4 Research Questions …… 10

1.5 Overview …… 11

Chapter 2 Taiwanese Tone Neutralization ……..... 12

2.1 Introduction …….. 12

2.2 Checked tone identification …….. 15

2.2.1 Method …….. 16

2.2.1.1 Participants ……. 17

2.2.1.2 Stimuli …… 17

2.2.1.3 Procedures …….. 19

2.2.1.4 Data coding and collection ……. 20

2.2.2 Results …… 21

2.2.2.1 Surface Tone Identification Rate ……. 21

2.2.2.2 Surface Tone Identification Point (IDP) …….. 24

2.2.2.3 Surface Tone Recognition Point (RP) ……. 27

2.2.3 Summary ……. 28

2.3 Corpus study …….. 29

2.3.1 Method …… 30

2.3.2 Results …… 32

2.3.2.1 Syllable duration ……. 32

2.3.2.2 F0 range ……..... 35

2.3.2.3 H1*-H2* ……. 39

2.3.3 Discussion ……. 40

Chapter 3 Tone Sandhi Group as a Prosodic Domain ……. 42

3.1 Introduction …….. 42

3.2 Part I: Comparisons between Word Boundary and TSG Boundary and between Word Boundary and IP Boundary …… 43

3.2.1 Method ……43

3.2.1.1 Participants ……43

3.2.1.2 Stimuli …….. 43

3.2.1.2.1 Word boundary vs. Intonation Phrase boundary ……44

3.2.1.2.2 Word boundary vs. Tone Sandhi Group boundary …… 46

3.2.2 Procedures ……47

3.2.3 Analysis ……... 48

3.2.4 Result ……... 48

3.2.4.1 Accuracy ……... 48

3.2.4.1.1 Word boundary vs. IP boundary …… 50

3.2.4.1.2 Word boundary vs. TSG boundary …… 50

3.2.4.2 Sensitivity (d' score) ……50

3.2.4.3 Bias ……... 52

3.2.4.4 Interim summary …….. 52

3.3 Part II: Comparison between TSG boundary and from IP boundary ……. 53

3.3.1 Method ……53

3.3.1.1 Participants ……53

3.3.1.2 Stimuli …….. 53

3.3.1.3 Procedures ……54

3.3.2 Analysis ……... 54

3.3.3 Result ……... 55

3.3.3.1 Accuracy ……... 55

3.3.3.1.1 Normal stimuli ……... 57

3.3.3.1.1 Filtered stimuli ……...57

3.3.3.2 Sensitivity (d' score) ……... 57

3.3.3.3 Bias …….. 58

3.4 Summary …… 59

Chapter 3 Perceived Boundary Strength in Taiwanese……..…. 61

4.1 Introduction ……61

4.1.1 Background ……..61

4.1.2 Swedish Boundary Detection:

Carlson, Hirschberg, and Swerts (2005)…….…….. 62

4.2 Method …… 65

4.2.1 Stimuli ……..65

4.2.2 Subjects …… 67

4.2.3 Procedures ……67

4.2.4 Acoustic measures ……... 68

4.3 Results ……72

4.3.1 Responses ……72

4.3.1.1 English listeners listening to the Swedish stimuli ……74

4.3.1.2 Taiwanese listeners listening to the Swedish stimuli……76

4.3.1.3 English listeners listening to the Taiwanese stimuli ……77

4.3.1.4 Taiwanese listeners listened to the Taiwanese stimuli …… 78

4.3.2 Acoustic differences ……80

4.3.3 Multiple linear regression …… 82

4.3.3.1 English listeners listening to the Swedish and Taiwanese normal stimuli …82

4.3.3.2 English listeners listening to the Swedish and Taiwanese filtered stimuli… 83

4.3.3.3 Taiwanese listeners listening to the Swedish and Taiwanese normal stimuli…….. 84

4.3.3.4 Taiwanese listeners listening to the Swedish and Taiwanese filtered stimuli …… 85

4.4 Summary ……86

Chapter 5 Taiwanese Sentence Disambiguation ……..88

5.1 Introduction ……88

5.2 Literature Review ……89

5.2.1 Prosodic cues ……89

5.2.2 Gating experiments ……..90

5.3 Method …… 92

5.3.1 Participants …….. 92

5.3.2 Stimuli …… 93

5.3.3 Procedures ……94

5.3.4 Data analysis …… 95

5.3.4.1 Response coding ……95

5.3.4.2 Acoustic measures ……96

5.3.4.3 Statistical analysis ……96

5.4 Results ……97

5.4.1 Choice Score …… 97

5.4.2 Confidence ……..….. 101

5.5 Acoustic results compared to perception results …… 102

5.5.1 Final lengthening …… 103

5.5.2 Final pitch declination …… 104

5.5.3 Final glottalization / aperiodicity …… 107

5.6 Multiple linear regression between the scores and the acoustic measures ……. 109

Chapter 6 General discussion …… 111

6.1 Are Taiwanese sandhi tone and citation tone neutralized? ……. 111

6.2 Is TSG boundary perceivable and distinct from other prosodic boundaries? ……. 114

6.3 How do native speakers disambiguate sentences? ……..…. 115

6.4 Conclusion ……..……. 116

LIST OF FIGURES

1-1. Taiwanese tones spoken in isolation. From Peng (1997)…….. 2

1-2. Taiwanese Tone Sandhi rules, including the Tone Sandhi Circle for non-checked syllables (i) and for the checked syllables (ii). …….. 3

1-3. Syntactic configuration and prosodic phrasing of the sentence "The old lady doesn't believe that parrots can talk" (from Chen 1987). ……. 6

1-4. Taiwanese prosodic hierarchy……... 7

2-1. The gating sequence, illustrated by the syllable /p^hak^{31}/. …… 19

2-2. Surface tone identification rate. …… 23

2-3. Identification Point (msec) with separate graphs given for each variable that showed a main effect. ……..... 26

2-4. Surface tone identification point interactions …….. 27

2-5. Boxplot comparing the duration for seven different surface tones in Taiwanese… 32

2-6. Mean duration of the seven tones at (a) IP boundary (b) TSG boundary……. 35

2-7. Mean log F0 range of the seven tones at (a) IP boundary (b) TSG boundary…….. 38

2-8. Mean H1*-H2* of the seven tones at (a) IP boundary (b) TSG boundary……40

3-1. Double histograms of participants' responses in Identification tasks: (a) Word boundary vs. IP boundary (b) Word boundary vs. TSG boundary ……49

3-2. Double histograms of participants' responses in Part II:

(a) TSG boundary vs. IP boundary in normal speech

(b) TSG boundary vs. IP boundary in filtered speech ……. 56

4-1. Mean perceived upcoming boundary strength. ……... 64

4-2. A screenshot of what the listeners saw during the task. …….. 68

4-3. An example spectrum. From Shue (2010). …….. 72

4-4. English listeners' average log perceived boundary strength for Swedish stimuli……. 75

4-5. Taiwanese listeners' average log perceived boundary strength for Swedish stimuli……. 77

4-6. English listeners' average log perceived boundary strength for Taiwanese stimuli……. 78

4-7. Taiwanese listeners' average log perceived boundary strength for Taiwanese stimuli……. 80

5-1. From Grosjean and Hirst (1996). Take the sentence *Earlier my sister took a dip* for example. ……92

5-2. The choice scores averaged across listeners for one pair of test sentences…….. 98

5-3. Average choice score across Gates and Boundary Conditions. ……..….. 100

5-4. Average confidence ratings across the Positions and Boundary Conditions…… 102

5-5. Mean duration of the two Boundary Conditions at the 9 Positions…….. 104

5-6. Average values of the two F0 measures: (a) F0 mean (b) F0 median. …… 105

5-7. Average values of the two F0 measures: (a) F0 range (b) F0 slope. ……106

5-8. Average H1*-H2* across the two Boundary Conditions at the 9 Positions. ……107

5-9. Average HNRs across the two Boundary Conditions at the 9 Positions. …… 108

5-10. Average CPP across the two Boundary Conditions at the 9 Positions. …… 109

LIST OF TABLES

1-1. Lexical and sandhi tones with /si/. ……3

2-1. Schematization for different Positions and Environments with the example test syllable /pak^{53}/ “to tie”. 18

2-2. Average identification to the last-gate stimuli. ……22

2-3. Coefficients in the linear mixed-effect model predicting the surface tone identification point from three variables……..… 26

2-4. Taiwanese tones. …… 31

2-5. Mixed-design ANOVA results for duration for each tone. …… 34

2-6. Mixed-design ANOVA results for log F0 range for each tone. …… 37

2-7. Mixed-design ANOVA results for H1*-H2* for each tone. ……..... 39

3-1. Response Proportions in the first task: IP vs. Word. ……49

3-2. Response Proportions in the second task: TSG vs. Word. ……...… 51

3-3. Response Proportions: IP vs. TSG in normal speech. ……..58

3-4. Response Proportions: IP vs. TSG in filtered speech. ……..58

4-1. Repeated measures ANOVA table for English listeners’ strength ratings Swedish stimuli. …… 75

4-2. Repeated measures ANOVA table for Taiwanese listeners’ strength ratings for Swedish stimuli. …… 76

4-3. Repeated measures ANOVA table for English listeners’ strength ratings for Swedish stimuli. …… 78

4-4. Repeated measures ANOVA table for Taiwanese listeners’ strength ratings for Taiwanese stimuli. …… 80

4-5. ANOVA results for the acoustic measures of the last syllable in the stimuli in normal speech. …….. 81

4-6. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese normal stimuli ……..83

4-7. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese filtered stimuli ……. 84

4-8. Results of multiple regression analysis for Taiwanese listeners listening to the Swedish and Taiwanese normal stimuli ……. 85

4-9. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese filtered stimuli ……. 86

5-1. Repeated measures ANOVA table for Choice Score. …… 99

5-2. Average choice scores and standard deviations. ……. 100

5-3. Repeated measures ANOVA table for Confidence ……. 101

5-4. Results of multiple regression analysis for choice scores vs. acoustic measures and for confidence vs. acoustic measures……. 110

6-1. Summary of the experiments with the Taiwanese listeners. …… 114

ACKNOWLEDGEMENTS

My deepest gratitude goes to my advisor, Pat Keating, who has always been so caring and supportive. This dissertation would have been nearly impossible without her guidance and pearls of wisdom. My gratitude to her is beyond words. Pat, I treasure you dearly with much respect and love.

I would like to thank my committee, Sun-Ah Jun, Megha Sundara and Jody Kreiman for their valuable time, sharp-witted comments, and continuous encouragements. I am also grateful for my teachers in the Linguistics Department, all of whom inspired me in different ways: Bruce Hayes, Robert Daland, Kie Zuraw, Nina Hyams and Susie Curtiss.

Many thanks are due to Henry Tehrani for his invaluable support with recordings and coding. Thanks also go to Professors Ho-hsien Pan and Sin-Horng Chen for insightful discussions and generous support. I am also thankful to my undergraduate research assistants: Irene Chou, Hannah Chu and Spencer Lin for their time and efforts with the data processing.

Special thanks go to my wonderful friends in the P-Lab and in the Department: Kristine Yu, Jianjing Kuang, Craig Sailor, Byron Ahn, Jason Bishop, Marc Garellek, Jamie White, Yun Kim, Chad Vicenik, Adam Chong, Victoria Thatte, Yu Tanaka, Tomoko Ishizuka and Melanie Lynn.

Many friends have helped me through these years – Yen-Jung Chang, Shao-Yi Chen, Jerry Weng, Fang-Ying Hsieh, Fei Wu, Callie Chen, Stasia Su, Frank Wang, Chris Hsieh and many others.

Last but certainly not the least, thank you to Father God, my dearest parents and grandmothers for their boundless love and support over these years, and to my amazing one-of-a-kind sister Chen-Ling Kuo for always being sweet, warm and loving.

VITA

1981	Born, Tainan, Taiwan
2003	B.A., Foreign Languages and Literatures National Taiwan University Taipei, Taiwan
2006	M.A., Linguistics National Tsing Hua University Hsinchu, Taiwan
2010	M.A., Linguistics University of California, Los Angeles
2009 – 2011	Research Assistant Department of Linguistics University of California, Los Angeles
2008 – 2013	Teaching Assistant Department of Linguistics University of California, Los Angeles

PUBLICATIONS

Kuo, G. 2013. Perceived prosodic boundaries in Taiwanese and Swedish. *Proceedings of Meetings on Acoustics (POMA),* Vol. 19, pp. 060228.

Kuo, G. and Vicenik, C. 2012. The Intonation of Tongan. *UCLA Working Papers in Phonetics*, 111, 63-91.

Keating, P. and Kuo, G. 2012. Comparison of speaking fundamental frequency in English and Mandarin. *Journal of the Acoustical Society of America* 132 (2): 1050-1060.

Kuo, G. and Yu, K. M. 2012. Mandarin Quantifiers. Chapter 12 in *Handbook of Quantifiers in Natural Language*, ed. Edward Keenan and Denis Paperno, Springer Press.

Kuo, G. 2012. Perceived prosodic boundaries in Taiwanese and their acoustic correlates. *InterSpeech 2012 Proceedings.*

Kuo, G. 2011. Syllable contraction in Taiwan Mandarin. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS-17).*

Kuo, G. 2011. Prosodic boundaries and the Taiwanese tone sandhi domain. *UCLA Working*

Papers in Phonetics, 109, 40-59.

Kuo, G. 2010. Production and perception of Taiwan Mandarin syllable contraction. *UCLA Working Papers in Phonetics*, 108, 1-34.

Keating, P. and Kuo, G. 2010. Comparison of speaking fundamental frequency in English and Mandarin. *UCLA Working Papers in Phonetics*, 108, 164-187.

Chapter 1: Introduction

1.1 Taiwanese Tone Sandhi

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. Take (1) for instance. (1a) shows the lexical tone for each syllable, and (1b) shows the surface tone for each syllable. Tone Sandhi Group boundaries are denoted by #, and the underlined tones in (1b) are sandhi tones. As can be seen, there are two Tone Sandhi Groups in (1). The syllables at the Tone Sandhi Group boundaries, *gi51* and *tiau33*, retain their lexical tones, whereas all the other syllables carry a new tone, i.e. sandhi tone.

(1) " I enjoy studying Taiwanese tone sandhi."

a. *Lexical:* goa51 chin55-ai31 gian24-kiu31 tai24-gi51 # e24 pian31-tiau33 #

b. *Surface:* goa55 chin33-ai51 gian33-kiu51 tai33-gi51 # e33 pian51-tiau33 #

c. *Gloss:* I enjoy study Taiwanese GEN tone sandhi

There had been extensive studies on Taiwanese tone sandhi (also known as Xiamen tone sandhi) (Wang 1967; Cheng 1968; Yip 1980; Wright 1983; Chen 1987; Du 1988; King 1988; Hsiao 1991; Tsay 1994; Lin 1994, and among others). A Tone Sandhi Group is not confined to a phonological word or a phonological phrase; its size could vary from a disyllabic phrase to a full sentence with multiple syntactic phrases. In (1), the first Tone Sandhi Group consists of "I – enjoy – study – Taiwanese", which could be an independent sentence, meaning "I enjoy studying

Taiwanese". The addition of the second Tone Sandhi Group, which contains three syllables, changes the object of the sentence from "Taiwanese" to "Taiwanese tone sandhi".

Figure 1-1 provides the F0 contours of the seven Taiwanese tones produced in isolation. Two of these appear only in checked syllables (= CVC, where the coda can only be a voiceless stop /p, t, k, ʔ/), hence are called checked tones. Checked tones are relatively short and the tonal values are traditionally described as *53* and *31* as shown in the figure. Other than the checked tones, there are two level tones (high level 55 and mid level 33), two falling tones (high falling 51 and low falling 31) and one rising tone (24). These non-checked tones appear in CV(C) syllables where the coda is either a nasal /m, n, ŋ/ or a glide /j, w/.

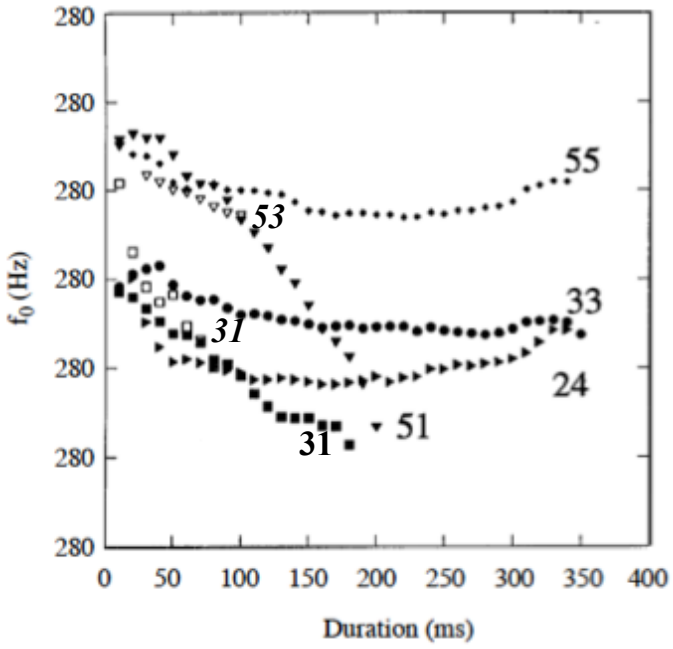


Figure 1-1. Taiwanese tones spoken in isolation by a native female speaker. *53* and *31* are two checked tones (which appear in CVC syllable). Figure revised from Peng (1997, p. 374.

The Taiwanese tone sandhi system is unusual – not only do all tones participate in sandhi, but also sandhi is structure-preserving – each lexical tone's sandhi form is another lexical tone. The

selection of a sandhi tone is subject to a Tone Sandhi Circle. Figure 1-2 illustrates the Tone Sandhi Circle and the direction of the arrow shows the selection of the surface sandhi form. For instance, the sandhi form of a lexical high level tone (55) is a mid level tone (33), the sandhi form of a lexical mid level tone (33) is a low falling tone (31), and so on.

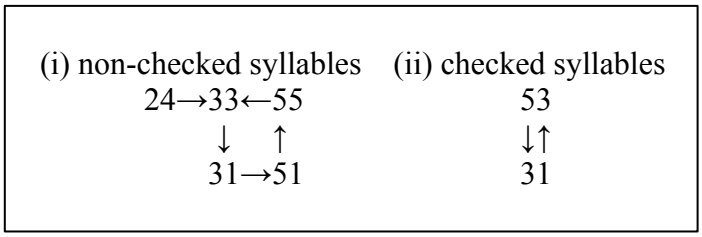


Figure 1-2. Taiwanese Tone Sandhi rules, including the Tone Sandhi Circle for non-checked syllables (i) and for the checked syllables (ii).

Table 1-1 demonstrates the lexical tones and their corresponding sandhi tones with the syllable /si/. The word "poetry" has a high level (55) lexical tone (i.e. underlying tone). When it appears in a non-final position in a phrase (i.e. sandhi position), it surfaces as a mid level tone (33), as shown in the phrase "psalm".

Table 1-1. Lexical and sandhi tones with /si/.

Lexical tone	Sandhi tone (=surface tone)
si 55 "poetry"	***si33****-pian55* "psalm"
si 51 "to die"	***si55****-lang24* "dead person"
si 31 "four"	***si51****-tiam51* "four o'clock"
si 24 "time"	***si33****-kan55* "time span; time"
si 33 "temple"	***si31****-ĩ33* "monastery"
sit31 "to lose"	***sit53****-bong33* "disappointed"
sit53 "real"	***sit31****-but53* "food"

Because the tone changes never create new tones in Taiwanese, the f0 contour of the mid level surface form in sandhi position is similar to a lexical mid level tone in a citation position. The mid level surface forms of the two sources are neutralizing. Therefore, each surface tone is

potentially ambiguous in Taiwanese. For example, the surface tone [si33] might be the lexical form of "temple", or the sandhi form of "poetry" or "time".

Furthermore, words and sentences can also be ambiguous because of the ambiguity in surface tones. For instance, (2a) is a sentence with two readings. The tones transcribed in (2a) are surface tones. The parentheses in (2b) and (2c) denote Tone Sandhi Group domains. Because sandhi tones and lexical tones share the same tone forms, (2a) is ambiguous for multiple available parses, as shown in (2b) and (2c). In (2b), *tiuN33* is in domain final position, therefore the surface *33* is a lexical tone. In (2c), *tiuN33* is not in domain final position, therefore, the surface *33* is the sandhi form of *tiuN24*. Therefore, the ambiguity in (2a) can be resolved by the prosodic grouping, as shown in (2b) and (2c).

(2) *a. Surface:* i33 tiuN33 bin33 koh53 khah53 bai51 ma31 m33 kiaN55

b. Lexical: (σ55 σ33)(σ33)(σ σ σ)(σ σ σ)

uncle face more ugly even not afraid

"Uncle is not afraid of his face becoming more and more ugly."

c. Lexical: (σ55 σ24 σ33)(σ σ σ)(σ σ σ)

s/he scene more bad even not afraid

"S/he is not afraid of the scene becoming any worse."

Since there are multiple prosodic groupings for a sentence, a sentence could be massively ambiguous considering all the possible prosodic groupings. However, Taiwanese listeners are easily able to understand words and sentences. Then questions regarding the prosodic groupings are brought up:

(i) How do listeners identify the tones, and process Taiwanese utterances, given that all utterances could be potentially ambiguous? How do people "disambiguate" unambiguous sentences?

(ii) How do listeners process ambiguous utterances where top-down knowledge is not present? If they are able to do so, what acoustic cues are available to them?

There are at least two possibilities – (a) Perhaps tones are not actually completely *neutralized*. Listeners can distinguish, say lexical 33 from sandhi 33. (b) Listeners can distinguish citation *positions* from sandhi *positions*. If they could recognize the position a syllable is in, they would know what kind of tone it must be. In this dissertation, both possibilities will be tested.

Regarding ***tone neutralization***, some previous studies have shown that duration and f0 measures were not completely identical between lexical tones and the corresponding sandhi tones in Taiwanese. These studies are reviewed in Chapter 2.

For listeners to be able to recognize positions from acoustic cues, these positions would have to be ***prosodic positions*** (i.e. positions in prosodic domains). Section 1.2 and section 1.3 provides a brief review of the studies on Taiwanese prosodic and prosodic boundary cues. The prosodic boundary cues will be examined in the experiments in alter chapters. Chapter 3 and 4 scrutinize the detection of prosodic positions in read speech and spontaneous speech. The prosodic positions would be signaled with prosodic correlates, namely, the prosodic boundary cues.

1.2 Prosodic Domains in Taiwanese

When it comes to prosodic boundaries in Taiwanese, we first need to know what the Taiwanese prosodic domains are. It is known that prosodic grouping often coincides with

syntactic phrasing even though there is not a one-to-one correspondence between prosodic and syntactic boundaries. The connection is stronger in Taiwanese because the prosodic domain where tone sandhi applies is mostly determined by the syntactic structure. Nonetheless, the example in Figure 1-3 from Chen (1987) demonstrates 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 phrase boundaries. The upper configuration with syntactic Inflectional Phrases ("IP" in the figure) shows how the sentence is syntactically parsed, and the lower configuration with Tone Sandhi Groups ("TG" in the figure) shows how the sentence is prosodically parsed. The "Lexical tone" transcribes the lexical tone of each syllable, and the "Surface tone" transcribes their surface realizations after sandhi.

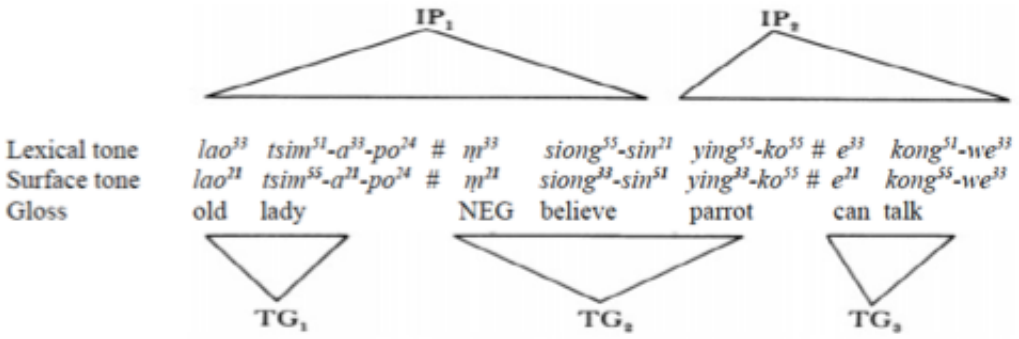


Figure 1-3. Syntactic configuration and prosodic phrasing of the sentence "The old lady doesn't believe that parrots can talk" (from Chen 1987). IP in this figure stands for the syntactic Inflectional Phrase, and TG in this figure refers to Tone Sandhi Group. # denotes the Tone Sandhi Group boundary.

Taiwanese speech can be organized into a hierarchy of three major prosodic levels – Intonation Phrase ("IP"[1], hereafter), Tone Sandhi Group ("TSG", hereafter) and Word/Syllable. Peng and Beckman (2003) proposed a ToBI framework with a similar break indexes – b4 for IP boundary, b3 for TSG boundary, and b2 for Word/Syllable boundary[2]. This dissertation

[1] The IP here is not the same as the IP (Inflectional Phrase) in Figure 1-3.
[2] The index b1 is for resyllabification (e.g. /hit31 e/ → [he53 le], which is not our concern here.

investigates the relationship between the perception and the acoustic analysis of Taiwanese prosodic domains based on this assumption.

The prosodic structure of Taiwanese is schematically represented in Figure 1-4. The IP is the largest prosodic level and can cover an entire sentence. IP immediately dominates the TSG, which is exhaustively contained into the IP. The number of TSGs contained in an IP may vary, and TSGs never straddle the IP boundaries. Word/Syllable is the lowest prosodic level and each has its own lexical tone. Therefore, in Figure 1-4, σ_3 and σ_5 are the last syllables of two TSGs. Since σ_5 is also the last syllable of IP, it is a syllable at the IP boundary. σ_3 is a syllable at the TSG boundary. The end of other syllables (i.e. σ_1, σ_2 and σ_4) marks a word boundary.

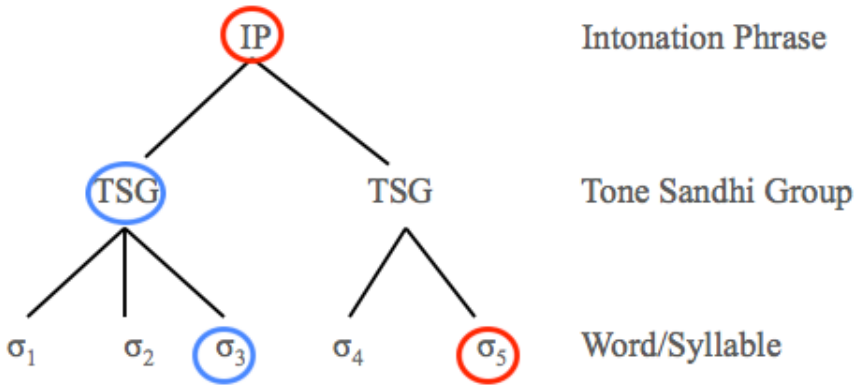


Figure 1-4. Taiwanese prosodic hierarchy

Nonetheless, there has been debate over whether or not the TSG should be incorporated into the prosodic hierarchy. The TSG mostly corresponds to the phonological phrase (PP); however, it can also correspond to a prosodic word (PrWord) or an Intonation Phrase (IP). Previous research analyzing a variety of articulatory and acoustic measures suggests that Taiwanese Tone Sandhi Group could be an independent prosodic domain and that speakers make distinctions between Tone Sandhi Groups and other prosodic domains during sentence speech production.

The measures included VOT and closure duration (Hsu and Jun 1998, though they did not directly incorporate TSG into the hierarchy); linguopalatal contact and articulatory seal duration (Keating, Cho, Fougeron and Hsu 2003; they named TSG as the "Small Phrase"), and degree of nasalization (Pan 2003). 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 can be identified perceptually, it would suggest that listeners not only are able to discriminate between prosodic domains, but also to exploit the relevant acoustic cues to process sentences correctly.

Pre-boundary lengthening and F0 declination are two useful acoustic measures that mark prosodic boundaries across languages, including Taiwanese. Previous studies on the perception of Taiwanese tones could be considered indirect studies on Taiwanese boundaries cues. That is because researchers were interested in the difference between Taiwanese lexical and sandhi tones, and the occurrence of the lexical the sandhi tones was closely related to the position where they were located – whether it was the tone of the last syllable of a Tone Sandhi Group (nearest to the boundary) or other non-final syllable within the sandhi domain. For example, Peng (1997) found F0 final lowering and final lengthening in a domain-final position (i.e. where lexical tones occur) as opposed to domain-initial and domain-medial positions (i.e. where sandhi tones occur). Pan (2006) found that smoother F0 slope was found in a domain-final position (i.e. lexical tone) than in a domain non-final position (i.e. sandhi tone).

1.3 Prosodic Boundary Cues

Ample experimental evidence has shown that both speakers and listeners use boundary cues to suggest or to locate possible prosodic boundaries. A reliable cue indicating clausal disjuncture is *the silent pause* (Lehiste 1979; Groz and Hirschberg 1992; Hirschberg and Nakatani 1996). However, it is not a necessary nor a sufficient cue in that it is not always observed at a prosodic boundary. Yoon et al. (2007) found that in the BU radio speech corpus, phrasal boundaries are signaled by the presence of a silent pause only about 40% of the time, whereas the remaining 60% of the boundaries occur with no silent pause.

In addition to the silent pause, the other major temporal cue that listeners use to locate the prosodic boundaries is *final lengthening*, with greater effects of lengthening at successively higher levels of prosodic domains (Lehiste 1973; Lehiste, Olive and Streeter 1976; Nooteboom, Scott 1982; Wightman, Shattuck-Hufnagel, Ostendorf and Price 1992; among others). The final rime in a prosodic domain is lengthened, which reflects a decrease of the articulation rate at the end of the phrase. Moreover, *speech rate* measured from durations is a similar indicator for boundaries, since it is negatively correlated with the boundary size (Hirschberg and Nakatani 1996).

Pre-boundary *pitch declination* and post-boundary *pitch reset* also characterize a prosodic domain (Cooper and Sorensen 1977; Streeter 1978; Beach 1991; Wightman et al. 1992; Pijper and Sanderman 1994). Final lengthening and the declining pitch contour are the most reliable cues for segmenting a continuous speech signal into prosodic constituents cross-linguistically.

Intensity drop has been found to serve as a possible cue for a prosodic boundary, even though it is less discussed. Hirschberg and Nakatani (1996) found that the RMS amplitude in a domain-final position is significantly smaller than in a domain-initial position.

Laryngealization has also been proposed as an indicator of a prosodic boundary (Lehiste 1979). Kreiman (1982) found that laryngealization sometimes appears at a sentential boundary, and listeners would posit a structural boundary if there is a change in voice quality from modal to creak.

1.4 Research Questions

There are several questions that we would like to address in this dissertation:

First of all, do lexical tones and sandhi tones sharing the same surface tone differ in terms of perception? In addition, given that a checked tone is 1/3 shorter than a non-checked tone (Tseng 1995), how well do listeners hear the difference between the lexical and sandhi checked tones sharing the same surface tone? In an attempt to provide the answer to this research question, a gating experiment with checked tones will be presented in the first part of Chapter 2.

Secondly, how do lexical tones and sandhi tones differ in terms of acoustic cues? Previous studies have shown that duration and F0 are salient boundary cues, therefore, the second part of Chapter 2 will present a corpus study with regards to the above acoustic measures at the two syllables in a disyllabic phrase – the first syllable is in domain non-final position so that it possesses a sandhi tone, whereas the second syllable is in domain-final position so that it possesses a lexical tone.

Thirdly, how do lexical tones and sandhi tones differ in longer fragments? Identification tasks will be presented in Chapter 3 to examine how lexical and sandhi tones are perceived in longer fragments. With utterances as short as 4 syllables, are listeners able to tell the size of the boundary that comes after (Word boundary vs. TSG boundary; Word boundary vs. IP boundary; TSG boundary vs. IP boundary)?

Fourthly, are people able to predict the size of the upcoming boundaries in spontaneous speech? It is known that spontaneous speech and read speech differ in terms of some acoustic properties (Blaauw, 1994), and we do not know if listeners can accurately predict the upcoming boundaries in spontaneous speech (though presumably the answer is yes). How much information do listeners need to make such judgments? Chapter 4 will present a boundary strength rating task with both native and nonnative speakers of Taiwanese, with two versions of the stimuli, normal and low-pass filtered.

Lastly, how do people disambiguate real ambiguous sentences? As we have mentioned earlier, every sentence is a potentially ambiguous sentence in Taiwanese. However, listeners seem to know how to disambiguate these unambiguous sentences. Therefore, in Chapter 5, an experiment with real ambiguous sentences will be presented.

1.5 Overview

This dissertation is divided into six chapters. Chapter 1 has introduced the dissertation with background research, research questions and the overview. Chapter 2 investigates Taiwanese tone neutralization in a lexical gating experiment and a corpus study. Chapter 3 examines the Taiwanese Tone Sandhi Group as an independent prosodic unit with an identification experiment. Chapter 4 investigates the detection of upcoming boundaries in Taiwanese and Swedish spontaneous speech with a boundary strength rating experiment in order to see if listeners discriminate the three prosodic boundaries – Word boundary, TSG boundary and IP boundary. Chapter 5 focuses on how listeners use the given acoustic cues to disambiguate Taiwanese sentences. Finally, Chapter 6 provides general discussion about the nature of the Taiwanese Tone Sandhi Group domain and the implication of the findings.

Chapter 2: Taiwanese Tone Neutralization

2.1 Introduction

This chapter presents an experimental study and a corpus study, which together explored whether lexical tones and sandhi tones differ perceptually and acoustically. In other words, this chapter examines the possible neutralization of Taiwanese tones that alternate by tone sandhi. In the experimental study, listeners were asked to identify two tones in checked syllables (i.e. syllables which end with a voiceless stop consonant) extracted from different prosodic positions (sandhi position vs. citation position) and different environments (in isolation vs. embedded in a sentence). In the corpus study, all but one[3] of the tone sandhi pairs of Taiwanese were compared acoustically to test for neutralizations.

Taiwanese tones are acoustically influenced by prosodic position. Peng (1997) examined the five Taiwanese tones in non-checked syllables placed in four different prosodic positions: phrase initial, phrase-medial, phrase-final and utterance-final. The 'phrase' and 'utterance' in her study are the two prosodic levels – TSG and IP in this dissertation. The tones in the latter two prosodic positions were assumed to retain their lexical tone ("citation tone", hereafter), whereas the tone in the former two prosodic positions had the sandhi forms as their output. In terms of duration, lengthening was found only at domain-final position (i.e. phrase-final and utterance-final). F0 also changed according to prosodic position: the F0 values were significantly lower in utterance-final position than in other prosodic positions; the F0 values were mostly the same in the two non-final positions. Peng's study thus showed that prosodic position substantially affects the duration and F0 values of Taiwanese tones in non-checked syllables; the tones in prosodically final positions tended to show final lengthening and final lowering.

[3] There is no Rising sandhi surface tone in Taiwanese.

Myers and Tsay had a series of production studies (1999, 2001 and 2008) on the issue of tone neutralization in Taiwanese. Their most recent paper compared the surface forms *kim33* in well-designed sentence pairs such as (1) and (2), where the tone in ***kim33*** is a lexical tone or a sandhi tone derived from *kim55*, respectively. The sentence pairs are syntactically ambiguous and the difference is revealed only when they are spoken. In (1), the aunt and the sister-in-law are different individuals and the words are separated by a TSG boundary, whereas in (2), *Akim* is the name of the sister-in-law, and is not at a TSG boundary.

(1) "These are aunt, sister-in-law and elder brother."

a. lexical: *che51 si33 a55-kim33 # che51-a0*[4] *# kap31 a55-hiaN55 #*

this is aunt sister-in-law and elder brother

b. surface: *che55 si31 a33-**kim33** # che51-a0 # kap53 a33-hiaN55 #*

(2) "These are sister-in-law Akim and elder brother."

a. lexical: *che51 si33 a55-kim55 che51-a0 # kap31 a55-hiaN55 #*

this is Akim sister-in-law and elder brother

b. surface: *che55 si31 a33-**kim33** che51-a0 # kap53 a33-hiaN55 #*

Myers and Tsay analyzed the surface output ***kim33*** in both (1b) and (2b). They found that in terms of F0, speakers did neutralize lexical and sandhi tones. However, lexical and sandhi tones were distinguished by syllable duration, with a strong effect of phrase-final lengthening: because lexical tones occur before a boundary, their syllables were lengthened. The authors argue that the observed duration effect is consistent with gradient incomplete neutralization of the lexical tone and the sandhi tone. That is, tone sandhi is neutralizing, but not categorically so. However, this

[4] Utterance final particle such as "a0" has a neutral tone and does not participate in tone sandhi. Therefore, *che51* in (1) and (2) is the last syllable within the Tone Sandhi Group which keeps its located in the citation position, hence does not undergo tone sandhi.

incompleteness results not from a failure to neutralize the tones per se, but rather from a phrasal cue to a prosodic position.

On the other hand, Myers and Tsay did find one case of incomplete neutralization of two tones, which had nothing to do with the prosodic position. In the Taiwanese tone sandhi circle, an underlying rising tone (24) and an underlying high level tone (55) have the same sandhi tone outcome – a mid level tone (33). Myers and Tsay compared sandhi tone 33 derived from 55 and from 24. This is the only case where two underlyingly different tones might be neutralized into the same surface tone in sandhi context. Their finding was that speakers produced overall higher F0 for sandhi tone derived from 55 than for sandhi tones derived from 24, and the tones derived from 55 were eight msec longer on average than the tones derived from 24. This result is interesting because the prosodic position of the two sandhi tones is the same, and thus any differences between them must be due to neutralization rather than phrasal prosody. The question then is again whether such small differences are perceivable.

It thus seems that there are at least two potential kinds of information that Taiwanese listeners might use in coping with tone sandhi. First, sandhi tones might be phonetically distinct from citation tones, e.g. in pitch. If a small phonetic difference is perceptible by listeners, such physical differences could be used to recover the lexical identity of a phonetic tone, and there would be no mystery at all in how Taiwanese listeners recognize words in the face of tone sandhi. In Section 2.2, I will present a perception study on Taiwanese checked tone recognition. Given that checked tones are relatively short and abrupt, it is worthwhile to see if listeners are able to recognize these short tones selected from different prosodic positions. Furthermore, tones being selected from different prosodic positions means that they are at different prosodic boundaries, which leads us to the second kind of information.

The second kind of information would be less direct, concerning the detection of prosodic boundaries. Since sandhi tones occur within the Tone Sandhi Group domain and citation tones occur before all kinds of phrasal boundaries, then information about phrasal boundaries is information about surface tone forms. For example, the tone before a full pause must be a citation form. Once a listener hears a pause, then it becomes clear that, say, a phonetic mid level tone right before the pause should be a lexical tone 33, and not the sandhi form of a lexical 55 or of a lexical 24. This information does not become clear until after the entire syllable is heard. In other words, if this were the only available boundary information, then recognition would be rather delayed. However, other prosodic information might provide cues to *upcoming* phrasal boundaries. If a boundary can be anticipated, then tonal recognition need not be delayed. Suppose, for example, that phrases always end with a drop in intensity, and that such drops never occur phrase-internally. In that case, an intensity drop would signal the end of a phrase, which is the position for a citation tone form. If a listener hears a phonetic tone 33 at the same time as the intensity drops, then the tone must be a lexical mid tone, and not the sandhi form of a lexical 55 or 24. Such prosodic boundary cues might be much easier to perceive than small F0 differences between incompletely-neutralized tone forms. In section 2.3, I will present a corpus-based study on the acoustic differences (duration, F0 and laryngealization in particular) between the sandhi tones and the citation tones in disyllabic phrases from different prosodic boundaries (TSG vs. IP).

2.2 Checked tone identification

Previous phonetic studies on Taiwanese tone sandhi mainly focused on studying tones in non-checked syllables (i.e. CV syllables) because (a) checked tones are each other's sandhi form and

do not participate the Tone Sandhi Circle, and (b) checked tones are relatively short, as shown in Figure 1-1 and suggested by Tseng (1995) where she found that checked tones are 1/3 shorter than non-checked tones in general. The current study therefore puts focus on the identification of the checked tones.

This experiment was designed to examine the two Taiwanese tones that occur in checked syllables ("checked tones", hereafter) – the high falling 53 and the low falling 31. These checked tones were on syllables taken from different prosodic positions (sandhi position vs. citation position) and different environments (in isolation vs. embedded in a carrier sentence).

2.2.1 Method

The gating paradigm is mostly used in spoken word recognition research. In the task, listeners hear increasingly longer fragments of a word, either in a sentence or on its own. The listeners are typically asked to propose (either a forced-choice or an open-choice response) the word being presented, and to give a confidence rating after each fragment. The two dependent variables in a typical gating experiment are the *identification point* and the *recognition point*. The identification point refers to the length of the word that the listeners need to hear in order to make the correct decision without further changes. The recognition point is the point in the word where the listeners not only make the correct choice without further changes but also give an 80% or higher confidence rating to their answer.

In this experiment, the gating paradigm is used in tone recognition. The listeners hear words with different tones in increments and they are asked to identify the tones they hear as quickly and accurately as possible, and to give a confidence rating.

2.2.1.1 Participants

Ten adult native Taiwanese speakers from the southern part of Taiwan participated as listeners in this experiment. None of them reported any hearing or language problem. They were compensated for their participation with ten dollars.

2.2.1.2 Stimuli

The stimuli contained eight real-word monosyllables, which are /pak53/ "to tie" and /pak31/ "belly", /kap53/ "to complain" and /kap31/ "to connect", /bat53/ "tight" and /bat31/ "to know", and /p^hak53/ "to expose under the sun" and /p^hak31/ "to lean over". As we can see, these eight syllables comprised four tonal minimal pairs, and the tones of interest are the low falling and the high falling checked tones. The onsets of all the stimuli were obstruents in order to eliminate any effect from tonal pitch patterns that begin during an initial sonorant.

A trained female Taiwanese speaker from the southern part of Taiwan, the author, recorded the eight syllables in the sound-treated booth in the UCLA Phonetics Laboratory at a sampling rate of 44 kHz. The eight syllables appeared in two positions, the first syllable of a disyllabic phrase ("sandhi position" hereafter) and the second syllable of a disyllabic phrase ("citation position" hereafter). The other syllable in each disyllabic phrase was controlled as carrying a mid level tone. The disyllabic phrases were recorded in two different environments – in isolation vs. embedded in the carrier sentence shown in (3). The transcribed tones in (3) are the surface tones. The pound signs (#) indicate the TSG boundaries, the underscores indicate syllables bearing sandhi tones, and ($\underline{\sigma}_1$ σ_2) shows where the disyllabic phrases go. Note that if, say, /pak53/ appears at σ_2, which is the citation position, then the surface tone is still high falling. If /pak53/ appears at $\underline{\sigma}_1$, which is the sandhi position, the surface tone will be low falling.

(3) *i33 tiaN33* (σ$_1$ σ$_2$) # *e33 siaN33-tiau33* #
s/he listen to POSS tone
'He listened to the tone of ($\sigma_1\sigma_2$).

In total, there were 32 tested syllables: 4 Syllables × 2 Surface Tones (high falling vs. low falling) × 2 Positions (citation vs. sandhi) × 2 Environments (in isolation vs. embedded in a sentence). Each syllable with one of the two checked tones was either in a citation position or a sandhi position, and it was either from the disyllabic phrase in the carrier sentence or was from the disyllabic phrase in isolation. Table 2-1 shows the Position × Environment conditions for each tone, with /pak^{53}/ "to tie" as the example.

Table 2-1. Schematization for different Positions and Environments with the example test syllable /pak^{53}/ "to tie". The tones transcribed here are the surface tones. When /pak53/ appears in the sandhi position, it has [pak^{31}] as the surface form. Listeners heard only the test syllable.

	Isolation environment	Sentence environment
Citation position	σ$_1$ **σ$_2$** bo33 **pak53** "to not tie"	σ σ σ$_1$ **σ$_2$** # σ σ σ i33 tiaN33 bo33 **pak53** # e33 siaN33-tiau33 "He listened to the tones of 'bo33 pak53'."
Sandhi position	**σ$_1$** σ$_2$ **pak31** u33 "had something tied"	σ σ **σ$_1$** σ$_2$ # σ σ σ i33 tiaN33 **pak31** u33 # e33 siaN33-tiau33 "He listened to the tones of 'pak31 u33'."

The recording was then transferred to Praat for viewing and editing. A Praat script was run to extract the gates from each syllable. The initial consonant of each syllable (i.e. the stop burst plus any aspiration, up to the onset of voicing) served as the first gate, and the following gates were formed in 5 msec increments, which is about one pitch pulse for this speaker. Figure 2-1 illustrates the first eight gates in the gating sequence for syllable /p^{h}ak^{31}/. There were 1108 gates in total in this experiment.

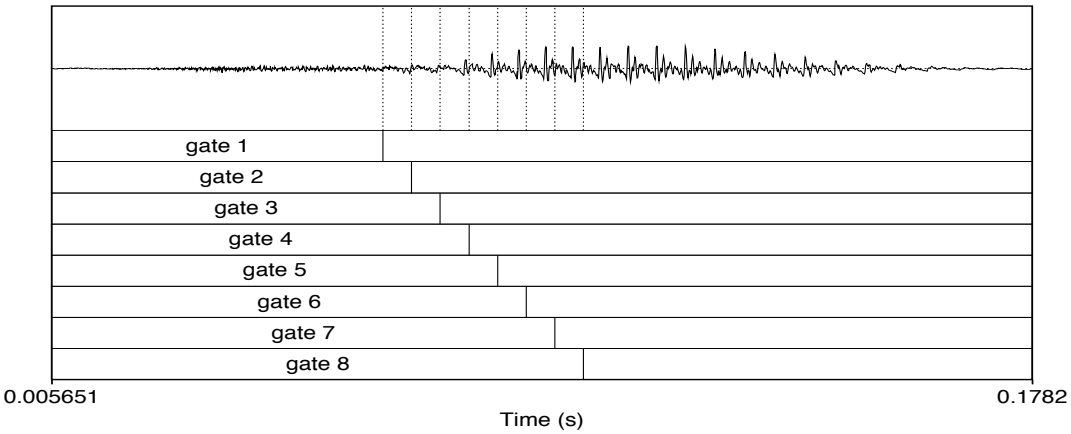


Figure 2-1. The gating sequence, illustrated by the syllable /p^hak^{31}/. The first gate included only the initial consonant, and the following gates were constructed in 5 msec increments. Only eight gates are shown here, but the increments continued through the entire syllable, such that the last gate comprised the whole.

2.2.1.3 Procedures

The subjects were tested individually in a quiet room: the experiment was presented on a laptop using Matlab and PsychToolbox. First, the subjects were asked to identify the tones of two characters on the screen (/kut31/ "bone" and /kut53/ "slippery"). Neither of these words was a test word in the experiment. They all agreed that the character of "bone" has a low falling tone and the character of "slippery" has a high falling tone. It is natural for native speakers to pronounce a stand-alone word with its lexical tone. The subjects were then told that their task was to identify the tone, either low falling or high falling, for each stimulus they would hear. They were also informed that all the stimuli came from real words, even though most of the time they would only hear parts of the words. They were explicitly told that the two characters on the screen were to be used as the signs for "low falling tone" and "high falling tone" respectively. Thus, in effect the listeners were implicitly encouraged to hear the tones of the stimuli as if they were all citation (isolation) tones. They then moved on to the test trials.

In each trial, they clicked on a speaker sign on the screen in order to hear the stimulus, and then they chose the tone, which was most likely for the presented auditory stimulus by a forced choice between the two Chinese characters on the screen. They then provided a confidence rating on a scale of 1 to 7 for each trial, and finally they pressed the space key in order to move on to the next trial.

2.2.1.4 Data coding and collection

Listeners' responses were coded into three variables – the surface tone identification rate (ID rate), the surface tone identification point (IDP), and the surface tone recognition point (RP).

The surface tone identification rate (ID rate) refers to the percentage of responses for the last gate (i.e. the full word), which matched the surface tone of the stimulus (whether lexical or sandhi) to the character with the same surface (here, its lexical) tone. For example, if listeners responded to [pak31] as if it matched the character /kut31/ "bone", it was coded as a 1, regardless of whether [pak31] came from /pak31/ or /pak53/. This will be referred to as the "match" response. The opposite response was coded as a 0; this is the "non-match" response. The response was coded with the assumption that sandhi tone and citation tone are neutralized so that listeners were expected to answer with, say, a high falling tone when they were presented with a lexical high falling as well as a sandhi high falling tone. The prediction is that the rate for the tone from citation position is closer to 1 because it is the lexical tone with which listeners are not expected to get confused. If the rate for the tone from sandhi position is closer to 1, it suggests a more complete neutralization from a listener's perspective; a sandhi high falling tone is perceived the same as a lexical high falling tone. If the rate is below chance, listeners are probably receiving some information that helps differentiate sandhi tones from citation tones.

The ID rates for the stimuli with high falling and low falling test syllables are expected to be the same. Similarly, the ID rate for the stimuli from the Isolation Environment is expected to be no different from those from the Sentence Environment in that the tones surrounding the test syllables were controlled with the same mid level tones.

The surface tone identification point (IDP) and the surface tone recognition point (RP) encoded how much of a word a listener needed to hear in order to make their decision. The IDP is the first gate (in msec) for which the correct identification was reached without further changes in response for later gates. The RP is the first gate (in msec) for which the identification was correctly reached without further changes and the confidence about that response was 80% or greater. It is predicted that listeners could identify and recognize the tones from the citation position earlier than the tones from the sandhi positions because citation tones are lexical tones. It is assumed the lexical tones retain more solid acoustic information from the beginning of the presentation. The tones with low falling and high falling surface tones are expected to show no difference in IDP and RP, and so are the tones from the Isolation and Sentence Environments.

2.2.2 Results

2.2.2.1 Surface Tone Identification Rate

The average responses for the "match" (1) and "no-match" (0) answers to full words are summarized in Table 2-2. An answer was considered 'match' when the listener's identification response at the last gate matched the surface tone of that syllable. An answer was a 'no-match' when listeners' response at the last gate did not match the surface tone.

Table 2-2. Average identification to the last-gate stimuli (on a scale from 0 from 1, where 1 response is a match to the surface tone and a 0 is a non-match). The standard deviations are in parentheses.

	Citation (*n*=160)	Sandhi (*n*=160)
Isolation		
Low falling tone	0.63 (0.49)	0.83 (0.38)
High falling tone	0.68 (0.47)	0.68 (0.47)
Sentence		
Low falling tone	0.80 (0.41)	0.78 (0.42)
High falling tone	0.60 (0.50)	0.83 (0.38)

The average surface tone identification rates for all last-gate stimuli (M = 0.73, SD = 0.45) were entered into a 2 (Surface Tone) × 2 (Position) × 2 (Environment) ANOVA. A main effect was only found for Position, $F(1, 9) = 10.44$, $p < .05$. The rates for the tones taken from sandhi position are significantly higher than the rate for tones from citation position. The sandhi tones sound more like the reference citation tones than the citation tones do.

Figure 2-2 shows the average rates. Even though the rates for the sandhi position and the citation position are both above chance, the fact that sandhi position had a significantly higher match than the citation position was unexpected. Neither Surface Tone nor Environment affected the rates, and there were no significant interactions.

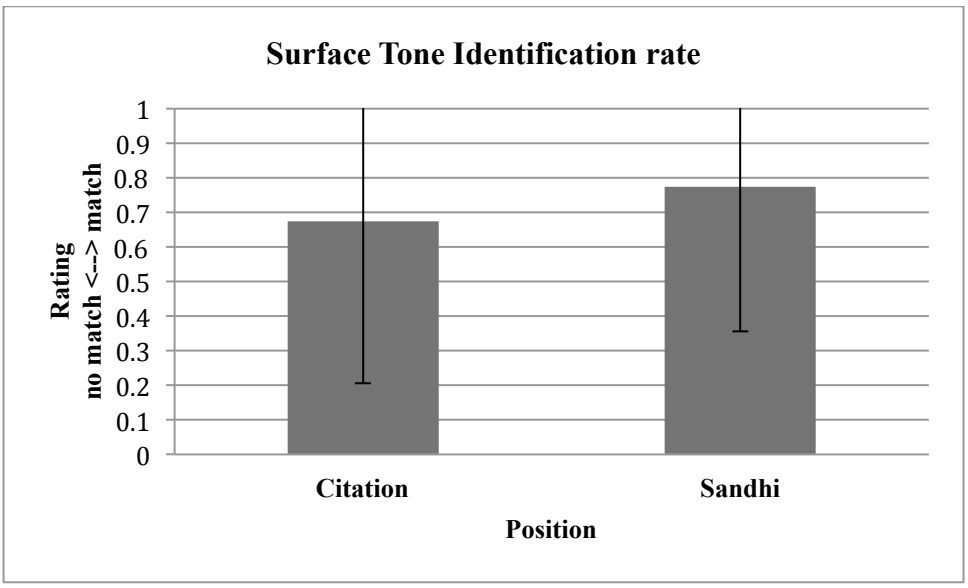


Figure2-2. Surface tone identification rate. The tone in the sandhi position (M=0.775, sd=0.42) has higher identification rate than tone in the citation position (M=0.675, sd=0.47).

In summary, the rates in each condition were better than chance, with a higher rate of success in the identification of sandhi tones. Combining all Surface Tones and Environments, the average identification rate is only 67.5% for citation tones but 77.5% for sandhi tones (as shown in Figure 2-2). Interestingly, even though the full syllable is heard, listeners did not always identify the words by taking the Surface Tone as the answer; overall correctness across all factors is only 73%. While the difference in rate shows that the tones at different positions sound somewhat different to listeners, the difference does not seem to be interpretable as lexical tone distinctions. Instead, the sandhi tones sound more like the surface citation tones, not like a different tone category.

Whether the disyllabic phrases were originally in isolation or embedded in a sentence, the tones in sandhi position is more like the intended tone, whereas the tone at the citation position is more readily perceived as the other tone. The preceding and the following tones in the recording

material were both controlled as mid level tones, and therefore tonal coarticulation should not be a large factor in tone identification.

The difference between the rate on citation tones and the rate on sandhi tones can also be used as a measure of the degree of neutralization. If neutralization is complete or near-complete, then these scores will be the same (whether both 100% or both 0%, or both 50%) and their difference will be 0. If neutralization is incomplete, listeners could get one category 100% correct and the other at chance, and then the difference is 50 (100%-50%). But if listeners can clearly tell the citation tones from the sandhi, undo the sandhi, and match the sandhi tones to their lexical tones, then the tones are completely distinct; the difference is 100 (100%-0%). Here, the difference between the rate on citation and sandhi tones is 10 (77.5%-67.5%), which suggests an incomplete neutralization, but barely so, and as discussed above, not interpretable in terms of distinct lexical tone categories.

2.2.2.2 Surface Tone Identification Point (IDP)

The definition of the surface tone identification point is the time where the tone was correctly identified without further changes in response. The data input therefore became an unbalanced design because some of the items were excluded for never being correctly identified by the listeners. Since repeated measures ANOVA with aov in R does not work for an unbalanced design, the linear mixed-effect model lme is applied to the IDP (in msec) with the three experimental variables – Position, Environment and Surface tone – as fixed effects, and the random factor was the listeners.

Table 2-3 summarizes the results. Main effects were found for all three factors, and are shown in Figure 2-3. Post-hoc analyses indicated that (i) tones taken from sandhi position were recognized earlier than tones from citation position; (ii) the IDP for the stimuli recorded in a

carrier sentence was earlier than that for the stimuli recorded in isolation; (iii) the IDP for the surface high falling tone (53) was earlier than that for the low falling tone (31). One explanation for the third finding about Surface Tone is that in both Environments and at both Positions, the syllable of interest was adjacent to either a prosodic boundary or a mid-level-tone syllable, which makes the occurrence of the high-pitched onset of the high falling tone quite distinctive. Finding (ii) about Environment suggests that for syllables recorded in a carrier sentence, some carryover or anticipatory cues must have been conveyed to the listeners so that they could identify those syllables earlier. Finding (i) is interesting because we would have expected that syllables from sandhi position would be more vulnerable and more dissimilar to the target tones since they were the products of tone sandhi. However, this finding could be explained with the same account that we had for the second finding, that is, the syllables at the sandhi position might have contained more information, which benefits listeners' tone identification. Another possible explanation would have to do with the duration ratio of the stimulus presentation. Previous studies on non-checked tones have shown that final lengthening is observed at phrase- and utterance-final position (Peng 1997), that is, the citation position. Therefore, the absolute value of the identification point (the raw duration in msec) does not seem to be a measure as ideal as the proportion of the presented stimulus to the overall stimulus for each Surface Tone. In this dissertation, we examined the raw IDP duration (as well as the raw RP duration reported in the later section), following what has been done in most of the previous Gating experiments. In future work, instead of the raw duration, the durational ratio should be a better measure assessed in this case given that the duration of sandhi tones and citation tones could be different.

Table 2-3. Coefficients in the linear mixed-effect model predicting the surface tone identification point (IDP) from three variables: Position, Environment and Surface tone.

	Coefficients: Estimate	Std.Error	t-value	p-value
(Intercept)	88.01904	8.520160 212	10.330680	0.0000*
position	-30.57026	9.327244 212	-3.277523	0.0012*
environment	-45.70340	9.425299 212	-4.849013	0.0000*
surfaceT	-31.09991	9.729958 212	-3.196304	0.0016*
position:environment	45.57888	12.574231 212	3.624785	0.0004*
position:surfaceT	24.24797	13.148963 212	1.844098	0.0666
environment:surfaceT	37.65057	13.343454 212	2.821651	0.0052*
position:environ:surfT	-36.59619	18.060693 212	-2.026289	0.0440*

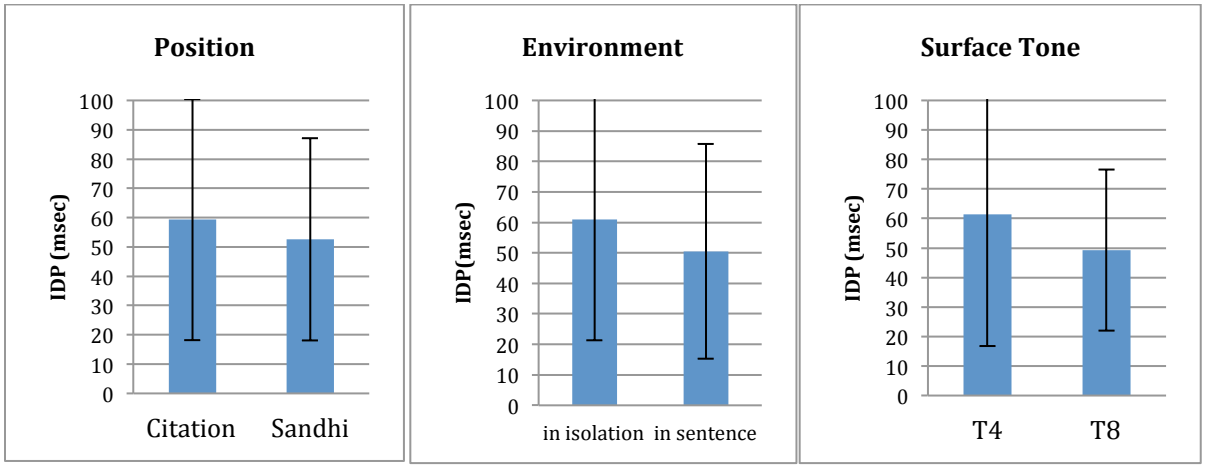


Figure 2-3. IDP (msec) with separate graphs given for each variable that showed a main effect. A lower value means earlier recognition. In the third graph, T4 stands for low falling tone, and T8 stand for high falling tone.

Moreover, the effects of these three factors are very similar – about 10-15 msec, which is equivalent to 2-3 gates. In addition, listeners were able to correctly identify a tone with only limited exposure of 50-60 msec to the stimuli.

The significant interaction between Position and Environment, and the significant interaction between Environment and Surface tone are shown in Figure 2-4. The post-hoc tests for the interaction between Position and Environment show that the IDP is significantly earlier when the tone was from citation position in a sentence than when the tone was from citation position but in

isolation. When the tones are both from isolated phrases, the one from sandhi position requires less time to be identified than the one from citation position. Both results suggest that listeners must have received some prosodic information from the surrounding context, and have used this information to help with the judgment.

The post-hoc tests for the interaction between Environment and Surface Tone show that when both tones were from in isolation, the identification point for the surface perceived high falling tone (53) is significantly earlier than the surface low falling tone (31). A possible explanation is that the onset of a high falling tone is a high register tone and this is relatively more salient.

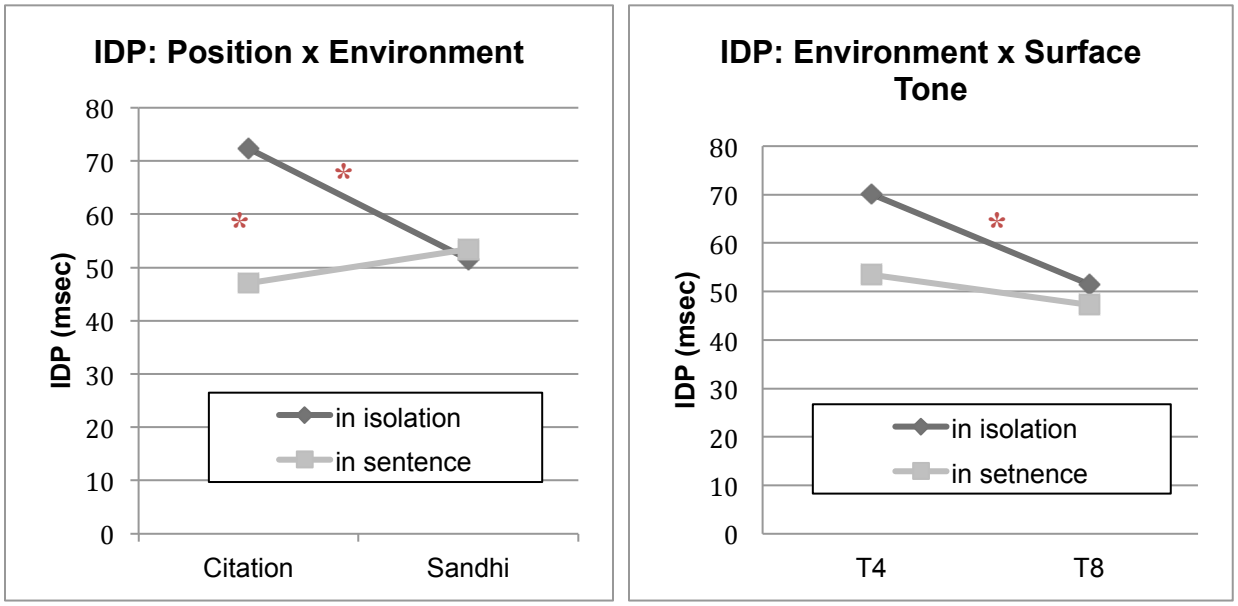


Figure 2-4. Surface tone identification point (msec), with two interactions: (a) Position × Environment; (b) Environment × Surface tone. Error bars reflect one standard deviation. The asterisks show the significant difference in post-hoc tests.

2.2.2.3 Surface tone Recognition Point (RP)

The definition of the surface tone recognition point (RP) is the point where the tone was correctly identified without further changes and the confidence rating was above 80%.

Essentially, the difference between the IDP and the RP was determined by listeners' confidence about their answers. The data were entered into a linear mixed effect model. No significant effect was found in either of the factors or their interaction. This result indicates that listeners take the same amount of time to confidently recognize tones from different positions and from different environments.

2.2.3 Summary

This study used the gating paradigm to examine how listeners identify the two Taiwanese checked tones. The result for the final gates shows that tones from the sandhi position received higher citation-response rates than the tones from the citation position, which suggests that listeners tended to match the sandhi tone to the lexical tone more than matching the lexical tone to itself. However, although there was a significant difference in the rates given to tones from these two positions, both sets of rates were above chance. So we can say the listeners clearly had a tendency to match the perceived surface tone to the corresponding lexical tone, e.g. when hearing a surface high falling tone whether it was a sandhi tone or a citation tone, the listeners would choose 'high falling' as the answer.

The IDP results reveal that all three factors (Position, Environment and Surface Tone) result in significant differences in the time required for identification. Interestingly, listeners needed less time to identify tones from sandhi position, tones from a sentence and tones with high-register onset. These findings suggest that some phonetic cues must have coarticulated with the stimuli, and these cues are relevant to "the location of the boundary", the "context" and the "F0 contour". Since the "context" and the "F0 contour" are both related to syllables outside of the tone sandhi group domain ("context": whether there are other syllables in this utterance; "F0

contour": whether the adjacent sounds have a mid level tone), the only thing that matters within a tone sandhi group is the cues concerning the location of the boundaries. In other words, lexical and sandhi tones can be neutralized, but only partially, because the occurrence of these tones depends on their position relative to the boundary.

Notice that the measure of the IDP was the raw duration of the stimulus presentation. As suggested earlier, it is very likely that sandhi tones and citation tones are intrinsically different in duration. Therefore, the ratio of the duration of the stimulus presentation to the duration of the entire stimulus might be a more valid variable to be assessed. I will leave this for future work.

2.3 Corpus study

In the previous section, we examined how tones extracted from sandhi position and from citation position are perceived. Listeners did not need to wait until the entire syllable was presented to make the correct choice. It is likely that some phonetic cues regarding the upcoming boundaries and surrounding environments were available at an early point in the syllables. In this section, we would like to make more comprehensive observations about the acoustic differences between sandhi tones and citation tones, using a Taiwanese speech corpus.

Final lengthening (a.k.a. *pre-boundary lengthening*) refers to the often-observed pattern that the final syllable preceding a prosodic boundary is longer than that syllable in a no-boundary position. Lengthening is a robust boundary marker in a variety of languages (English: Price 1991, Wightman et al. 1992, Turk and Shattuck-Hufnagel 2007; Mandarin: Shen 1992; Korean: Cho and Keating 2001, among others.) Peng (2007) also found final lengthening at utterance-final and phrase-final positions in Taiwanese. *F0* is another salient cue that has been proposed by a number of studies, especially studies with tone languages. *Creaky voice phonation* (a.k.a.

laryngealization) often accompanies a low F0 value and could be used as a boundary signal if it occurs more frequently before boundaries. In a study of Mandarin Chinese connected speech, Belotel-Grenié and Grenié (2004) found that the second syllable of a disyllabic word (therefore, at a prosodic boundary) is where the creaky voice was found because the latter syllable is associated with a lower F0 value.

In our corpus study, we selected disyllabic phrases that appeared at an intonation phrase boundary (IP, hereafter) and a Tone Sandhi Group boundary (TSG, hereafter) in read speech. Durational, F0 and phonation effects are expected to be found in the second syllables of the disyllabic phrases (the syllables in the pre-boundary, citation position) as opposed to the first syllables, and the effects are expected to be observed especially at the bigger boundary (i.e. the IP boundary).

2.3.1 Method

The data were taken from the Taiwanese corpus collected by the Signal Processing Laboratory at National Chiao Tung University[5]. A trained Taiwanese native male speaker read a story written in Chinese characters. The data included both the audio signals and Praat Textgrids which labeled three levels of prosodic boundaries using the Taiwanese ToBI conventions (Peng and Beckman 2003) – word/syllable boundaries, Tone Sandhi Group boundaries and Intonation Phrase (IP) boundaries. The syllable boundaries were first obtained through a forced alignment speech recognition system developed in the lab, and then were manually checked by at least two trained native speakers. Other levels of prosodic boundaries were manually labeled and were as well cross-checked among at least four trained native speakers.

[5] Many thanks go to Professor Sin-Horng Chen for generously sharing the corpus with me.

The data extracted from the corpus included disyllabic phrases with 42 possible tonal sequences (6 surface tones at sandhi position x 7 surface tones at citation position). The tones are shown in Table 2-4. The acoustic measures were taken from each syllable, which could appear at either of the two Positions (sandhi vs. citation), and from either of the two Prosodic Boundaries (IP vs. TSG). Note that a sandhi tone is always one syllable away from the boundary.

Table 2-4. Taiwanese tones. There is no Tone 6 in modern Taiwanese because it has merged with 51 and surfaced as high falling. 24 is a rising tone, which only appears in citation position, which means there is no rising sandhi tone. 31q and 53q are the two checked tones.

Names of the lexical tones	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 7	Tone 8
Citation form	55	51	31	*31q*	24	33	*53q*
Tone description	High level	High falling	Low falling	Low falling checked	Rising	Mid level	High falling checked
Sandhi form	33	55	51	*53q*	33	31	*31q*

The acoustic measures included the time span of the F0 contours (duration), the F0 magnitude (F0 range), and the voice measure H1*-H2*, across the voiced portion of each syllable. H1*-H2* refers to the difference in amplitude between the first and second harmonics, and is often correlated with creakiness. The asterisk shows that the values were corrected for formant frequencies and bandwidths. A lower value of H1*-H2* suggests a smaller glottal open quotient, as in a laryngealized voice quality such as creaky voice. These measurements were obtained using VoiceSauce (Shue *et al.* 2011) and Praat (Boersma *et al.* 2012).

2.3.2 Results

2.3.2.1 Syllable duration

The distributions of the durations for the seven surface tones are plotted in Figure 2-5. This gives us an overview of the duration for each tone, which here includes both citation and sandhi tones. That is, for instance, the median duration for surface high level tone 55 is around 150 msec, which was the average duration across the lexical 55 and the sandhi form of 51.

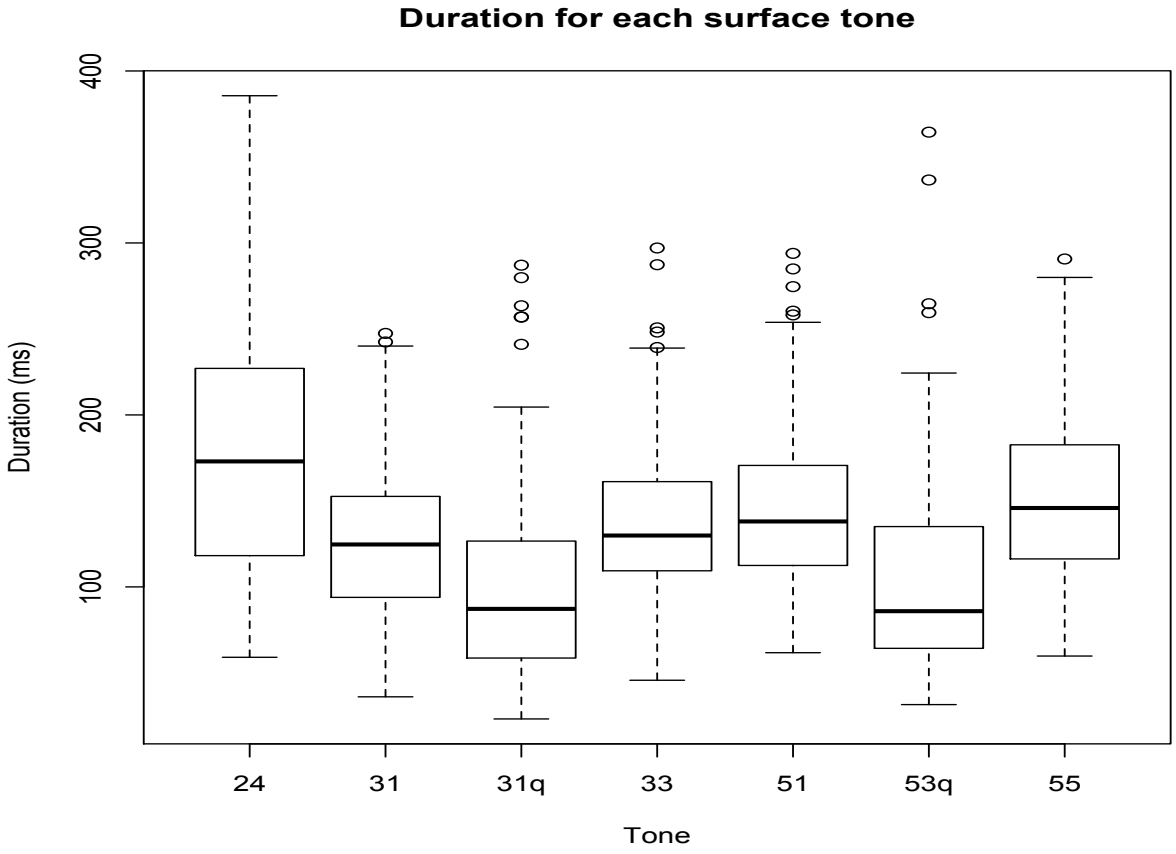


FIGURE 2-5. Boxplot comparing the duration for seven different surface tones in Taiwanese. The values on the axis are the tonal values for these tones. The line in each box indicates the median duration. 31q = low falling checked tone; 53q = high falling checked tone. These are average duration across citation tones and sandhi tones. Note that there is no surface rising sandhi tone, so 24 includes only citation tones.

The data were entered into a mixed design ANOVA (2 Positions × 2 Prosodic Boundaries × 7 Surface tones). Main effects were found for all three factors. First, citation tones (i.e. at a boundary) are longer than sandhi tones (i.e. not at a boundary); second, measurements from an IP boundary are longer than those at a TSG boundary. These two findings can be accounted for by the final lengthening effect. The differences in the Surface tones show that the rising tone (24) is the longest tone whereas the two checked tones (31q and 53q) are the shortest.

As a form of post-hoc comparison, the measurements *for each surface tone* were then entered into a series of 2 Positions × 2 Prosodic Boundaries mixed design ANOVAs. The results are given in Table 2-5 and Figure 2-6. For 55 (High level), 51 (High falling), 31q (Low falling checked), and 33 (Mid level), main effects were found for both Position and Prosodic Boundary. In other words, these tones are longer in syllables in citation position and disyllables at an IP boundary as opposed to sandhi position and TSG boundary. This presumably results from final lengthening. For 31 (Low falling) and 24, main effects were found for Prosodic Boundary, indicating that 31 and 24 was significantly longer in disyllables at IP boundary than at TSG boundary and that sandhi 31 and citation 31 were neutralized in terms of duration. For 53q (High falling checked), a main effect of Position was found; 53q was significantly longer in citation position than in sandhi position. In general, then, most of the second syllables are longer than the first syllables from these disyllables, and thus it is possible that the relative durations of tones in a sequence are informative to listeners. For 55 and 33, they both had interaction effects of Position and Prosodic Boundary. Tukey HSD tests for 55 show that citation tone at IP is longer than citation tone at TSG while sandhi tone at IP is the same as the sandhi tone at TSG. Tukey HSD tests for 33 show that citation tone at IP is longer than citation tone at TSG, and sandhi tone

at IP is also longer than sandhi tone at TSG. That is, final lengthening extends back to the sandhi –tone syllable.

TABLE 2-5. Mixed-design ANOVA results for duration for each tone. Rising tone (labeled as 24 here) never appears in a sandhi position, therefore, the ANOVA analysis of 24 is done separately with Prosodic Boundary as the sole factor. (c: citation; s: sandhi; I: IP; T: TSG)

	55	51	31	31q	24	33	53q
Position	c>s	c>s		c>s	n/a	c>s	c>s
Prosodic Boundary	I>T	I>T	I>T	I>T	I>T	I>T	
Position × Prosodic Boundary	√				n/a	√	

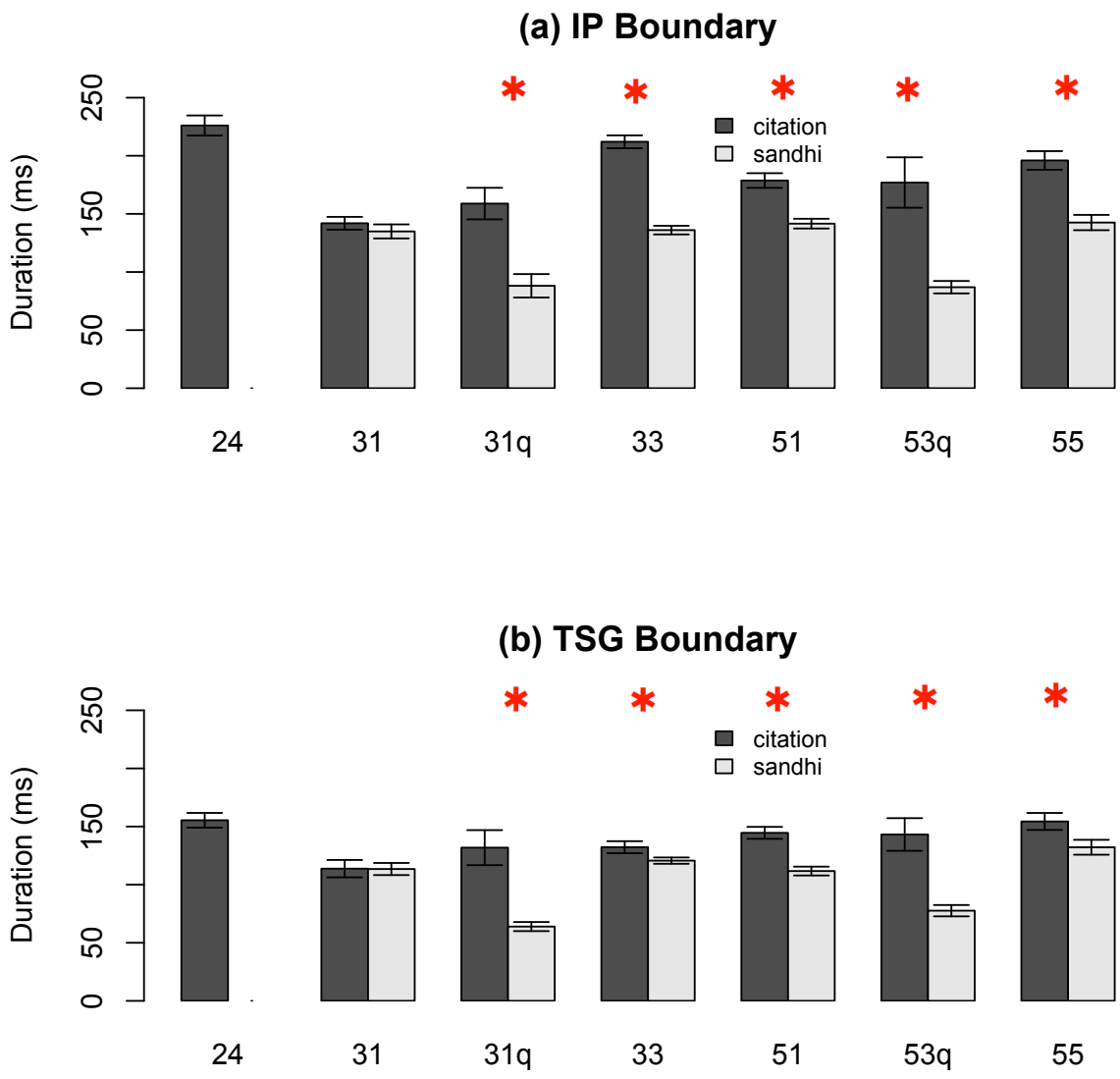


FIGURE 2-6. Mean duration of the seven tones at: (a) Intonation Phrase boundary; (b) Tone Sandhi Group boundary. Error bars reflect +1/-1 standard errors. 24 is a rising tone that only appears at citation position, which means there is no sandhi tone that has a rising tone as its output.

2.3.2.2 F0 range

The next measurement of interest is the F0 range of each tone. Pan's (2002) study found that the raw F0 range of the checked tones differs with the adjacent boundaries, i.e. word < TSG < IP.

The bigger the boundary, the wider the F0 range. To be comparable to previous studies, the raw F0 ranges were converted into the logarithmic F0 range. (Previous studies use logarithmic F0 range to eliminate the gender effect, i.e., women's pitch is higher than men's pitch.) The logarithmic F0 range is the log of the ratio between the minimum F0 and the maximum F0 within each syllable. The value of the logarithmic F0 range for *each surface tone* was examined with the same two factors – Position and Prosodic Boundaries.

Results show that both factors had main effects on the logarithmic F0 range, with (i) disyllables at an IP boundary having a wider F0 range than syllables at a TSG boundary, and (ii) citation tones having a wider F0 range than sandhi tones. Table 2-6 shows the ANOVA results for each tone with the two factors "Position" and "Prosodic Boundary". Figure 2-7 plots the logarithmic F0 range for each tone at different prosodic boundaries and from different positions. Post-hoc t-tests for each individual tone show that the difference between prosodic boundaries mainly came from 51 (High falling) in which the F0 range in disyllables at the IP boundary is bigger than that at the TSG boundary. On the other hand, the difference between Positions mainly came from 31 (Low falling) and 33 (Mid level), though these effects are contradictory: 31 has a wider F0 range for citation tones but 33 has a wider F0 range for sandhi tones. The only interaction effect was found in 33, and the Tukey HSD tests show that the citation tones at TSG and IP do not differ from each other, and neither do the sandhi tones at the two prosodic boundaries. It seems that the logarithmic F0 range is not a cue as strong as duration in that fewer significant effects were found.

TABLE 2-6. Mixed-design ANOVA results for logarithmic F0 range for each tone. Rising tone (labeled as 24 here) never appears in a sandhi position, therefore, the ANOVA analysis of 24 is done separately with Prosodic Boundary as the sole factor. (c: citation; s: sandhi; I: IP; T: TSG)

	55	51	31	31q	24	33	53q
Position			c>s		n/a	s>c	
Prosodic Boundary		I>T					
Position × Prosodic Boundary					n/a	√	

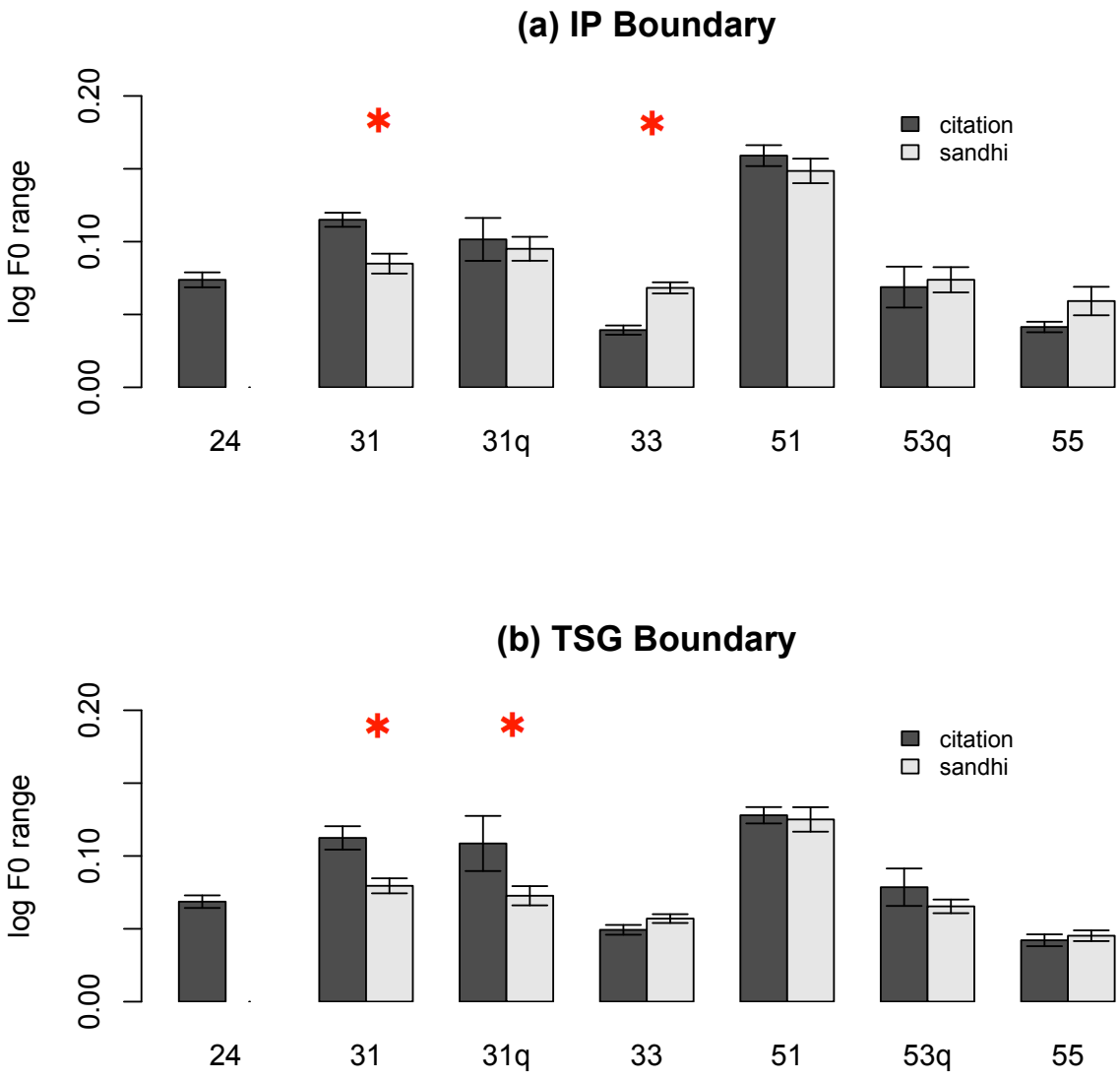


Figure 2-7. Mean logarithmic F0 range of the seven tones at: (a) Intonation Phrase boundary; (b) Tone Sandhi Group boundary. Error bars reflect +1/-1 standard errors. In either Prosodic boundary, the high falling tone (51) appears to have wider F0 range, followed by other falling tones (31, 31q, 53q), which in turns have wider F0 range than level tones (55 and 33). There is no sandhi tone whose output is a rising tone (24).

2.3.2.3 H1-H2**

Main effects were found for Position and Prosodic Boundary: (i) disyllables at an IP boundary have lower H1*-H2*, which suggests that IP boundaries tended to be creakier, and (ii) citation tones also tended to be creakier than sandhi tones.

The ANOVA analyses for the tones are shown in Table 2-7 and Figure 2-8. Interestingly, at the IP boundary, even the mid level tone (33) could be very creaky in that the H1*-H2* value is below zero. However, not every tone showed this effect; especially, 31 patterns differently from most other tones. The post hoc tests show that all the tones use H1*-H2* to differentiate IP boundary from TSG boundary; IP has creakier voice quality than TSG. For 51, 33 and 53q, they also use H1*-H2* to differentiate tones at citation position from tones at sandhi position; citation tones are creakier than sandhi tones.

TABLE 2-7. Mixed-design ANOVA results for H1*-H2* for each tone. The ANOVA for 24 tone data was analyzed separately with 'Prosodic Boundary' as the only factor. It is because rising tone (24) does not occur in sandhi position. The smaller the H1*-H2* values, the creakier the voice quality. (c: citation; s: sandhi; I: IP; T: TSG).

	55	51	31	31q	24	33	53q
Position		c<s			n/a	c<s	c<s
Prosodic Boundary	I<T	I<T	I<T	I<T	I<T	I<T	I<T
Position × Prosodic Boundary					n/a		

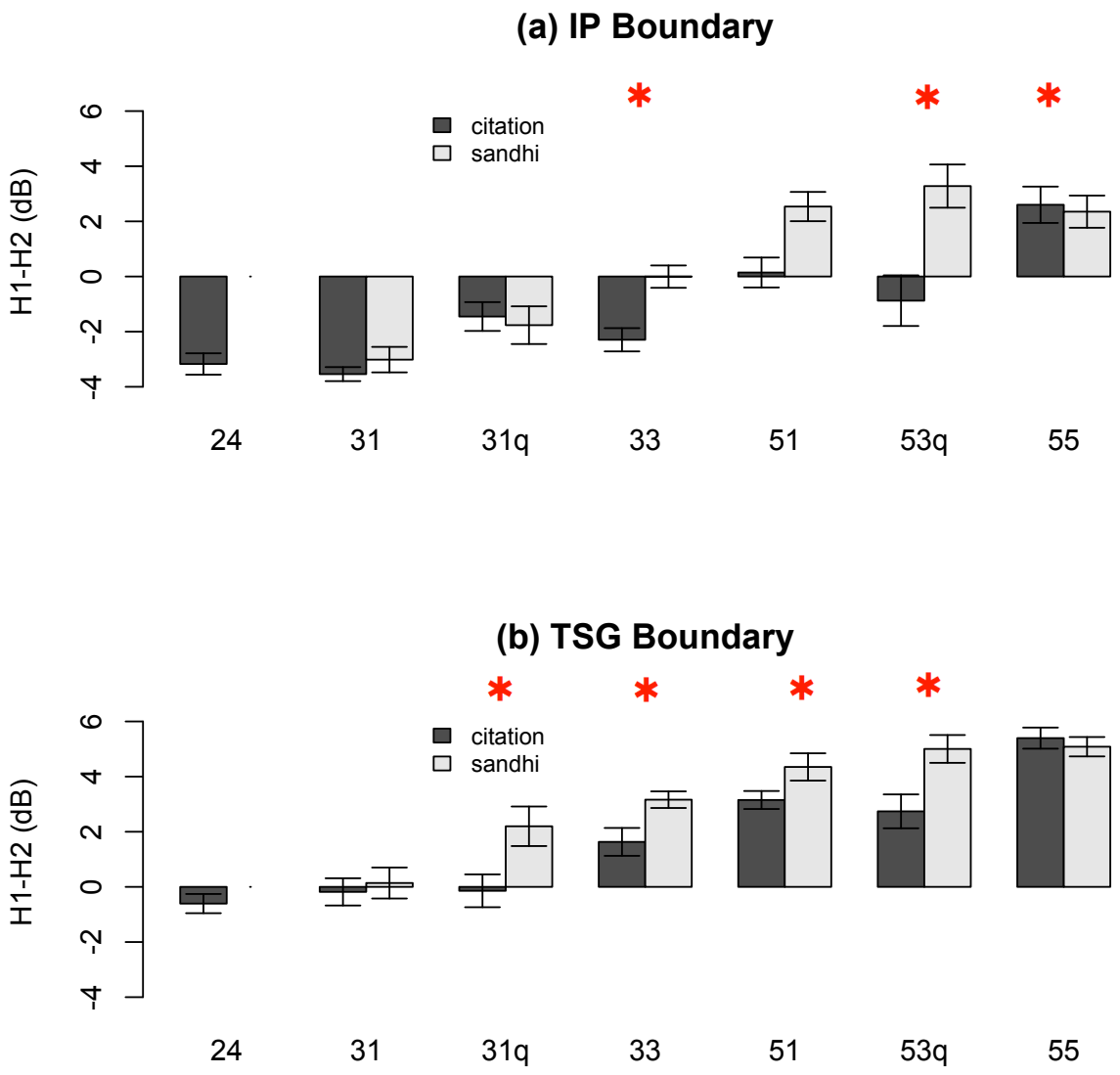


FIGURE 2-8. Mean H1*-H2* of the seven tones at: (a) Intonation Phrase boundary; (b) Tone Sandhi Group boundary. Error bars reflect two standard errors. There is no sandhi tone whose output is a rising tone (24).

2.3.3 Summary

This study was designed to examine tone sandhi and tone neutralization in a large corpus. Our results confirm that for the speaker who recorded the corpus, there are acoustic differences between the citation tones and the sandhi tones in terms of duration, F0 range and H1*-H2*.

More specifically, citation tones tended to have greater length, wider F0 range and creakier voice than sandhi tones, but not for all the tones all the time. Therefore, overall, citation tones and sandhi tones are not completely neutralized.

Since sandhi tones and citation tones are distinctive in duration, F0 and voice quality to some extent, when listeners hear a sandhi tone, they could recognize that this is not a citation tone because it lacks the properties of TSG-final positions.

Chapter 3: Tone Sandhi Group as a Prosodic Domain

3.1 Introduction

Chapter 2 indicated that the Taiwanese Tone Sandhi Group (TSG) is distinct from the Intonation Phrase (IP) in terms of duration, F0 and voice quality for one speaker at least. However, the utterances of interest in Chapter 2 were disyllabic phrases, rather than longer utterances. In this chapter, we examine listeners' behaviors when they are presented with longer fragments from read sentences. We test whether listeners distinguish Word boundary from a TSG boundary, as well as from IP boundary. In addition, the perceptual distinction between TSG and IP is also examined.

Previous studies have shown that listeners are able to detect major prosodic boundaries in meaningless speech stimuli, such as low-pass filtered speech (de Rooij 1975; Kreiman 1982) and hummed speech (t'Hart and Cohen 1990; Pan 2011). These are stimuli that do not contain lexical and syntactic information, and listeners could still make use of the prosodic information to discover the boundaries. The acoustic information in filtered speech and hummed speech includes duration and pitch, and some voice quality, all of which were found to vary with boundaries in Chapter 2.

The present study was designed to address these issues by establishing a more complete description of listeners' identification of sandhi and citations tones[6] in longer fragments. Accordingly, we assessed syllables at Word boundary, TSG boundary, and IP boundary.

This experiment contains two parts: the first part involves the comparison of the perception of three prosodic domains in filtered speech – Word boundary, TSG boundary and IP boundary. The amount of speech in each presented stimulus was four syllables. Specifically, we expected

[6] "Citation tone" refers to the lexical tone at the citation position. Since the lexical tone at the citation position retains its original tonal contour, "citation tone" and "lexical tone" are used interchangeably in this dissertation.

listeners to be able to predict the adjacent boundaries accurately with this limited amount of exposure in quality (filtered speech) and in quantity (four syllables).

The second part was the comparison of the shared part at a TSG boundary and an IP boundary in both filtered and unfiltered (normal) speech. This task is relatively more difficult because the whole TSG overlaps with part of the IP from the sentence onset. We expected to find that listeners could still use acoustic information to predict the adjacent prosodic boundary.

3.2 Part I: Comparisons between Word Boundary and TSG Boundary and between Word Boundary and IP Boundary

3.2.1 Method

3.2.1.1 Participants

A total of 30 adult native speakers of Taiwanese participated in the experiments as listeners. None reported hearing or vision problems and all were compensated for their participation with ten dollars. Two participants' results were removed from the analysis due to failure to complete the experiment within the allotted time.

3.2.1.2 Stimuli

For each task, carefully matched pairs of sentences were constructed to differ minimally in the size of the prosodic boundary occurring after a target syllable. The sentences were recorded by a trained female Taiwanese speaker, who had been informed of the boundary difference between the sentences in each pair. She first produced all the sentences with the IP boundary, followed by all the sentences with the TSG boundary, and then all the sentences with the Word

boundary. She was told to produce the sentences as naturally as possible and not to pause in the middle of a sentence.

A portion of each sentence was extracted to make a test stimulus, as described below. To eliminate the lexical cues that native speakers would otherwise use in the perception, the experimental stimuli were then low-pass filtered, using Praat, at a frequency cut-off of 400 Hz and with 50 Hz smoothing. The peak intensity of all stimuli was adjusted to 70 dB.

Five CV syllables were the target syllables in this study: /pi/, /u/, /pa/, /tə/ and /tʰe/. Each of the five syllables was combined with the five different non-checked tones (High level, mid level, high falling, low falling and rising) to form real disyllabic phrases with another syllable with mid level tone. This yields twenty-five test phrases. The sentence types recorded for each task in Part I are described below.

3.2.1.2.1 Word boundary vs. Intonation Phrase boundary

In this task, the 25 combinations of syllables and tones were placed before a Word boundary or an IP boundary. Stimuli consist of pairs of sentences like (1), which demonstrate a pair of test sentences with the target syllable *pi55*. The tones transcribed here are all surface tones, and the pound sign (#) denotes the tone sandhi group boundary. In both conditions, *a0* is an IP-initial particle that denotes the beginning of an IP[7]. The underlined tones are the tones of the syllable that undergo tone sandhi.

In the IP boundary condition as in (1a), the IP boundary is of course also a TSG boundary. Since the whole first tone sandhi group is followed by an IP-initial *a0*, it is certain that this is an IP boundary, which subsumes the TSG boundary. In the Word boundary condition as in (1b), the Tone Sandhi Group (and also IP) boundary occurs one syllable after the test syllable. In (1a),

[7] IP-initial *a0* has a neutral tone but it is different from the sentence-final particle *a0* as we've already seen in Chapter 2.

the test syllable is the second word of a disyllabic phrase ‘coffee’. In (1b), the test syllable is a determiner specifying the following noun, which has a close juncture with it. Therefore, the *pi55* in (1a) is located at an IP boundary and carries a citation tone, whereas the *pi55* in (1b) is located in the penultimate position relative to the IP boundary and carries a sandhi tone whose lexical tone was /pi51/.

(1) a. IP

*i33 tiaN33 ka33-**pi55** # a0 tan31-si31 be3- hiau55 sia55 chit5 nng31 ji33 #*

s/he listen to coffee A but cannot write the two character

“S/he listened to ‘coffee’. But (s/he) couldn’t write these two characters.”

b.Word

*i33 tiaN33-bo33 **pi55**-hoaN33 # a0 tan31-si31e31-hiau55 sia55 chit5 nng31 ji33#*

s/he listen to-not that riverbank A but can write the two character

“S/he didn’t understand ‘that riverbank’. But (s/he) could write these two characters.”

In this task, the listeners were presented visually on the screen with the written forms of the first tone sandhi group from a pair of sentences in each trial (e.g. up to the first # sign in (1a) and (1b)). They then heard just the highlighted 4-syllables from one of the two sentences. They were explicitly told what has been described above and that their task was to choose which written phrase corresponded to the audio stimulus (two-alternative forced-choice). Due to the large difference in prosody between the Word boundary and the IP boundary, this is expected to be a relatively easy task. Note that although the tone sandhi groups in the two sentences have

different lengths, the portions presented auditorily have the same lengths (here, four syllables). Because the four syllables were not exactly the same in each pair of sentences (e.g. *ka33* in (1a) and *bo33* in (1b)), the low-pass filtered stimuli were used in order to eliminate the segmental differences.

3.2.1.2.2 Word boundary vs. Tone Sandhi Group boundary

In this task, the same 25 combinations were placed before a Word boundary in the same way as before, or before a TSG boundary. (2) demonstrates a pair of test sentences with *pi55*. In (2a), the sentence begins like those above with an IP boundary (cf. (1a)), but the test syllable *pi55* is more closely related to the word that follows it, *e33,* and this makes the boundary here a TSG boundary. Therefore, *pi55* in (2a) is located at a TSG boundary and so carries a citation tone. However, the *pi55* in (2b) carries a sandhi tone because it is not adjacent to a TSG boundary.

(2) a. TSG

*i33 tiaN33 ka33-**pi**55 # e33 siaN33-tiau33 #*

s/he listen to coffee GEN tone

"S/he listened to the tones of 'coffee'."

b. Word

*i33 tiaN33 bo33 **pi**55-hoaN33 # e33 siaN33-tiau33 #*

s/he listen to not that riverbank GEN tone

"S/he didn't understand the tones of 'that riverbank'."

During the experiment, the listeners were presented visually with the written forms of the first TSG (up to the first # in (2a) and (2b)), and heard the 4-syllable auditory stimuli from either of the highlighted parts. Their task was to choose which sentence the auditory stimuli came from. Similarly, the stimuli were low-pass filtered so that listeners could not make the judgment based on their lexical knowledge of the words.

3.2.2 Procedures

The experiment comprised two sessions. The first session functioned as a practice session, with normal speech stimuli. The second session presented the filtered speech stimuli. In each session, both tasks were presented in a counterbalanced order, with a break between them. Listeners were told that they would hear a series of Taiwanese utterances over noise-cancelling headphones and would be offered two alternatives on the laptop screen for each utterance. They were asked to respond with their initial intuitions and choose "which highlighted part the speaker said."

The experiment was conducted with a Matlab program using Psych Toolbox. In each trial, the participants saw a speaker sign and the written forms of the sentence alternatives on the screen with the first four syllables highlighted. They had to click on the speaker sign when they were ready to listen to the utterance, and they had to click on the sentence of their choice as the answer. And then, they had to press the space key to proceed to the next trial. They could only hear each utterance once.

3.2.3 Analysis

The accuracy of the listeners' responses was analyzed. In addition, responses were converted to a d' score to assess the listeners' sensitivity to the Boundary conditions. The d' was calculated by subtracting the false alarm ratio from the hit ratio. The bias for each condition is calculated by having the sum of the hit ratio and false alarm ratio divided by the overall responses. The formulas are shown in (3) and (4). More detailed information about these analyses is presented in the results section below.

(3) d' = *Hit – False Alarm*

(4) Bias measure = (*Hit+False Alarm)/(Hit+False Alarm+Miss+Correct Rejection)*

3.2.4 Result

3.2.4.1 Accuracy

The results for correctness are demonstrated in Figure 3-1. Figure 3-1 (a) shows the comparison between IP boundary and Word boundary. Figure 3-1 (b) shows the comparison between TSG boundary and Word boundary.

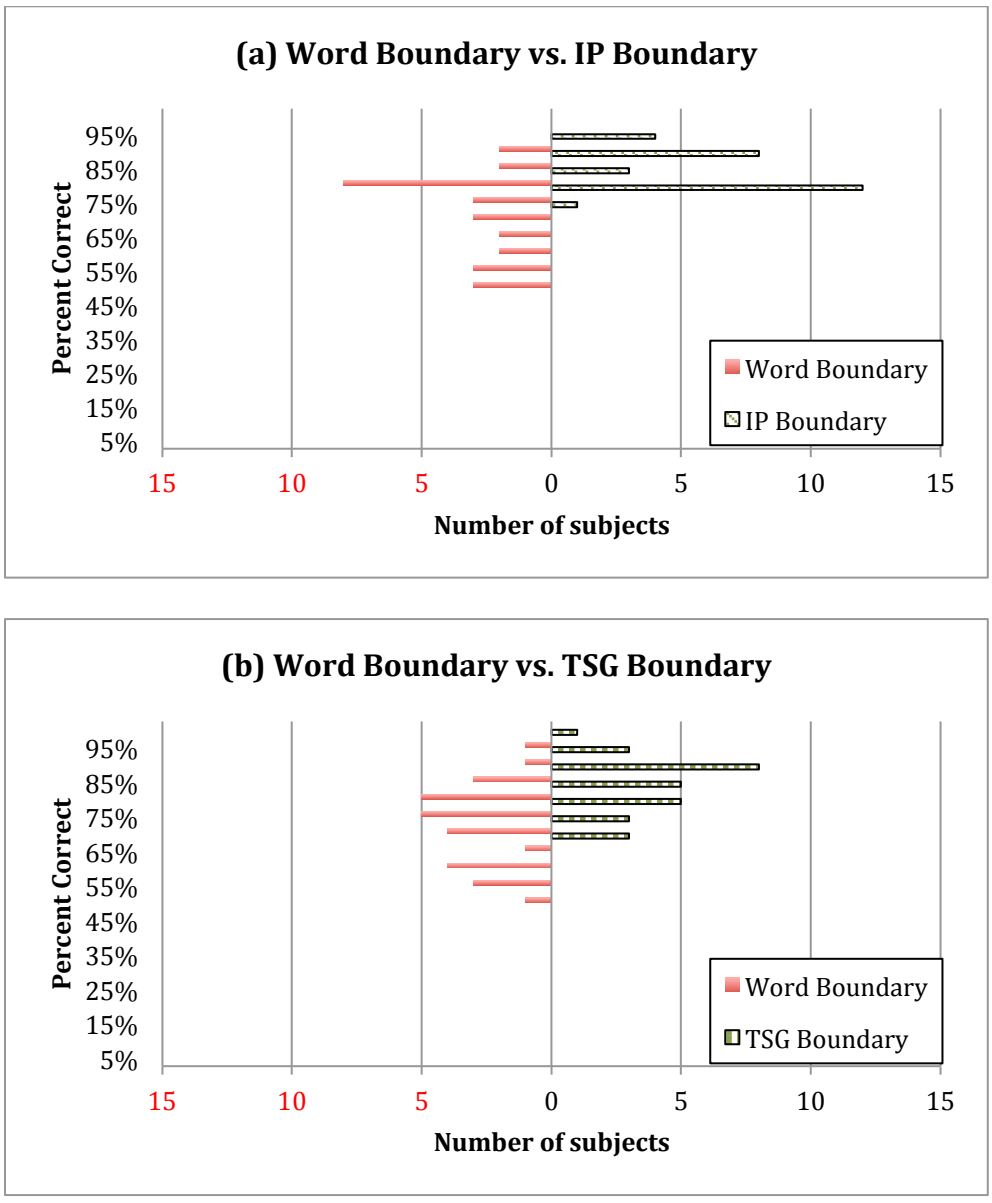


Figure 3-1. Double histograms of participants' responses in Identification tasks: (a) Word Boundary vs. IP Boundary (b) Word Boundary vs. TSG boundary. In each graph, the y-axis is the averaged percent correct, and the x-axis shows the number of subjects who got that score. For instance, in (a), four subjects got 95% correct when they heard the IP Boundary stimuli and two subjects got 90% correct when they heard the Word Boundary stimuli.

Table 3-1. Response Proportions in the first task: IP vs. Word

	Response: IP	**Response: Word**
Stimuli: IP	*Hit: 0.839*	*Miss: 0.161*
Stimuli: Word	*False Alarm: 0.307*	*Correct Rejection: 0.693*

3.2.4.1.1 Word boundary vs. IP boundary

Listeners chose Word boundary over IP boundary when presented with Word boundary stimuli on average in 69.3% of the cases (*t*-test against chance: $t(27)=7.87$, $p<.001$). In addition, they chose IP boundary over Word boundary when presented IP boundary stimuli on average in 83.9% (*t*-test against chance: $t(27)=28.85$, $p<.001$). Listeners correctly chose IP boundary significantly more than they chose correct Word boundary (paired *t*-test: $t(27)=5.15$, $p<.001$).

3.2.4.1.2 Word boundary vs. TSG boundary

When given Word boundary stimuli, listeners chose Word boundary over TSG boundary on average 69.9% (*t*-test against chance: $t(27)=8.94$, $p<.001$). When given TSG boundary stimuli, they chose TSG over Word boundary on average 82.3% (*t*-test against chance: $t(27)=20.13$, $p<.001$). Listeners correctly TSG boundary significantly more than Word boundary (paired *t*-test: $t(27)=3.91$, $p<.001$).

3.2.4.2 Sensitivity (d' score)

The listeners were fairly accurate at the task. They had overall accuracy of 77% for the first task and 76% for the second task. In both tasks, listeners were more accurate at judging stimuli from the condition with bigger boundaries: 83.9% for the IP Boundary in the first task and 82.3% for the TSG boundary in the second task. From these results, it can be concluded that either the listeners were better at stimuli from bigger boundaries, or they were a priori biased towards that response.

To assess accuracy apart from bias, responses were converted to d' score to assess the sensitivity of the Boundary conditions in each task. Take the first task for example, the "signal" in this task was the condition whose boundary was bigger (i.e. IP Boundary) and the "noise" is the condition with the smaller boundary (i.e. Word Boundary). The outcomes are schematized in Table 3-1. Since the signal is IP, the *Hit* ratio is equivalent to the averaged proportion correct when listeners chose 'IP Boundary' when they were presented with stimuli of IP Boundary condition. *False Alarm* comes from the averaged percent correct when listeners chose 'IP Boundary' when they were presented with stimuli of Word Boundary condition.

The d', without standardizing the scores, was then calculated by subtracting the *False Alarm* ratio from the *Hit* ratio, which is 0.532 ((*Hit 0.839*) – (*False Alarm 0.307*)) in the first task. The larger the d' value, the better the listeners' sensitivity was. With this calculation, a d' value of zero means that listeners cannot distinguish stimuli with the IP Boundary from stimuli with the Word Boundary. A d' of 1 indicates a perfect ability to distinguish between IP Boundary and Word Boundary conditions. In the first task, the d' (=0.523) shows that listeners could distinguish stimuli of two Boundary conditions, though not perfectly.

Table 3-2 shows the outcomes in the second task. In this task, the "signal" is TSG. Similarly, the d' was the difference between the *Hit* ratio and the *False Alarm* ratio, which is 0.522. In this task, listeners could also distinguish the TSG condition from the Word condition, yet not perfectly.

Table 3-2. Response proportions in the second task: TSG vs. Word

	Response: TSG	**Response: Word**
Stimuli: TSG	*Hit: 0.823*	*Miss: 0.177*
Stimuli: Word	*False Alarm: 0.301*	*Correct Rejection: 0.699*

3.2.4.3 Bias

In order to explore whether listeners had some bias in favor of a particular response in the two tasks, bias measures were calculated. Bias is calculated simply by dividing the total proportion correct responses by 2. No bias is 0.5. A bias closer to 0 or 1 indicates more bias toward that particular response.

In the first task, the bias toward IP Boundary response is 0.575 ((*Hit+False Alarm)/(Hit+False Alarm+Miss+Correct Rejection))*. The results revealed that listeners did not have a strong bias toward either response since the bias measure was close to 0.5.

In the second task, the bias toward TSG Boundary response is 0.562. Likewise, there was no strong bias shown toward either of the responses.

3.2.4.4 Interim summary

The correctness in Part I reveals that listeners are more accurate at identifying bigger boundaries (i.e. IP Boundary and TSG Boundary). The sensitivity as well as the bias results suggested that they could distinguish the bigger boundaries from the smaller boundary (i.e. Word Boundary) and that they did not have a strong bias toward a particular boundary response.

One thing worth remembering was that the lexical information in the stimuli was all removed. The filtered speech hid the lexical information from the listeners; however, their responses suggest that listeners could accurately identify the sizes of the boundaries even in a degraded signal.

3.3 Part II: Comparison between TSG Boundary and IP Boundary

3.3.1 Method

3.3.1.1 Participants

The same 30 adult native speakers of Taiwanese participated in Part II. None reported hearing or vision problems and all were compensated for their participation with ten dollars. Only 28 participants' results were analyzed because of the two subjects' failure to complete the experiment within the allotted time.

3.3.1.2 Stimuli

The stimuli in this experiment involved sequences with greater length and more complicated structures, as shown in (5). In (5a), the only TSG covers the whole sentence and so the following boundary is an IP. In (5b), the same TSG covers the subject, verb and partial object of the relative clause in the sentence, and the boundary is merely a TSG. This pair of sentences was to examine whether listeners were able to know what to expect after they sequences which were potentially complete utterances – the end of a sentence, as in (5a), or part of the relative clause, as in (5b). During the experiment, the listeners would read both (5a) and (5b) on the laptop screen, and were only presented with highlighted audio utterance from one of the sentences.

(5) a. IP

thou55-san55 tiam31 u31 teh5 be31 be31-ge33-ko55 #

souvenir store have ASP sell maltose syrup

" The souvenir store sells maltose syrup."

b. TSG

thou55-san55-tiam31 u31 teh5 be31 be31-ge33-ko55 # cho31 # e33 thng33-a51 #

souvenir store have ASP sell maltose syrup make GEN candy

"The souvenir store sells candy made of maltose syrup."

3.3.1.3 Procedures

The experiment comprised two sessions. The first session presented the normal speech stimuli, and the second session presented the filtered speech stimuli. Results from both sessions were analyzed Listeners were told that they would hear a series of Taiwanese utterances over noise-cancelling headphones and would be offered two alternatives on the laptop screen for each utterance. They were asked to respond with their initial intuitions and choose "which highlighted part the speaker said." They could only hear each utterance once.

3.3.2 Analysis

Listeners' responses to both the normal speech and the filtered speech were analyzed. The specific measures were correctness, sensitivity (d' score) and bias. As in Part I, the bigger boundary, which is IP Boundary here, is considered the 'signal' whereas TSG boundary was considered the 'noise'. The d' score and bias for each condition were calculated by the following formula:

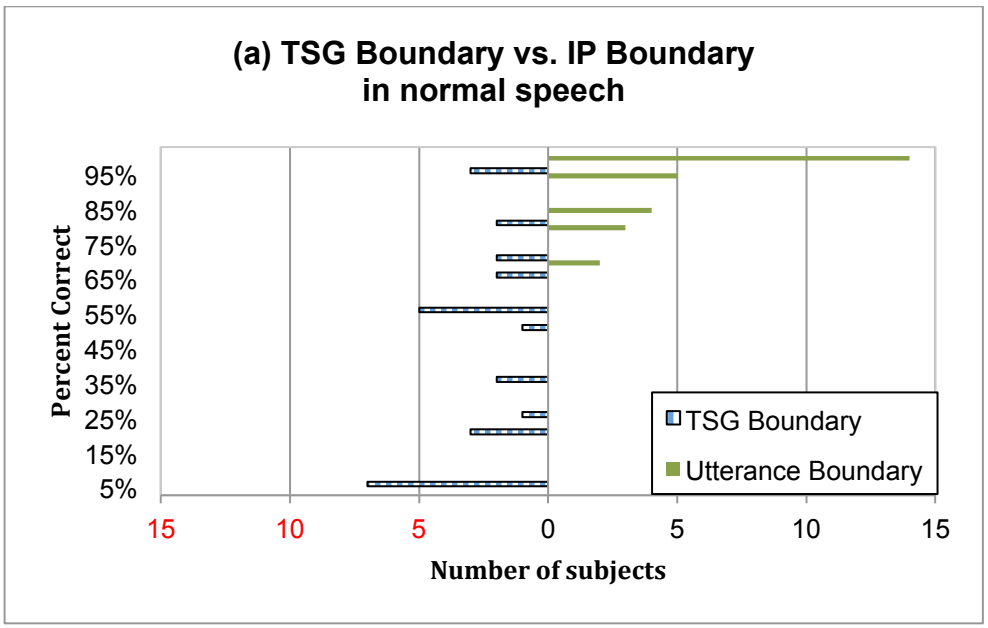
(6) d' = *Hit-False Alarm*

(7) Bias measure = *(Hit+False Alarm)/(Hit+False Alarm+Miss+Correct Rejection)*

3.3.3 Results

3.3.3.1 Accuracy

The accuracy results are shown in Figure 3-2. Figure 3-2 (a) presents the percentages of correctly chosen TSG Boundary and IP Boundary in normal speech. Figure 3-2 (b) shows the comparison in filtered speech.



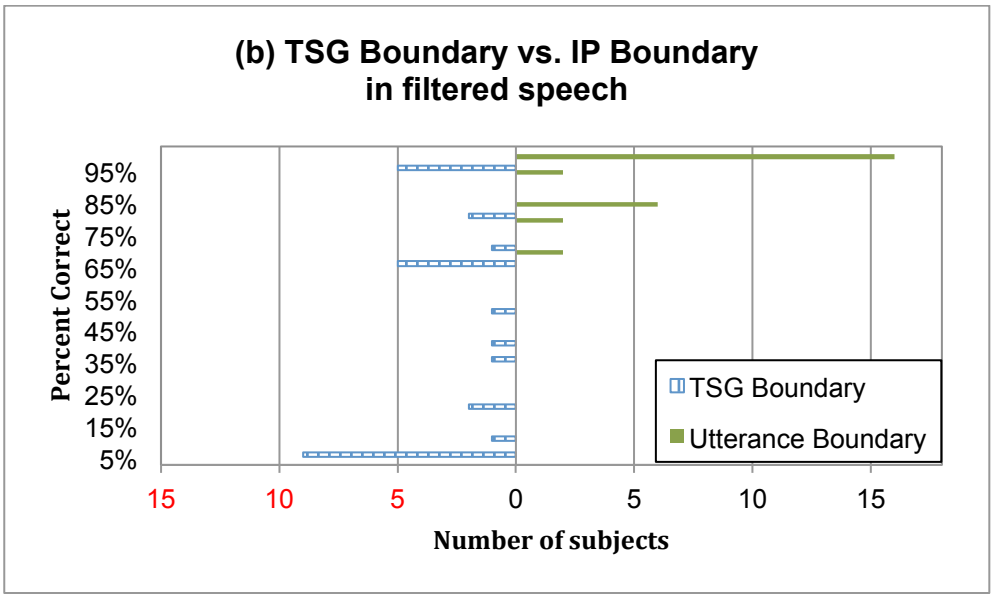


Figure 3-2. Double histograms of participants' responses in Part II: (a) TSG Boundary vs. IP Boundary in normal speech (b) TSG Boundary vs. IP Boundary in filtered speech. In each graph, the y-axis is the averaged percent correct for each subject, and the x-axis shows the number of subjects obtaining that percent correct. For instance, five subjects in (b) had 95% correct for TSG boundary.

3.3.3.1.1 Normal stimuli

Listeners did not choose TSG boundary over IP boundary when given TSG stimuli (41.2%: *t*-test against chance: $t(27)=1.44$, $p=0.2$). However, when they were presented with the IP boundary, listeners chose IP boundary over TSG boundary 91.8% of the time, which is above chance (*t*-test against chance: $t(27)=21.58$, $p<.001$). Listeners correctly chose IP boundary significantly more than TSG boundary (paired *t*-test: $t(27)=7.93$, $p<.001$).

3.3.3.1.2 Filtered stimuli

The results of the filtered stimuli are very similar to that of the normal stimuli. Listeners did not choose TSG boundary over IP boundary when given TSG stimuli (40.9%: *t*-test against chance: $t(27)=1.31$, $p=0.2$). However, listeners chose IP boundary over TSG boundary when they were presented with the IP 92.3% of the time, which is above chance (*t*-test against chance: $t(27)=21.83$, $p<.001$). Listeners chose correct IP boundary significantly more than TSG boundary (paired *t*-test: $t(27)=6.67$, $p<.001$).

In both normal and filtered speech stimuli, listeners were more accurate choosing IP, and their choice for correct TSG boundaries was below chance, which can be observed from their diverse choices for TSG boundaries.

3.3.3.2 Sensitivity (d' score)

The detection outcomes are shown in Table 3-3 and Table 3-4. IP is considered the 'signal', and TSG is the 'noise'. The d' scores were both the normal speech and the filtered speech were calculated based on the values in the tables.

Table 3-3. Response Proportions: IP vs. TSG in normal speech

	Response: IP	**Response: TSG**
Stimuli: IP	*Hit: 0.918*	*Miss: 0.082*
Stimuli: TSG	*False Alarm: 0.588*	*Correct Rejection: 0.412*

Table 3-4. Response Proportions: IP vs. TSG in filtered speech

	Response: IP	**Response: TSG**
Stimuli: IP	*Hit: 0.923*	*Miss: 0.077*
Stimuli: TSG	*False Alarm: 0.591*	*Correct Rejection: 0.409*

The d' was then calculated by subtracting the *False Alarm* ratio from *Hit* ratio. In the normal speech, the d' score is 0.33. The d' score in the filtered speech is 0.332. That is, once response bias is factored out, the listeners are shown to be not especially sensitive to the Boundary difference in either normal speech or filtered speech.

3.3.3.3 Bias

In order to verify that listeners had some bias in favor of a particular Boundary response, the bias measures were calculated. It is calculated simply by dividing the percent correct of their responses by their overall responses. No bias is 0.5. A bias closer to 0 or 1 indicates more bias toward that particular response.

In the normal speech, the bias toward IP Boundary response is 0.753 ((*Hit+False Alarm)/(Hit+False Alarm+Miss+Correct Rejection))*. The results revealed that listeners have stronger bias toward the IP Boundary response.

In the filtered speech, the bias toward IP Boundary response is 0.757. Similarly, there was a strong bias for the IP boundary response.

3.4 Summary

The aims for this chapter were to examine: (i) How well do Taiwanese listeners do when they are asked to identify an upcoming boundary in utterances bigger than two syllables? (ii) How well do they do with filtered speech?

Part I included two tasks asking listeners to distinguish IP boundary from Word boundary, and TSG boundary from Word boundary in filtered speech. The results showed that listeners are more accurate at identifying bigger boundaries (i.e. IP Boundary and TSG Boundary). In addition, listeners showed no bias toward either response in either task, and were fairly sensitive in identifying IP boundaries and TSG boundaries in the two tasks respectively.

Part II asked listeners to identify IP boundary from TSG boundary in normal speech, then in filtered speech. The results revealed that in both normal and filtered speech, the accuracy of the IP boundaries was higher because the listeners had greater response bias toward IP, they were relatively insensitive to the boundary difference.

Therefore, to answer the question in (i), it seems that listeners were able to distinguish Word boundaries from bigger boundaries, such as TSG and IP, by being more sensitive to bigger boundaries. However, they were relatively insensitive to the TSG vs. IP distinction. One possibility to this is that the task in Part II is a relatively complicated task since the TSG condition could potentially be a sentence itself. If this is true, then it suggests that TSG in the TSG condition is not different from the TSG in the IP condition.

Listeners' responses in Part 1 shows that the degraded signal did not stop them from giving accurate responses. It is more obviously seen in Part II where the responses to normal speech and filtered speech were compared and found to be similar. It is known that filtered speech has removed all the segmental information and the cues left were duration and pitch. Taiwanese

listeners use pitch in lexical distinction, however, this did not make them less able to participate in prosodic groupings.

Chapter 4: Perceived Boundary Strength in Taiwanese

4.1 Introduction

4.1.1 Background

During sentence processing, listeners need not only process the meaning of each lexical item, but also the ordering and arrangement of these items. The latter is about the hierarchical structure of the sentence, including both the syntactic hierarchy and the prosodic hierarchy. One way to examine how listeners process the prosodic hierarchy is to observe their detection of the prosodic boundaries. In this chapter, the research question is whether people are able to predict the upcoming boundaries at different prosodic levels in *spontaneous speech*, and if they are, which acoustic correlates are propitious for boundary detection. A similar question was asked in Chapter 3. However, there are several differences in the two experimental designs: (a) in Chapter 3, the stimuli were from read speech where as the stimuli in this chapter were from spontaneous speech. (b) The response in Chapter 3 required a categorical choice whereas this chapter examines a continuous decision made by the participants.

Previous studies have shown that read speech and spontaneous speech differ in the distribution and realization of prosodic boundaries. For example, Blaauw (1994) found that for minor boundaries, though lengthening was observed in both spontaneous speech and read speech, a combination of lengthening and pause only appeared in spontaneous speech. Phrase-internal boundaries were characteristic of spontaneous speech. For major boundaries, a falling boundary tone was more frequently observed in read than in spontaneous speech. In the current study, the stimuli are taken from spontaneous speech, and the three boundaries under study are Word/Syllable boundary, phrase boundary/TSG boundary, and sentence/IP boundary.

For boundary cues, most previous studies of speech prosody have focused on duration and F0. Only a few (e.g. Choi et al. 2005) analyzed voice source parameters in connected speech. In this study, we not only examine listeners' use of duration and F0, but also their use of other voice cues in order to see whether and to what extent do listeners' judgments depend on voice quality.

Previous research has shown that listeners are able to predict upcoming prosodic boundaries. For example, Grosjean and Hirst (1996) had the subjects listen to some part of an English sentence and asked them to predict how long the remaining sentence was (see next chapter for detailed review). The results showed that English listeners were very accurate at predicting the amount of the rest of the sentence, but French listeners could only tell if a sentence ended, unable to differentiate between different amounts to come.

4.1.2 Swedish Boundary Detection: Carlson, Hirschberg, and Swerts (2005)

Carlson, Hirschberg, and Swerts (2005) had a perceptual study on listeners' predictions of the upcoming boundary strengths in Swedish. Their stimuli were taken from the spontaneous speech of a Swedish female politician as she was being interviewed. The whole interview was prosodically labeled by three researchers who relied on their perception instead of any visual tool to do the labeling (Strangert and Heldner 1995 for more detail). Carlson *et al.* (2005) selected their stimuli fragments of speech that were followed either by a strong break (i.e. sentence boundary), a weak break (i.e. phrase boundary) or no break (i.e. word boundary). All stimuli came in two lengths, 2-second fragments and one-word fragments. From each 2-second fragment, the last word was extracted to be the one-word fragment. There were 120 stimuli in total (20 items × 3 breaks × 2 lengths). Their subjects were asked to listen to each stimulus and express their judgment about the upcoming boundary on a 5-point scale. In other words, the

subjects' judgments were relatively continuous compared to the categorical choice in the previous chapters.

Their subjects came from two language groups – native speakers of Swedish and English. The English speakers knew no Swedish. Their boundary strength rating results are presented in Figure 4-1. The authors found that English listeners were able to predict the strengths of the upcoming boundaries as well as Swedish listeners, whether presented with a 2-second fragment or a one-word fragment. Since English listeners could not use the meanings of the fragments to predict possible phrasing, they must have made their judgments based on the prosodic cues about the phrasings contained in the signal – information that is not specific to Swedish apparently. However, a follow-up pilot study found that Mandarin listeners could hear different Swedish boundaries only when presented with 2-second fragments, indicating that language background affects the listeners' judgments. That is, there must be similar prosodic boundary cues in Swedish and English, but not all these cues are used in Mandarin. Only information over a longer span of speech is similar enough between Mandarin and Swedish to be useful to Mandarin listeners of Swedish.

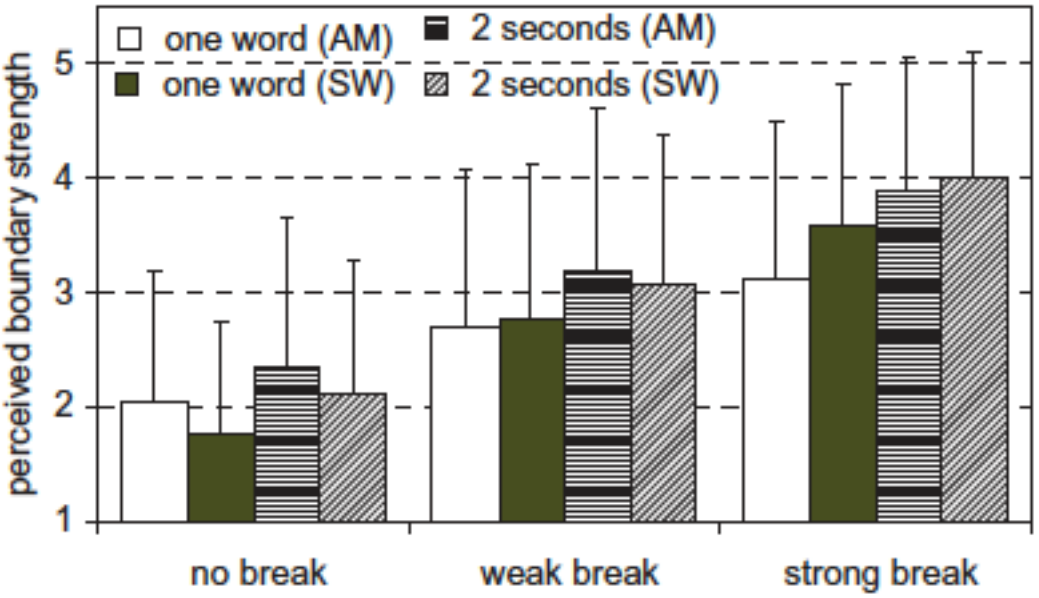


Figure 4-1. Mean perceived upcoming boundary strength. Adopted from Carlson *et al.* (2005), p 329.

Furthermore, Carlson *et al.* investigated the prosodic cues that listeners used by testing for correlations between acoustic measures of the one-word fragments in the stimuli and the ratings of the Swedish and English listeners. There were significant correlations between ratings and median F0, and between ratings and the presence of final creak. In addition, the fact that the crucial information for Mandarin listeners (whatever it may be) is not found in the pre-boundary word suggests that they pay less attention to F0 cues; as tone language listeners, they might be attending to F0 only for lexical tone recognition.

In the current study, we replicate and extend Carlson *et al.*'s (2005) experiment. The entire stimulus set consisted of both Carlson *et al.*'s Swedish stimuli, and similar Taiwanese stimuli, plus low-pass-filtered versions of these. American English listeners and Taiwanese listeners were recruited to participate in a rating task. The experiment was like Carlson *et al.*'s except that the rating scale was more continuous. The results of the rating task and the acoustic measures of the

stimuli will allow us to see whether and how the prosodic boundaries in Taiwanese and Swedish were perceived by Taiwanese listeners and English listeners.

4.2 Method

4.2.1 Stimuli

The Swedish stimuli were the same stimuli Carlson *et al.* used in their experiment[8]. As mentioned earlier, 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 a strong or weak break or as not followed by a phrasal boundary (i.e. no break). If there was a disagreement about a perceived boundary, a majority voting strategy was used. To make sure that the fragments were all from a comparable syntactic position, the selected fragments were all followed by the word "*och*" (and) in the original context, and were cut before the silent pause (if any) preceding "*och*". In their paper, they did not specifically mention paying attention to lexical pitch accent when they selected the stimuli for their experiment. The 60 utterances had three boundary types (based on their labeling): 20 word boundaries (labeled as "no break"), 20 phrase boundaries ("weak break") and 20 IP boundaries ("strong break").

In order to make comparable Taiwanese stimuli, utterances (n=60) were extracted from an interview with a female Taiwanese preacher who was recorded for this experiment. As in Carlson *et al.*, boundary presence and boundary strength were labeled by the author and another researcher based on the criteria in TW-ToBI developed by Peng and Beckman (2003). These stimuli also had three boundary types – here, word boundary ("no break"), TSG boundary ("weak break") and IP boundary ("strong break"). It is known that a tone sandhi group can cover

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

as much as a whole sentence. To make sure that "strong break" did have a bigger boundary, all the stimuli with the IP boundary were followed by a pause, which in turn was followed by a sentence-initial particle '*a0*'[9] as the beginning of the sentence that came after. Also as in Carlson *et al.*, the stimuli came in two lengths, either an approximant 2-second fragment or a one-word fragment. Similarly, each one-word fragment was the very last word in the corresponding 2-second fragment. In Taiwanese, each word is one syllable long, hence the one-word fragments are fragments with only one syllable. However, in Swedish, a word could have more than one syllable, so the one-word fragments are longer than in the Taiwanese stimuli. In fact, the average length of the one-word fragments in Swedish is 506 msec and the average length of the one-word fragments in Taiwanese is 214 msec.

Thus, both the Swedish and Taiwanese stimuli are taken from recordings of natural speech, without any manipulations such as tapering the amplitudes at the edge. One of the implications in Carlson *et al.*'s study was that, since English listeners could accurately predict the boundaries in Swedish, they can use prosodic information instead of semantic/lexical information. However, as long as both segmental and suprasegmental cues are available, it is impossible to say what information the native listeners are using. Therefore, in the current study, filtered versions of the Swedish and Taiwanese stimuli were used in addition to the normal (full-frequency) stimuli. The filtered stimuli were generated by low-pass filtering the normal stimuli at a frequency cut-off of 400 Hz and 50Hz smoothing, and the intensity was adjusted to 70 dB. The entire manipulation was done with a Praat script. With low-pass filtering, most of the segmental information will be removed, yet the prosodic information, such as duration and F0, will remain intact.

Therefore, there were 480 utterances in total (20 items × 3 breaks × 2 fragment lengths × 2 speech qualities × 2 languages). Each language had 20 items with "no break", 20 with "weak

[9] "*a*0" is a sentence-initial particle whose carries a neutral tone. The neutral tone was transcribed with 0.

break" and 20 with "strong break". Each item had a 2-second fragment and a one-word fragment, with each one-word fragment being the end of a 2-second fragment. Each fragment had a normal- quality and a low-pass-filtered version.

4.2.2 Subjects

Eighteen Taiwanese native speakers and eighteen American English native speakers participated in this experiment. The Taiwanese native listeners were undergraduate students recruited at National Kaohsiung Normal University in Taiwan, with no previous knowledge about Swedish. The American English listeners were undergraduate students recruited at UCLA and had no prior experience with Swedish, nor with Taiwanese or any other tone language. They received either a monetary compensation or class credit for their participation. None of them had hearing or language problems according to their self-report.

4.2.3 Procedures

The experiment was implemented with a Matlab PsychoToolbox program on a PC or an iBook. The subject 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 phase, in which the experimenter made sure that each subject fully understood the task. The experimenter gave verbal instructions in Mandarin Chinese when speaking to the Taiwanese listeners, and she used English when speaking to the

English listeners. The instructions included examples of a word boundary, phrase/tone sandhi group boundary and intonation phrase boundary in the listeners' native language. The English version of the instructions is given in Appendix A. Before they started the task, the subjects read the same instructions presented on the laptop screen again.

During the task, the subjects could choose to hear each stimulus more than once, but were encouraged to make their judgments by instinct. Listeners had to provide a judgment in order to proceed to the next trial. No feedback was given on their responses. The position of the slider bar was recorded by the Matlab script on a scale from 0-100 though the listeners only saw the two words "small" and "big" at the two bar ends, as shown in Figure 4-2. These numerical values were the responses that were analyzed.

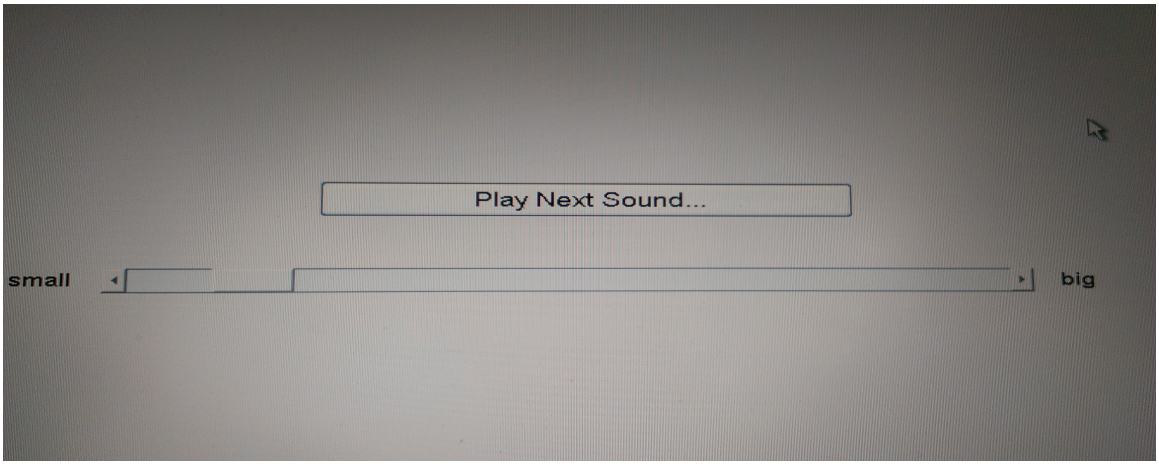


Figure 4-2. A screenshot of what the listeners saw during the task.

4.2.4 Acoustic measures

In an attempt to identify the prosodic cues that could contribute to accurate boundary strength judgments, we examined six categories of acoustic measures from the *last syllable* of each stimulus: duration, pitch, harmonic amplitude/spectral tilt, harmonic-to-noise ratios, CPP and

Energy. The last syllable of each stimulus was labeled in Praat (Boersma et al. 2012) and then the acoustic measures for the labeled portions were obtained using VoiceSauce (Shue *et al.* 2011). For the Swedish stimuli, the last syllable was usually part of a word, whereas the last syllable in the Taiwanese stimuli was an entire word.

Durational measurements include normalized vowel duration and speech rate. Vowel duration was normalized for two reasons (i) the absolute duration might vary with the vowel height (i.e. a low vowel is longer than a high vowel); (ii) according to Tseng (1995), the duration of a Taiwanese checked tone is about 1/3 to ¼ of the duration of a Taiwanese non-checked tone. Therefore duration was normalized with respect to vowel height and syllable structure. Previous research on English (Wightman *et al.* 1992) has found that "segmental lengthening in the vicinity of prosodic boundaries is restricted to the rhyme of the syllable preceding the boundary." In their study, nucleus and coda showed final lengthening individually, instead of the entire rime. Therefore, vowel duration was the main duration measure used in this study. Normalized vowel duration was calculated based on observed segment durations, using the normalization method employed in Wightman *et al.* (1992). The phone-based normalization formula is given in (1), where *d(i)* is the vowel duration of word *i*, and μ_p and σ_p are the mean and standard deviation of the duration of that vowel segment observed over the *entire* set of Swedish stimuli or Taiwanese stimuli.

$$\tilde{d}(i) = [d(i) - \mu_p] / \sigma_p \quad (1)$$

Speech rate was the other duration measure used in this study: it was calculated as the reciprocal of the raw total rime duration (vowel duration + coda duration if there is any). 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 range, F0 slope, F0 median, and F0 mean measured in the vocalic rime. F0 range was calculated as the difference between the maximum and minimum raw F0 over the vocalic rime. F0 slope was calculated by dividing the F0 range by the raw duration of the vocalic rime. Pan (2006) found that F0 range was greatest before an IP boundary, followed by a WRD boundary, and finally TSG boundary. In addition, F0 slope was slowest before an IP boundary, followed by TS, then WRD, and finally SYL. Therefore, in the present study, F0 slope is expected to decrease as the boundary becomes bigger, and F0 range is expected to be the widest before the IP boundary, and narrowest before the TSG boundary. F0 mean and F0 median were the average and the midpoint values across each rime.

The harmonic amplitude/spectral tilt measurements include H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, and H1*-A3*. The values presented here were all formant-corrected values (indicated by *) obtained from VoiceSauce (Shue *et al.* 2011). By using corrected values, different vowels were comparable. H1*-H2* refers to the difference in amplitude between the first and second harmonics. A lower value of H1*-H2* is often correlated with a more constricted glottis in a laryngealized voice, such as creaky voice, while a higher value is found in breathy voice (Shue *et al.* 2011, Gordon & Ladefoged 2001, Hanson *et al.* 2001, Blankenship 2002). A1, A2 and A3 are the amplitudes of the harmonics nearest 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. The basis of H2*-H4* is uncertain in the literature, and it is used here for future reference.

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

CPP (Cepstral Peak Prominence) is another harmonic-to-noise measure, which has been shown to correlate with the perception of breathiness by English listeners (Hillenbrand and Houde 1996). Some speakers tend to have a breathy voice instead of a creaky voice when approaching a prosodic boundary. In this case, CPP could be an appropriate measure for the distinction. CPP is predicted to show smaller values if the signal tends to be breathier.

Energy is a measure of the intensity in a series of windows across each utterance, and it is expected to have a smaller value closer to a prosodic boundary.

The filtered speech yielded fewer measurements regarding voice quality. 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 4-3), 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; H1*), HNR05 (frequency range <500 Hz) and CPP.

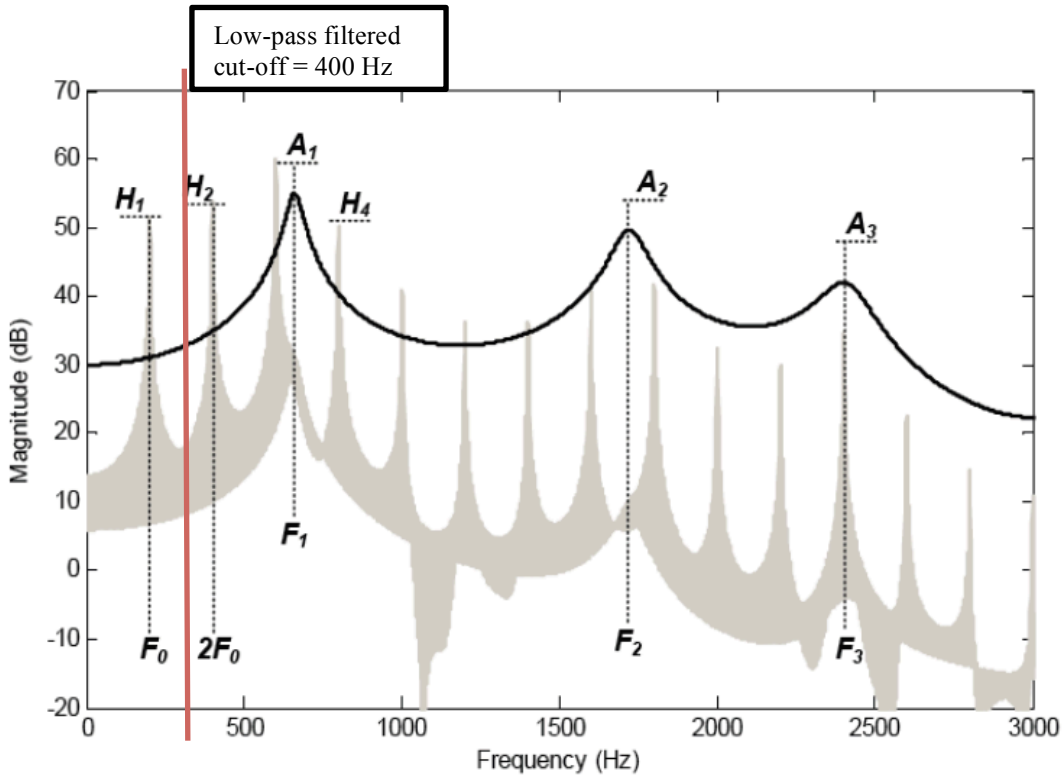


Figure 4-3. An example spectrum. Adopted from Shue (2010). Note that the envelope in shadow is the actual output from VoiceSauce. The corrected values are obtained from this envelope (i.e. corrected values).

4.3 Results

4.3.1 Responses

Listeners' perceptual judgments for boundary strength were converted into logarithmic strengths. The use of the logarithmically transformed strength ratings, rather than the raw ratings, reduces the skewing in the distribution. Without the transformation, the outliers might dominate the outcome, which might obscure the main trend in the data.

The strength ratings were entered into the repeated measures ANOVA analysis in R, with the between-subject factor "(listener's) native language" (Taiwanese listeners vs. English listeners) and the within-subject factors, "break" (no break vs. weak break vs. strong break), "stimulus language" (Swedish vs. Taiwanese) "speech quality" (normal vs. filtered) and "fragment length" (2-second vs. one-word). The results show that (i) there is no significant difference between the

ratings of the Taiwanese listeners and of the English listeners; (ii) there is no remarkable difference in rating for normal stimuli vs. filtered speech; (iii) there is also no difference between the ratings for the Swedish stimuli and the ratings for the Taiwanese stimuli. However, (iv) listeners tended to give higher average strength ratings for bigger boundaries F (2, 70) = 47.9, p < .05. Other than these main effects, some interaction effects were also found (as shown in Appendix B). The between-subject factor "native language" did not show main effect alone; however, its interactions with other factors are mostly significant.

In the following sections regarding the strength ratings, the results are presented by listeners' native language and stimulus language, so that the different predictions for the different languages can be addressed. It is predicted that when English listeners listened to the Swedish stimuli (as in section *4.3.1.1*), they should be able to differentiate the three breaks from one another in both 2-second and one-word normal fragments, as shown in Carlson *et al.*'s study. When Taiwanese listeners listen to the Swedish stimuli (section *4.3.1.2*), it is predicted that they could make such distinction only when presented with 2-second fragments, reminiscent of Carlson *et al.*'s pilot study with Mandarin speakers. The same prediction about English listeners listened to the Swedish stimuli was made for English listeners when they were presented with the Taiwanese stimuli (section *4.3.1.3*) because English listeners were expected to be good at predicting prosodic boundaries, even in a foreign language. When Taiwanese listeners were presented with the Taiwanese stimuli (section *4.3.1.4*), it is predicted that they could accurately predict the upcoming boundary only with the 2-second fragments with the assumption that tone language listeners require longer fragments to make accurate judgments.

4.3.1.1. English listeners listening to the Swedish stimuli.

These results are presented in Table 4-1 and Figure 4-4. For the Swedish normal stimuli, repeated measures ANOVA reveals significant effects of "break" [$F(2, 34)=25.28$, $p < .05$] and "length" [$F(1, 17)=8.47$, $p < .05$]. A Tukey HSD post hoc test for Break shows that their ratings for the three breaks are significantly different from one another ($p < .01$) and this three-way distinction can be found in the 2-second fragments as well as the one-word fragments. This distinction found in each length is illustrated with the lines and asterisks above the bars in the figure. In addition, the finding that English listeners gave significantly higher ratings for 2-second fragments than for one-word fragments in normal stimuli is the same as Carlson *et al.*'s finding.

For the Swedish filtered stimuli, significant effects were found not only in "break" [$F(2, 34) = 16.77$, $p < .05$] and "length" [$F(1, 17) = 12.07$, $p < .05$], but also in their interaction [$F(2, 34) = 9.01$, $p < .05$]. The Tukey HSD post hoc test for "break" shows that listeners' ratings for the three breaks are significantly different from one another, yet this distinction only held for 2-second fragments, but not for the one-word fragments. When presented with the one-word fragments, the English listeners only made a binary choice – the boundary was either big or small.

Table 4-1. Repeated measures ANOVA table for English listeners' strength ratings for Swedish stimuli.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
(a) normal					
break	2	63.14	31.57	25.28	< .01*
length	1	98.3	98.3	8.47	< .01*
break × length	2	0.58	0.29	1.24	0.30
(b) filtered					
break	2	19.73	9.86	16.77	< .01*
length	1	202.7	202.7	12.07	< .01*
break × length	2	2.26	1.13	9.01	< .01*

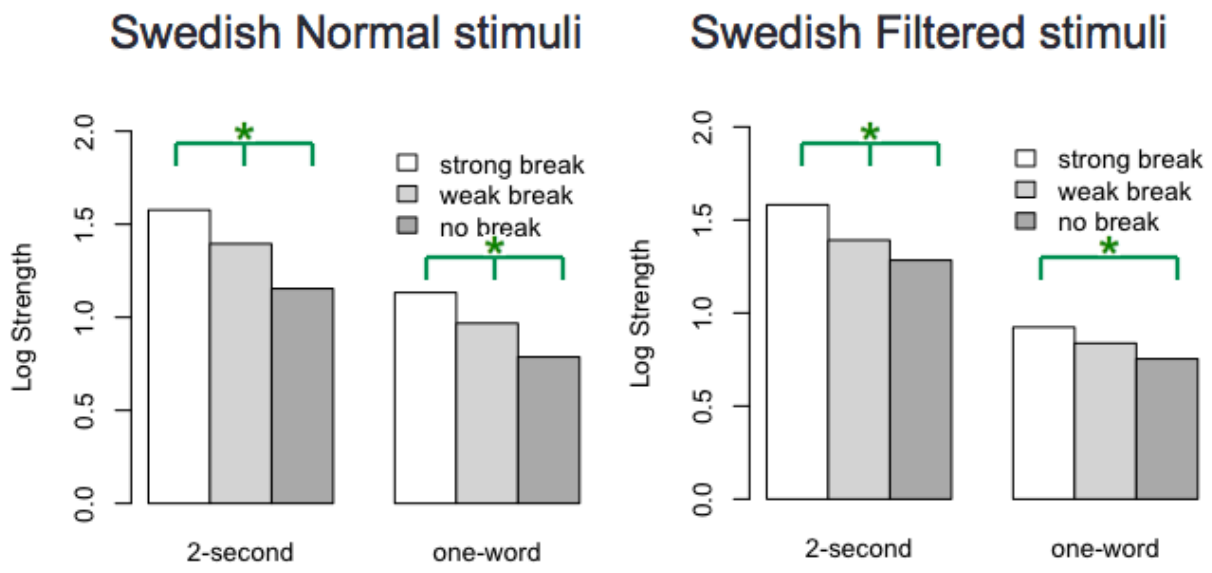


Figure 4-4. English listeners' average logarithmic perceived boundary strength for Swedish stimuli. A significant difference between "breaks" of either length is indicated with a line and asterisk above the bars.

4.3.1.2 Taiwanese listeners listening to the Swedish stimuli.

These results are presented in Table 4-2 and Figure 4-5. For the Swedish normal stimuli, significant effects were found of "break" [$F(2, 34)=18.33$, $p < .05$], "length" [$F(1, 17)=20.93$, $p < .05$], and their interaction [$F(2, 34)=4.89$, $p < .05$]. The Tukey HSD post hoc test of Taiwanese listeners' responses showed that the log strengths of the three breaks are different from one another ($p<.01$) but only for the 2-second fragments. Thus, Taiwanese listeners, like the Mandarin Chinese listeners in Carlson et al. (2005), tended to give different ratings for the three different breaks only when presented with 2-second fragments. When listeners heard one-word fragments, the "weak break" stimuli were heard like the "no break". The post hoc test for Length showed that Taiwanese listeners, like English listeners, gave higher ratings for longer fragments.

For the Swedish filtered stimuli, both "break" [$F(2, 34)=14.91$, $p < .05$] and "length" [$F(1, 17)=28.23$, $p < .05$] the showed significant effect for the boundary strength ratings. No significant interaction was found in the filtered speech. The Tukey HSD post hoc test of "break" reveals that the Taiwanese listeners had different ratings for the three different breaks in Swedish filtered speech, and this distinction came mainly from the 2-second fragments. However, the three breaks were not distinctive in the post-hoc comparison of the one-word fragments.

Table 4-2. Repeated measures ANOVA table for Taiwanese listeners' strength ratings for Swedish stimuli.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
(a) normal					
break	2	57.95	28.97	18.33	< .01*
length	1	93.38	93.38	20.93	< .01*
break × length	2	3.01	1.51	4.89	< .05*
(b) filtered					
break	2	15.27	7.63	14.91	< .01*
length	1	195.9	196	28.23	< .01*
break × length	2	3.22	1.61	3.22	0.0523

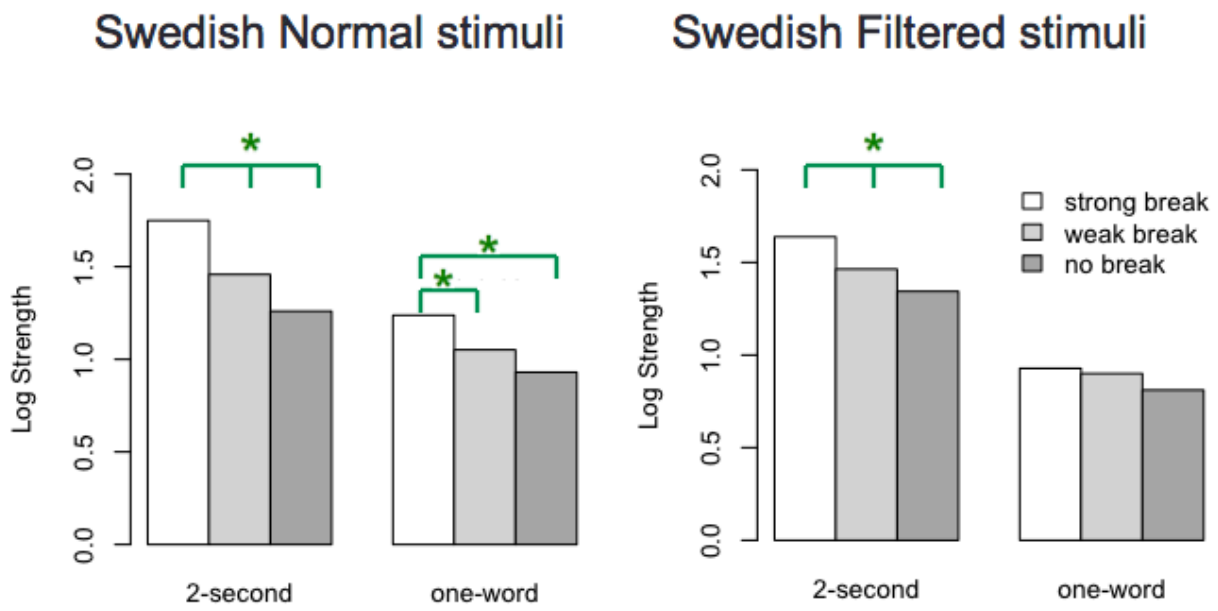


Figure 4-5. Taiwanese listeners' average logarithmic perceived boundary strength for Swedish stimuli.

4.3.1.3 English listeners listening to the Taiwanese stimuli.

The results are presented in Table 4-3 and Figure 4-6. For the Taiwanese normal stimuli, significant main effects were found in "break" ($F(2, 34)=10.21$, $p < .01$) and "length" ($F(1, 17)=16.09$, $p <.01$). No significant interaction was found. The Tukey HSD post hoc test on "break" revealed that English listeners did not give a distinctive rating for the "weak" break" vs. the "strong break", for either 2-second or one-word fragments. Their ratings for the 2-second fragments were significantly higher than those for the one-word fragments.

For the Taiwanese filtered speech stimuli, significant main effects of "break" ($F(2, 34)=13.64$, $p <.01$) and "length" ($F(1, 17)=19.25$, $p <.01$) were found. The Tukey HSD post hoc test on "break" showed that the English listeners did not distinguish the "weak break" and the "strong break" when presented with 2-second fragments. However, when they were presented with the one-word fragments, they used the same strategy as they listened to one-word fragments in Swedish filtered speech – they heard either a big break or a small break. Again, the ratings for

the 2-second fragments were significantly higher than the ratings for the one-word fragments. In a word, for English listeners, the Taiwanese TSG boundary is the same as the IP boundary, and except in the one-word filtered fragments, TSG and IP boundaries are both different from Word boundary.

Table 4-3. Repeated measures ANOVA table for English listeners' strength ratings for Taiwanese stimuli.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
(a) normal					
break	2	11.87	5.94	10.21	< .01*
length	1	228.2	228.19	16.09	< .01*
break × length	2	0.64	0.32	1.70	0.20
(b) filtered					
break	2	10.13	5.06	13.64	< .01*
length	1	299.4	299.37	19.25	< .01*
break × length	2	0.08	0.04	0.24	0.79

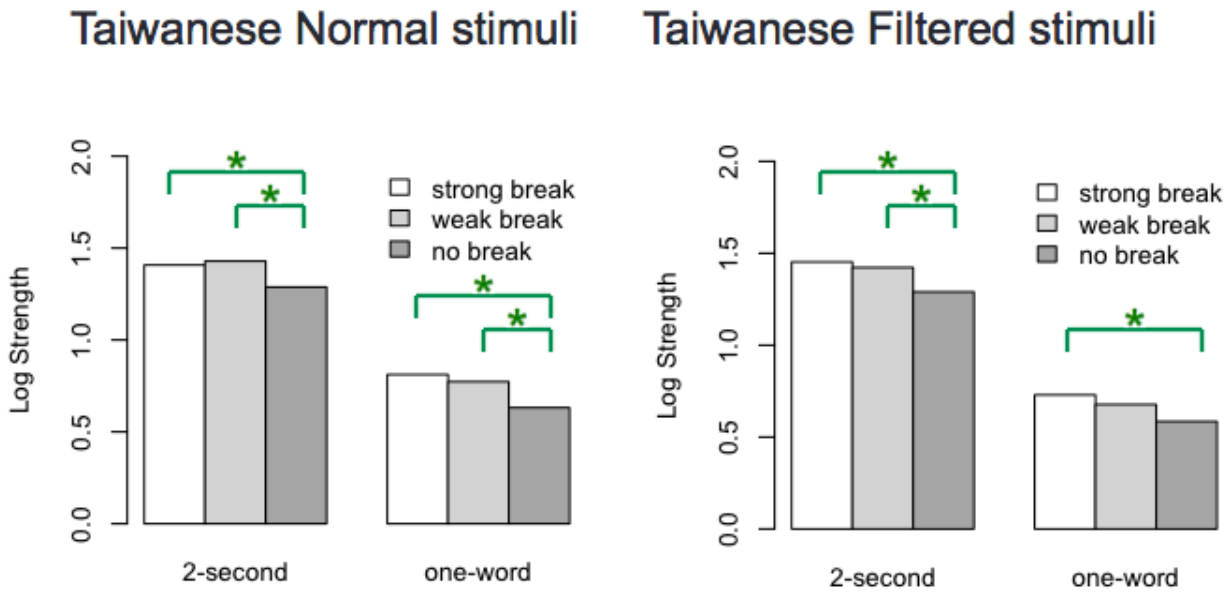


Figure 4-6. English listeners' average logarithmic perceived boundary strength for Taiwanese stimuli.

4.3.1.4 Taiwanese listeners listening to the Taiwanese stimuli.

The results are presented in Table 4-4 and Figure 4-7. For the Taiwanese normal stimuli, a two-way repeated measures ANOVA revealed significant main effects of "break" (*F*(2,

34)=22.12, p<.05) and "length" (F(1, 17)=35.58, p <.05), as well as of their interaction (F(2, 34)=11.74, p<.05). The Tukey HSD post hoc test of the Taiwanese listeners' responses shows that the log strengths of the three breaks are different from one another (p<.01) and the difference came from the 2-second fragments only. When they listened to one-word fragments, they gave the same rating for those with "weak break' and those with "no break". This is exactly the same as the finding we obtained in section *4.3.1*.2, when they listened to the Swedish normal stimuli. Thus, Taiwanese listeners seemed to use the same strategy for Swedish and Taiwanese normal stimuli when they were asked to predict the upcoming boundary. Their ratings for the 2-second fragments were significantly higher than those for the one-word fragments. In a word, Taiwanese listeners can hear a three-way distinction, but only in 2-second normal speech; this indicates that crucial information is spread out in time and frequency. More specifically, information about TSG vs. Word boundary is spread out in time, and information about TSG vs. both Word and IP is spread out in frequency.

For the Taiwanese filtered speech stimuli, significant main effects were found in "break" (F(2, 34)=10.54, p <.05) and "length" (F(1, 17)=25.54, p <.05). No significant interaction was found. The Tukey HSD post hoc tests on "break" reveals that Taiwanese listeners couldn't distinguish the weak break from the other two breaks in terms of their ratings with either length of fragment. Therefore, with the segmental and other high-frequency information removed. Taiwanese listeners only used binary choice to determine the upcoming boundary. Again, their ratings for the 2-second fragments were significantly higher than those for the one-word fragments.

Table 4-4. Repeated measures ANOVA table for Taiwanese listeners' strength ratings for Taiwanese stimuli.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
(a) normal					
break	2	66.13	33.07	22.12	< .01*
length	1	432.6	432.6	35.58	< .01*
break × length	2	12.62	6.31	11.74	< .01*
(b) filtered					
break	2	10.40	5.20	10.54	< .01*
length	1	313.9	313.9	25.54	< .01*
break × length	2	0.62	0.31	1.25	0.30

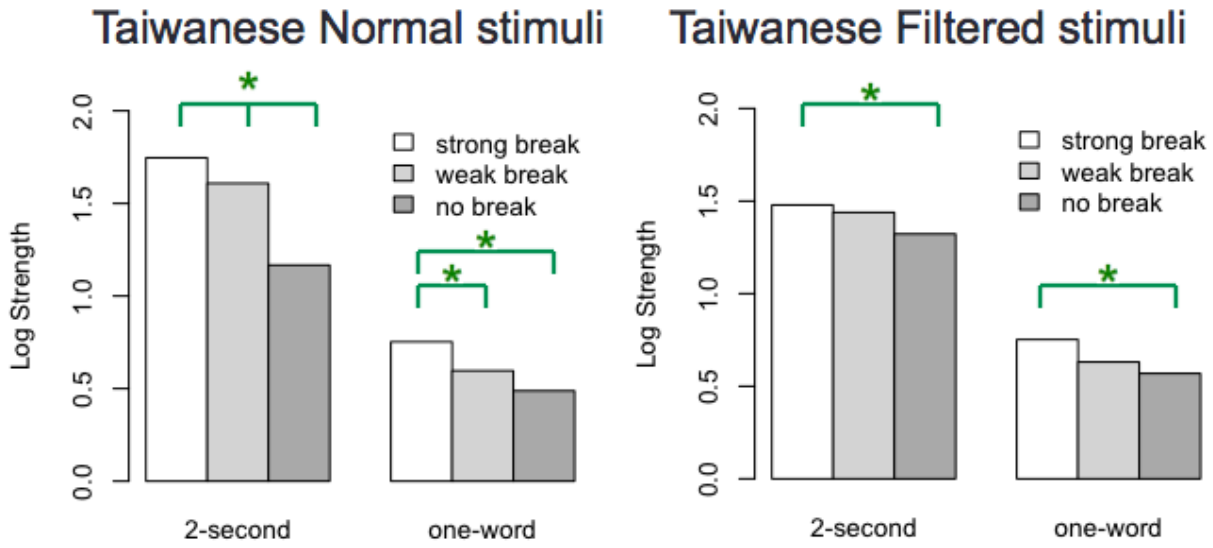


Figure 4-7. Taiwanese listeners' average logarithmic perceived boundary strength for Taiwanese stimuli.

4.3.2 Acoustic differences

The acoustic measures on normal speech were made on the rime of the last syllable. Recall that the Swedish words were often longer than one syllable, while the Taiwanese words were always one syllable, so we examine only the last syllable in both languages. Table 4-5 presents the ANOVA results for the acoustic measures with "break" as the predictor variable. The results

revealed that except for H1*-A1* in the Taiwanese stimuli, all the other 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 and Swedish stimuli, Tukey HSD post hoc tests were performed. H1*-H2*, H2*-H4*, HNR15 and CPP were good predictors of the three-way distinction in the Taiwanese stimuli. On the other hand, all the F0 measures and most of the HNRs were good predictors in the Swedish stimuli. In other words, for the Taiwanese stimuli, the bigger boundary tended to show greater creakiness (H1*-H2*), and greater aperiodicity or noise (HNR15 and CPP). For Swedish stimuli, the presence of the bigger boundary tended to show not only wider F0 range, larger F0 mean, larger F0 median and steeper F0 slope, but also greater aperiodicity (HNRs).

Table 4-5. ANOVA results for the acoustic measures of the last syllable in the stimuli in normal speech. $p < .05$ is marked with a tick (✔). If the significant acoustic measures showed a three-way distinction in terms of "break", then it is marked with a double tick (✔✔).

Measures	Taiwanese stimuli	Swedish stimuli
Normalized duration	✔	✔
Speech rate	✔	✔
F0 range	✔	✔✔
F0 slope	✔	✔✔
F0 mean	✔	✔✔
F0 median	✔	✔✔
H1*-H2*	✔✔	✔
H2*-H4*	✔✔	✔
H1*-A1*		✔
H1*-A2*	✔	✔
H1*-A3*	✔	✔
HNR05	✔	✔
HNR15	✔✔	✔✔
HNR25	✔	✔✔
HNR35	✔	✔✔
CPP	✔✔	✔✔
Energy	✔	✔

4.3.3 Multiple linear regressions

4.3.3.1 English listeners listening to the Swedish and Taiwanese normal stimuli

In order to see which acoustic measures contribute to the prediction of the boundary size, regressions between the acoustic measures and the logarithmic boundary strength ratings are made. Table 4-6 shows which of the acoustic measures contributed significantly to the regression equation (p <.05) and what proportion of the variance in the boundary strength ratings could be explained by the acoustic measures together. The multiple regression analysis was carried out for each type of stimulus (considering the stimulus language, speech quality and fragment length) separately.

For English speakers listening to normal speech, the proportion of explained variance in the acoustic measures was higher for Swedish stimuli than for Taiwanese stimuli. Speech rate was a useful measure in the normal stimuli in both languages, which suggest that English listeners paid particular attention to final lengthening. In addition to speech rate, correlations were found between listeners' ratings and f0 slope, creakiness (H1*-A2*) and Energy for Taiwanese normal stimuli. On the other hand, there were correlations between their ratings and F0 range and aperiodicities (HNRs) when they were presented with the Swedish stimuli. English listeners showed a three-way distinction of breaks in ratings when presented with 2-second fragments (as shown in section 4.3.1.1) in either language. The regression results here suggest that Speech Rate, aperiodicities (HNR25, HNR35) and Energy were cues listeners consistently rely on in 2-second stimuli. Almost all cues could explain some of the variance in boundary strength rating when they were entered into the regression equation, but the total variance accounted for is very small.

Table 4-6. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese normal stimuli; marks show which acoustic measures contributed significantly to the regression equation; the proportion of explained variance (R^2) is given, and the significance of this result is indicated.

	English listener			
	Swedish stimuli		Taiwanese stimuli	
	2-sec	1-wrd	2-sec	1-wrd
Duration				
Rate	✔	✔	✔	✔
F0 range	✔	✔		
F0 slope		✔	✔	✔
F0 median	✔			✔
F0 mean	✔			✔
H1*-H2*		✔		
H2*-H4*				
H1*-A1*			✔	
H1*-A2*		✔	✔	✔
H1*-A3*				
HNR05				✔
HNR15	✔	✔		✔
HNR25	✔	✔	✔	
HNR35	✔	✔	✔	
CPP		✔		
Energy	✔		✔	✔
R^2	0.13	0.09	0.05	0.04
F	19.38	10.97	7.99	5.21
p	< .05	< .05	< .05	< .05

4.3.3.2 English listeners listening to the Swedish and Taiwanese filtered stimuli

The multiple regression analysis results on English subjects listening to filtered speech stimuli are presented in Table 4-7. When English subjects listened to filtered speech, the proportion of explained variance in each speech type was consistently no more than 5%. F0 mean, H1* and HNR05 explain some of the variance in Swedish stimuli, whereas speech rate, F0 slope, F0 median and H1* were cues that could explain some of the rating variance in Taiwanese stimuli. It seems that duration and F0 cues were not as powerful as the voice cue H1*. English listeners

rated the boundary strength in filtered speech based on their perception of the voice quality instead duration nor pitch.

Table 4-7. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese filtered stimuli; marks show which acoustic measures contributed significantly to the regression equation; the proportion of explained variance (R^2) is given, and the significance of this result is indicated.

	English listener			
	Swedish stimuli		Taiwanese stimuli	
	2-sec	1-wrd	2-sec	1-wrd
Duration			✔	
Rate	✔		✔	✔
F0 range				✔
F0 slope			✔	✔
F0 median			✔	✔
F0 mean	✔	✔		
H1*	✔	✔	✔	✔
HNR05	✔	✔		
CPP				
R^2	0.05	0.04	0.05	0.03
F	13.82	15.96	11.47	6.19
p	< .05	< .05	< .05	< .05

4.3.3.3 Taiwanese listeners listening to the Swedish and Taiwanese normal stimuli

Table 4-8 shows which acoustic measure contributed significantly to the boundary strength ratings for Taiwanese listeners. The proportion of variance accounted for was slightly bigger for the longer fragments than the shorter fragments. It seems that F0 range contributes to the explanation across stimulus languages and the fragment lengths. Other than F0 range, for Swedish stimuli, F0 median, HNR25 and Energy were particularly helpful in that they could explain the variance in both long and short fragments. On the other hand, F0 range, F0 slope and HNR05 were the cues that help explain the distribution of the boundary strength ratings in Taiwanese stimuli.

Table 4-8. Results of multiple regression analysis for Taiwanese listeners listening to the Swedish and Taiwanese normal stimuli; marks show which acoustic measures contributed significantly to the regression equation; the proportion of explained variance (R^2) is given, and the significance of this result is indicated.

	Taiwanese listener			
	Swedish stimuli		Taiwanese stimuli	
	2-sec	1-wrd	2-sec	1-wrd
Duration			✔	
Rate	✔			✔
F0 range	✔	✔	✔	✔
F0 slope		✔	✔	✔
F0 median	✔	✔		
F0 mean	✔		✔	
H1*-H2*			✔	
H2*-H4*				
H1*-A1*		✔	✔	
H1*-A2*		✔	✔	
H1*-A3*				
HNR05		✔	✔	✔
HNR15		✔	✔	
HNR25	✔	✔	✔	
HNR35		✔	✔	
CPP		✔	✔	
Energy	✔	✔	✔	
R^2	0.14	0.07	0.09	0.04
F	28.66	6.93	8.43	10.04
p	< .05	< .05	< .05	< .05

4.3.3.4 Taiwanese listeners listened to the Swedish and Taiwanese filtered stimuli

Table 4-9 shows that Taiwanese listeners did not use only one cue to boundary strength in filtered speech. It seems that more acoustic measures contributed significantly to the variance in Swedish 2-second speech, including duration, pitch and voice quality. But for Swedish one-word fragments, duration did not seem to be helpful. For Taiwanese stimuli, few cues correlated with boundary strength.

Table 4-9. Results of multiple regression analysis for English listeners listening to the Swedish and Taiwanese filtered stimuli; marks show which acoustic measures contributed significantly to the regression equation; the proportion of explained variance (R^2) is given, and the significance of this result is indicated.

	Taiwanese listener			
	Swedish stimuli		Taiwanese stimuli	
	2-sec	1-wrd	2-sec	1-wrd
Duration	✔			
Rate	✔			✔
F0 range	✔	✔		✔
F0 slope				
F0 median	✔	✔	✔	
F0 mean	✔			✔
H1*	✔		✔	
HNR05	✔	✔		
CPP		✔		
R^2	0.08	0.03	0.05	0.03
F	13.35	6.89	27.07	11.07
p	< .05	< .05	< .05	< .05

4.4 Summary

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

The distribution of the perceived boundary strengths shows that for Swedish stimuli, English listeners showed a three-way distinction in breaks in normal (both 2-second 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 finding are consistent with Carlson *et al.*'s findings.

For Taiwanese stimuli, TSG boundary is not distinguishable from IP boundary in both normal (both 2-second and one-word) and filtered (both 2-second and one-word) stimuli for English

speakers. However, for Taiwanese listeners, TSG is not distinguishable from Word boundary in normal one-second stimuli and filtered stimuli (both 2-second and one-word). In other words, for Taiwanese native speakers, TSG boundary is different from IP boundary but similar to Word boundary, whereas for English native speakers, TSG boundary is the same as IP boundary. This suggests that English listeners differentiate breaks at a prosodically final position (both TSG and IP boundary is at a tone sandhi group boundary) from breaks at a prosodically non-final position (Word boundary). However, Taiwanese speakers treat TSG boundary as the same as Word boundary.

Most of the acoustic measures of final syllables in normal stimuli show significant difference in terms of "break". Among these acoustic measures, H1*-H2*, H2*-H4*, HNR15 and CPP show a three-way distinction for Taiwanese stimuli, and F0 measures, HNRs and CPP in Swedish show a three-way distinction. However, none of these measures, separately or together, are strongly predictive of listeners' judgments of boundary strength. It remains unknown how the listeners in this study made their judgments.

Chapter 5: Taiwanese Sentence Disambiguation

5.1 Introduction

This chapter presents a perceptual study, which explored the use of prosody in resolving syntactic ambiguity in Taiwanese. Previous perception studies (Lehiste, Olive, and Streeter 1976; Price, Ostendorf, Shattuck-Hufnagel, and Fong 1991; Beach 1991; Speer, Kjelgaard, and Dobroth 1996; Allbritton, McKoon and Ratcliff 1996; Kjelgaard and Speer 1999; Schafer, Speer, Warren and White 2000) have shown that some global and most temporary syntactic ambiguities can be resolved by prosodic structure. For example, for the ambiguous sentence *Someone shot the servant of the actress who was on the balcony*, more interpretations of the relative clause (RC) as modifying *the servant* were perceived when there was a prosodic boundary between *the actress* and the RC.

The present study uses pairs of phonetically similar sentences in Taiwanese as the stimuli. Each pair contains one sentence with an early boundary and another with a late boundary. Without boundary information, the sentences were ambiguous. This manipulation makes it possible to examine Taiwanese listeners' use of prosodic phrasing to interpret ambiguous sentences. In addition, analysis of the correspondence between the phonetic attributes of the prosodic structures and the perceived meanings suggests some reliable acoustic cues associated with the process of disambiguation through prosodic phrasing.

This chapter is organized as follows: 5.2 reviews some previous work on sentence disambiguation using prosody; 5.3 describes the current experiment, including the use of the gating paradigm; 5.4 provides the results and their implications.

5.2 Literature Review

Prosodic phrasing and prominence are the two 'prosodic devices' that we use to disambiguate sentences. We use prosodic phrasing to disambiguate sentences like (1), whose ambiguity results from user's attachment preference in the level of syntactic structure, whereas we usually use prominence to differentiate the two interpretations in (2), which deals with the ambiguity at the level of information structure. This study focuses on ambiguous sentences that result from different prosodic phrasings.

(1) *Someone shot the servant of the actress who was on the balcony.*

a. High attachment's reading: the servant was the one who's on the balcony.

b. Low attachment's reading: the actress was the one who's on the balcony.

(2) a. ***I** bought a motorcycle.*

b. *I bought **a motorcycle**.*

5.2.1 Prosodic cues

Previous studies have shown that *speakers* reliably manipulate the two pre-boundary cues, duration and f0, to signal interpretations of an ambiguous sentence. A great deal of data also indicates that *listeners* make use of this prosodic information in sentence comprehension. Lehiste (1973) and Lehiste, Olive and Streeter (1976) found that durational differences were often associated with the ambiguous constituents of the sentence or the ambiguous boundaries of the constituents. Streeter (1978) found that both duration and intonation (= pitch) had significant effects in changing perceived meanings, but intensity showed a significant effect only when it was combined with the other cues. Scott (1982) found that lengthening could signal the

occurrence of a syntactic boundary, and thus change the perceived meaning of a sentence. Price, Ostendorf, Shattuck-Hufnagel and Fong (1991) found that naïve listeners could reliably use duration and intonation to separate structurally ambiguous sentences. Beach (1991) indicated that prosody has an immediate influence on listeners' expectations about upcoming syntactic structures, and the information of duration and pitch was interactively processed in interpreting the ambiguous sentences. Geers (1978) and Speer, Kjelgaard, and Dobroth (1996) showed that when the syntactic boundaries and prosodic boundaries are in conflict, prosody interferes with the syntactic parse. If the two boundaries coincided, the prosodic structure facilitated the comprehension of the syntactic structure.

Based on these findings, the present study on parsing Taiwanese ambiguous sentences tests the hypothesis that the durational pattern and the pitch contour would be correlated with listeners' processing of the syntactic ambiguity. In addition, the results in the previous chapters show that voice quality plays a role in identifying TSG and other prosodic boundaries, therefore, some of the voice measures are also examined here.

5.2.2 Gating experiments

The gating paradigm (Grosjean 1980) is mostly used in spoken word recognition research and has been considered particularly useful in examining moment-to-moment recognition processes and in assessing the amount and location of acoustic-phonetic information needed for the correct identification of a word (as shown in Chapter 2). Gating can also be used in sentence recognition except now the gates are as big as a syllable or a word. Grojean and Hirst (1996) conducted a perceptual study with the gating paradigm in order to see whether listeners could predict the length of an entire sentence at any point within the sentence, or whether they must hear the

potentially last word of that sentence. For example, the sentence *Earlier my sister took a dip* could end on *dip* (+ 0 word), or could continue with *in the pool* (+ 3 words), or could continue with *in the pool at the club* (+ 6 words). They gave listeners parts of the sentence syllable-by-syllable, up to the potentially last word *dip*, and at each gate, the listeners had to decide whether there was more to come, and if so, how much more (the choices are +0/+3/+6). This was inspired by an earlier study by Grosjean (1983) where he had the listeners hear the sentence through the word *dip*, yet *dip* was presented in fragments of increasing duration. The result of this earlier study was that English listeners were very accurate at predicting how much material was missing and their predictions got better as they progressed through the potentially last word *dip*. The result of Grosjean and Hirst's (1996) later study, as shown in Figure 5-1, revealed that listeners estimated a longer length of the sentence as they progressed through all versions of the sentence, and a differentiation between the three ending choices (+0/+3/+6) was only found when the listeners heard the potentially last word *dip.*

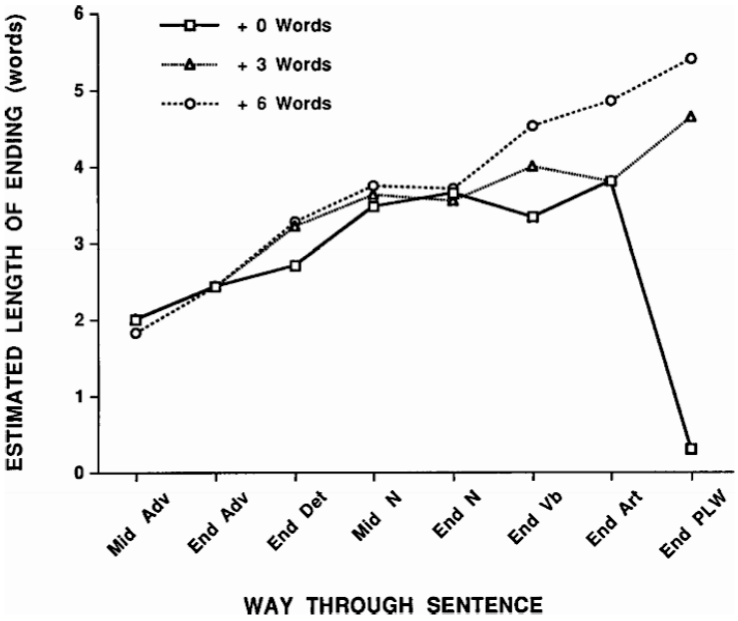


Figure 5-1. From Grosjean and Hirst (1996). Take the sentence *Earlier my sister took a dip* for example. The correspondent gates would be *Ear-lier-my-sis-ter-took-a-dip*.

In the present experiment, a similar sentence-gating task was employed with Taiwanese ambiguous sentences. These sentences were considered ambiguous in that the sentences in each pair were phonetically similar (i.e. their sequences of surface tones were identical), and listeners would need to rely on their knowledge about Taiwanese prosodic phrasing in order to resolve the ambiguity.

5.3 Method

5.3.1 Participants

Fourteen native adult Taiwanese speakers participated in this experiment. They were undergraduate students at the China Medical University, located in Taichung City in Taiwan. All participants identified themselves as Taiwanese native speakers and had no hearing, vision or language related problems. They were compensated with $10 for their participation. Four

subjects' results were excluded from the analyses because some sentences remained ambiguous to them at the end of the sentence.

5.3.2 Stimuli

Eight pairs of test sentences, given in Appendix C, were used in this study. (3) and (4) show one pair of sentences, with the underlines marking the sandhi tones and # marking the TSG boundaries. The tones transcribed here are the surface tones. These are sentence pairs with temporary ambiguity; the ambiguity will resolve by the end of each sentence, leaving only one interpretation. The sequences of surface tones in the two sentences are identical; however, different parsing leads to different interpretations. For instance, the first tone sandhi group in (3) contains five syllables, which makes the high falling tone in *kou51* a citation tone; whereas, the first tone sandhi group in (4) contains six syllables and the high falling tone in *kou51* is the penultimate syllable before the tone sandhi group domain boundary. Presumably, listeners could disambiguate the sentences as they realize that the syllable they hear is directly adjacent to the tone sandhi group boundary (i.e., bearing the citation tone). The point where the sentence becomes unambiguous is referred as the "disambiguation point". Thus, in this example, *kou51* is the first disambiguation point that comes from the Early Boundary condition, and *su33* is the second disambiguation point that comes from the Late Boundary condition.

(3) Early Boundary condition

i33 chin33 gau33 kong55 ***kou51*** # *su33-lang33 po55-sit31-pan55* # *long55 beh31 chhiaN31 i0* #

he very good-at tell folklore private tutoring-center entirely want hire him

" He is good at telling folklores, so the private tutoring center wants to hire him."

(4) Late Boundary condition

*i33 chin33 gau33 kong55 kou51-**su33** # lang33 po55-sit31-pan55 # long55 beh31 chhiaN31 i0 #*

he very good-at tell story the tutoring-center entirely want hire him

" He is good at telling stories, so the tutoring center wants to hire him."

All the test sentences were recorded by a female native Taiwanese speaker, who was a professional political fundraiser host and was not aware of the experimental purpose. The sentences were read at a normal speech rate (=3.6 syllables/sec) and with a natural intonation. Each sentence was then segmented into syllables so that a series of gates could be generated. In each sentence, the first gate contains only the first syllable, and the second gate contains the first two syllables, and so on. The last gate contains the entire sentence. Most of the syllable boundaries could be easily identified by inspection of the waveforms and spectrograms. In the cases where the syllable boundaries were undetermined at first, they were identified with the additional consideration of intensity contours.

5.3.3 Procedure

All the listeners first passed a pre-test before participating in the actual experiment. They were asked to read all the test sentences (written in Chinese characters) verbally to the experimenter. The purpose of this pre-test was to (a) make sure the listeners used the same lexicon as they would hear in the recording, and (b) familiarize the listeners with the test sentences.

The experiment is a two-alternative forced-choice design. During the experiment, the two alternative sentences in each trial were visually presented with Chinese characters side-by-side on the laptop screen, counterbalanced for appearance on the right and left sides of the screen. The listeners saw the sentences first, and then heard a sentence gate. Their task was to listen to

each gate and to determine after each presentation as quickly and accurately as possible whether the sentence gate that had been presented came from the sentence on the right or the sentence on the left. They were asked to click on their answer, and to indicate how sure they felt about their choice with a slider confidence rating scale. The listeners had to give a confidence rating for each trial in order to proceed to the next trial. The listeners saw a slider whose left end and right end were labeled "very unsure" and "very sure" respectively. There were no numbers labeled on the slider, but the listeners were explicitly told that the scale is gradient, not binary or categorical; they could give ratings anywhere on the scale. The gates were presented in an increasing word-blocked fashion; that is, all the gates with only one word were presented first, followed by all the gates with two words, three words, and so on. The last gate for each sentence corresponds to the entire sentence.

5.3.4 Data analysis

5.3.4.1. Response coding

Two dependent variables were measured in this experiment. The first was listeners' choice of response. The listener's choice was given a score of '1' if the listener chose the Late Boundary condition for the answer and '0' if the listener chose the Early Boundary condition. In other words, a score close to '1' for the Late Boundary condition sentences and a score close to '0' for the Early Boundary condition sentences indicate higher correctness.

The other dependent variable was listeners' confidence rating regarding each gate. This is to examine how confident they felt about their decision. Listeners' ratings were given using a slider bar, but the program converted bar positions to values on a 1-100 scale.

5.3.4.2 Acoustic measures

In this study, we measured the rime duration, F0 mean, F0 median, F0 range, F0 slope, H1*-H2*, HNRs and CPP of all the syllables in the sentences; values were obtained from VoiceSauce (Shue *et al.*2011). These are the acoustic measures that revealed a three-way distinction in "break" in normal Taiwanese speech in Chapter 4.

5.3.4.3 Statistical analysis

There are two independent variables in the experimental design: "Boundary Condition" (Early Boundary vs. Late Boundary) and "Gate". Because the various sentences have different numbers of gates preceding the disambiguation points, we only analyzed the "choice scores" and the "confidence ratings" at six selected gates in each sentence. These were the first gate (henceforth: ***First gate***) and the last gate (henceforth: ***Last gate***), the first disambiguation point in the Early Boundary Condition (henceforth: ***Early DP gate***), the second disambiguation point in the Late Boundary Condition (henceforth: ***Late DP gate***), the gate that directly precedes the *Early DP* (henceforth: ***pre-early gate***) and the gate that immediately follows the *Late DP* (henceforth: ***post-late gate***). For instance, (5) and (6) are one pair of test sentences, and the six selected points refer to the same syllables in (5) and (6). The meanings of these sentences and their choice score results are shown in Figure 5-2 in the next section.

Early DP ***Late DP***

↓ ↓

(5) Early condition: *tai31-pak3 u31 nng31-e33 tiam51 -bin33 # tai33 ko55 bo33 mua55 i31 #*

(6) Late condition: *tai31-pak3 u31 nng31-e33 tiam51 # bin33 tai33 # ko55 bo33 mua55 i31 #*

↑ ↑ ↑ ↑

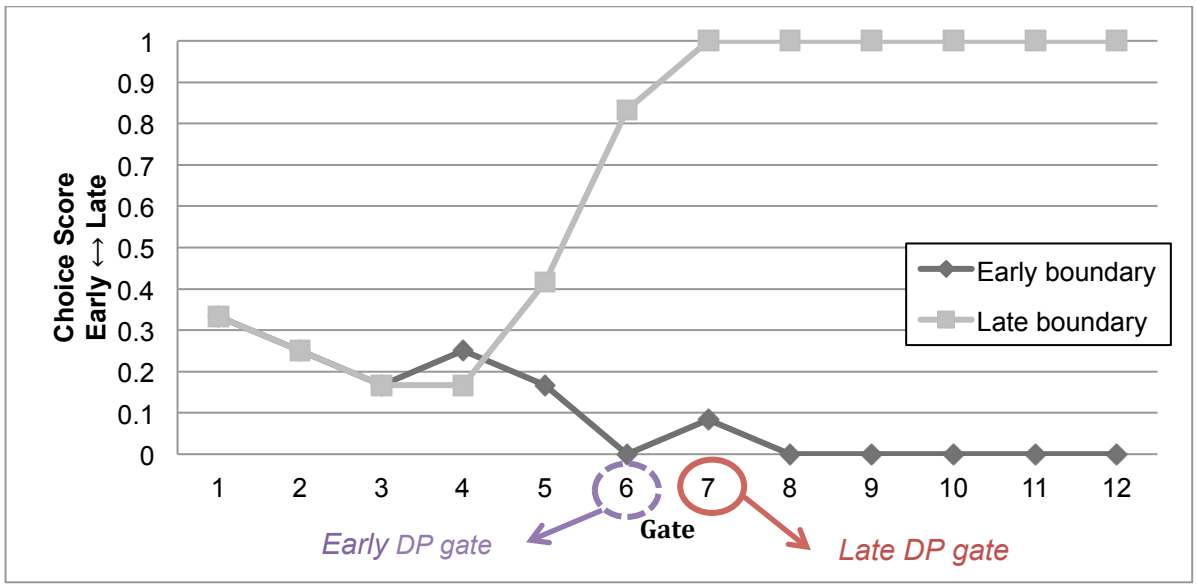
First ***pre-early*** ***post-late*** ***Last***

The choice scores and the confidence ratings were entered separately for repeated measures ANOVA in R with the factors of "Gate" (First, pre-early, Early DP, Late DP, post-late, and Last) and "Boundary Condition" (Early vs. Late).

5.4 Results

5.4.1 Choice Score

Figure 5-2 is the averaged choice score results across listeners for one pair of the test sentences – (5) and (6) in the last section. The two sentences are provided again below the figure and the tone sandhi group boundaries are indicated with #. In the Early Boundary Condition (=(5) in the previous section), the syllable at Gate 6, *tiam51*, is the disambiguation point (i.e. *Early DP gate*) because it is located at the tone sandhi group boundary. In the Late Boundary Condition (=(6) in the previous section), the syllable at the tone sandhi group boundary is Gate 7, *bin33*, which is the disambiguation point (i.e. *Late DP gate*).



a. Early Boundary: *tai31-pak3 u31 nng31- e33 tiam51 # bin33- tai33 ko55 bo33 mua55 i31 #*

"Mr. Tai is not satisfied with these two stores in Taipei."

b. Late Boundary: *tai31-pak3 u31 nng31- e33 tiam51 - bin33 # tai33-ko55 bo33 mua55 i31 #*

"The representative is not satisfied with just having two locations in Taipei."

FIGURE 5-2. The choice scores averaged across listeners for one pair of test sentences (Late Boundary =1; Early Boundary =0).

The statistical results are presented in Table 5-1. Figure 5-3 presents the choice scores averaged across all listeners and sentences in the two Boundary Conditions (Early vs. Late) and at the six Gates (*First, pre-early, Early DP, Late DP, post-late* and *Last*). Significant effects were found for Boundary Condition ($p < .01$) and the interaction between Boundary Condition and Position ($p<.01$). As can be seen in Figure 5-3, the choice score increases across gates for the Late Boundary Condition, while it decreases across gates for the Early Boundary Condition.

Table 5-1. Repeated measures ANOVA table for Choice Score.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BoundaryCondition	1	16.02	16.02	134.94	< .01 *
Gate	5	0.26	0.05	0.45	0.817
BoundaryCondition × Gate	5	9.14	1.83	15.41	< .01*
Residuals	936	111.1	0.12		

Post hoc comparisons of the scores for each Boundary Condition across the Gates revealed that (i) for the Late Boundary Condition, the score at *First* did not differ from *pre-early*, and their scores were significantly lower than *Early* DP which in turn had a lower score than *Late DP, post-late* and *Last*; (ii) for the Early Boundary Condition, the scores of *First* and *pre-early* were significantly lower than *Early DP*, *Late DP*, *post-late* and *Last*. The positions with statistically the same scores are enclosed in the same ellipses in Figure 5-3.

The critical position is *Early DP*. In the Early Boundary Condition, *Early DP* is when listeners detected a boundary. In the Late Boundary Condition, *Early DP* is when listeners detected the absence of a boundary, and they detected the appearance of a boundary at *Late DP*.

In addition, listeners did not start giving different scores until they reached the first disambiguation point (*Early DP*). They did not show any difference in scores even at *pre-early*. This suggests that not enough boundary cues were provided before the disambiguation point and this possibility will be addressed in the later section on acoustic measures.

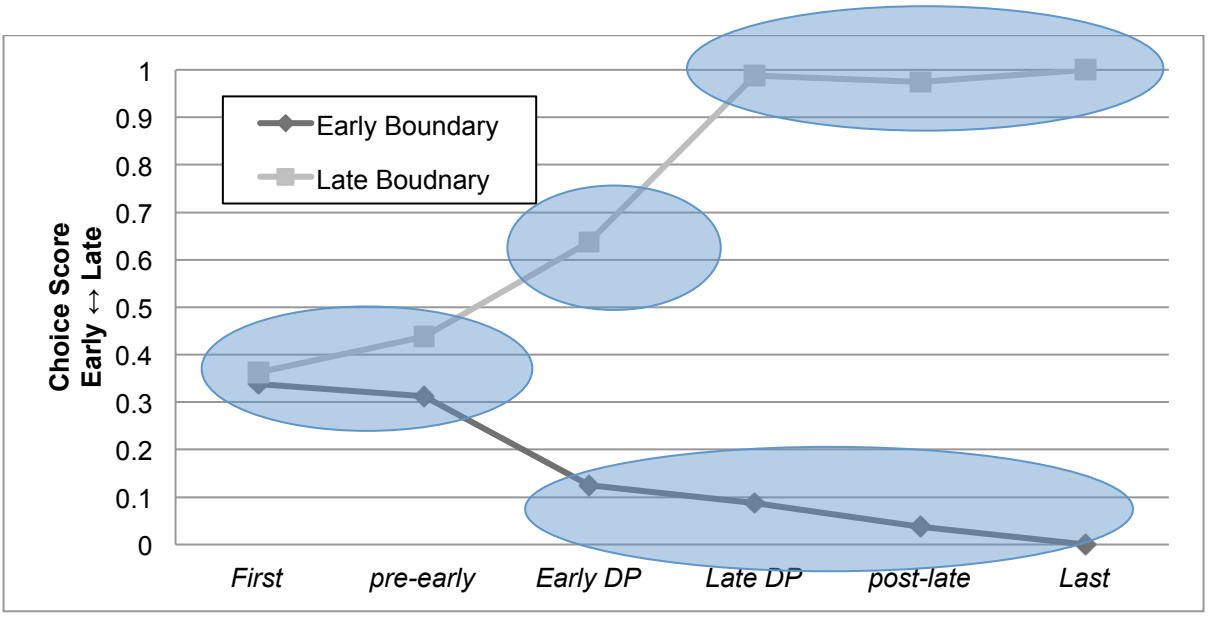


FIGURE 5-3. Average choice score across Gates and Boundary Conditions.
For the Late Condition: *First = pre-early < Early DP < Late DP = post-late = Last*
For the Early Condition: *First = pre-early < Early DP = Late DP = post-late = Last*
Ellipses indicate no significant difference within a condition.

Table 5-2 presents the average choice scores as well as the standard deviation. In the Early Boundary Condition, the standard deviation gradually declined as the listeners progressed through the sentence. In the Late Boundary Condition, the variation in score showed a sudden reduction after *Early DP*. It seems that for the Late Boundary Condition, listeners were not so sure about their choice (i.e., more variation in score) as they detected a possible absence of a boundary at *Early DP*; however, the variation got reduced when they detected a real boundary at *Late DP*.

TABLE 5-2. Average choice scores (mean) and standard deviations (sd) across Positions and Boundary Conditions.

	First	Pre-early	Early DP	Late DP	Post-late	Last
	mean (sd)	mean(sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
Late	0.36 (0.48)	0.44 (0.50)	0.64 (0.48)	0.99 (0.11)	0.98 (0.16)	1.00 (0)
Early	0.34 (0.48)	0.31 (0.47)	0.13 (0.33)	0.09 (0.28)	0.04 (0.19)	0 .00(0)

5.4.2 Confidence

The choice score results in the previous section suggest that listeners made accurate judgments only after they reached the *Early DP*, whether the sentence was from the Early Boundary Condition or the Late Boundary Condition. Their confidence ratings for these judgments are presented in Figure 5-4. As the listeners encountered longer sentence fragments, they felt more confident about their responses - equally so, across both Boundary Conditions. The repeated measures ANOVA result is shown in Table 5-3. A significant main effect was only found for Gate (p <.01), but not for Boundary Condition.

Table 5-3. Repeated measures ANOVA table for Confidence

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BoundaryCondition	1	3	3	0.004	0.947
Gate	5	118796	23759	42.183	< .01*
BoundaryCondition × Gate	5	936	187	0.332	0.893
Residuals	936	527190	563		

In a typical gating experiment, researchers often examine a 'recognition point' which is the point when listeners make a correct choice without further changes and their confidence rating about the choice is 80% or higher. With this threshold (=80%) in mind, we find that listeners' confidence rating went beyond 80% after they reached *Late DP* in both conditions, as shown with the dashed line at 80% in Figure 5-4. Therefore, in terms of confidence, the critical point was *Late DP*. This is the position when listeners had detected a non-boundary and a boundary in both Boundary Conditions. It seems that whether they detected a boundary or not determines their confidence about their choice.

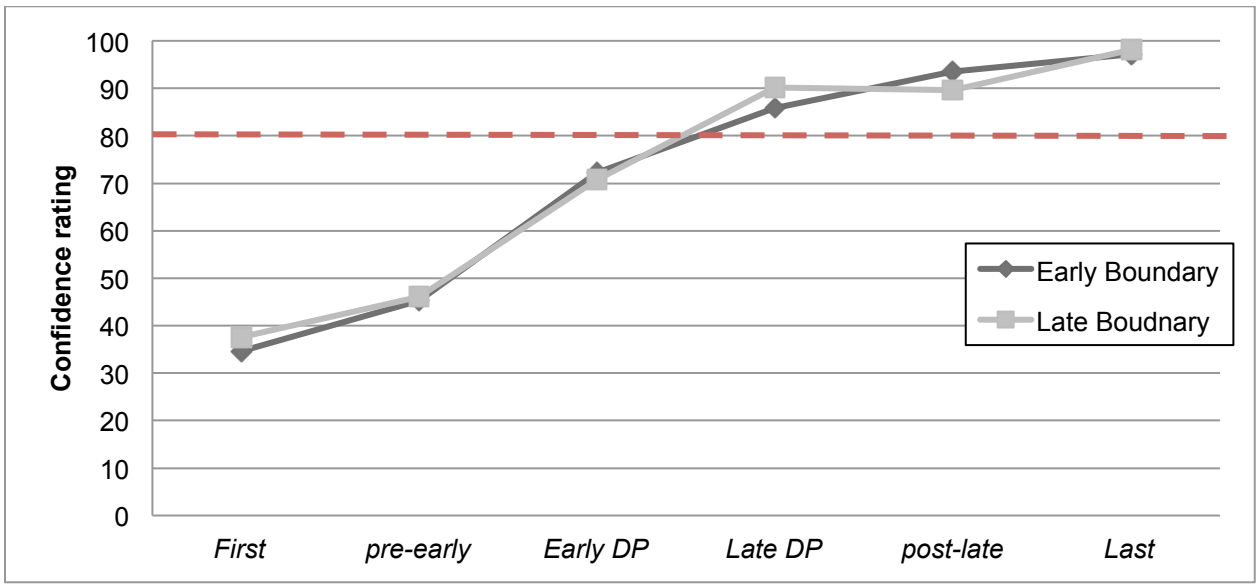


FIGURE 5-4. Average confidence ratings across the Positions and Boundary Conditions.

In a word, the more gates the listeners heard, the more confident they were about their choice scores. To conclude from the choice score and the confidence results, it was only at *Early DP* and *Late DP* that the listeners clearly discriminated between the two boundary conditions. A similar result was found in Grosjean's (1983) experiment. In addition, listeners showed greater confidence in their choices after they reached *Early DP*, which clearly showed them whether there was a boundary or not. To tell whether there was a boundary, listeners would need to perceive some acoustic cues which signal the appearance of a prosodic boundary. In the next section, several acoustic cues were investigated compared with the perception results.

5.5 Acoustic results compared to perception results

Given that listeners started to give accurate judgments right at the *Early DP* or *Late DP*, we are interested in the acoustic cues that they received before and at these disambiguation points in each condition. Thus, in all the following figures, the syllables are right-aligned; ' 0' is the DP of the early boundary condition and '1' is the DP of the late boundary condition. The acoustic

measures were obtained from the voiced portion of each syllable before and including *Late DP*. For '1', '0', '-1' and '-2', each has eight pairs of values; '-3' has seven pairs of points; '-4' has 6 pairs; '-5' has three pairs; '-6' and '-7' has one pair. Therefore, at position "-5", "-6" and "-7", paired t-tests could not be employed.

5.5.1 Final lengthening

The results for rime duration are presented by Boundary Condition in Figure 5-5. The duration at "1" is the average duration of the *Late DPs* from all sentences. The duration at "0" is the average duration of the *Early DPs* from all sentences. With the Boundary Condition (Early vs. Late) as the independent variable, paired *t*-tests at each Position revealed that durations were only significantly different at the *Early DP* ($t(7)=10.71$, $p < .05$) and the *Late DP* ($t(7)=5.68$, $p < .05$). The increase in duration at the *Early DP* and at the *Late DP* suggests that there was a duration contrast caused by final lengthening in each Boundary Condition. The final lengthening was only realized at the last syllable in each condition, instead of increasing gradually through the sentences. This reminds us of Klatt's (1975) and Wightman *et al.*'s (1992) studies where they found that lengthening occurred only in the phrase-final syllable. Thus, final lengthening could have been a cue to make these two Boundary Conditions distinct from each other.

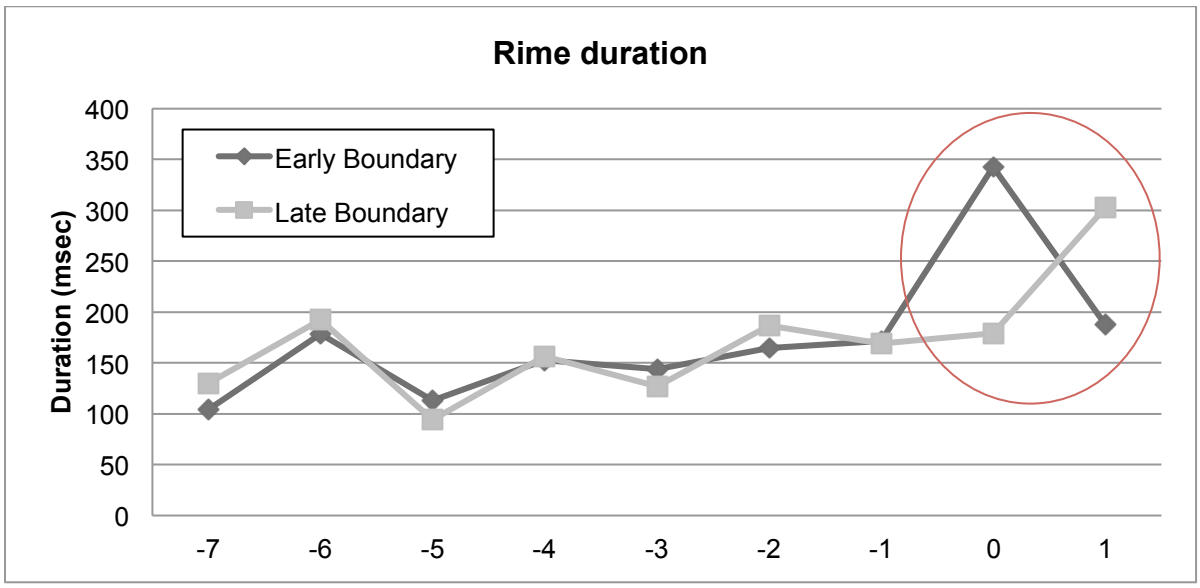


FIGURE 5-5. Mean duration (msec) of the two Boundary Conditions at the nine Positions. "1" and "0" denote the *Late DP* and the *Early DP* respectively. "1" and "0" show the averaged duration at *Late DP* and *Early DP* across sentences.

5.5.2 Final pitch declination

Four F0 measures were made from each syllable up to the *Late DP* – F0 range, F0 slope, F0 median and F0 mean. The results are presented in Figure 5-6 and Figure 5-7. For each F0 measure in each condition, the pitch contour shows a declination pattern as the sentence proceeds. Paired *t*-tests at each Position with Boundary Condition as the independent variable reveal that F0 mean and F0 median in the two Boundary Conditions differ significantly at *Late DP* (F0 median: $t(7)=6.77$, $p < .05$; F0 mean: $t(7)=5.61$, $p < .05$). This is because *Late DP (= point "1")* in the Early Boundary Condition began a new tone sandhi group and so involved F0 reset.

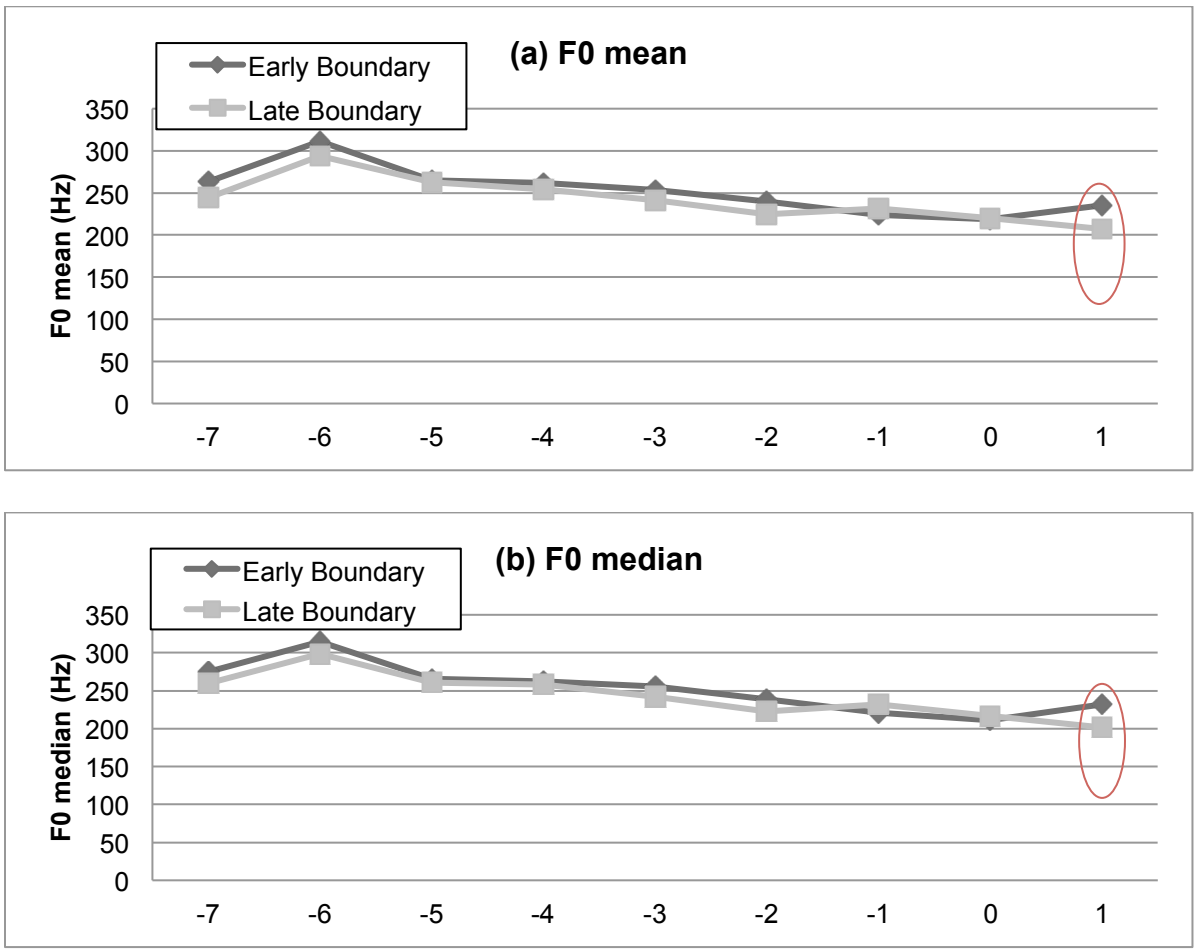


FIGURE 5-6. Average values of the two F0 measures: (a) F0 mean, (b) F0 median. The x-axis denotes the position of the gates: '0' and '1' correspond to the *Early DP* and the *Late DP*, respectively.

On the other hand, F0 range in the two Boundary Conditions differs at *Early DP* ($t(7)=2.93$, $p<.05$) but not at *Late DP*. The finding that F0 range is wider in the Early Boundary Condition at *Early DP* (= point "0") is consistent with previous studies, which found that syllables at a prosodic boundary tend to have a wider F0 range. For the same reason, F0 with a wider magnitude was expected to be observed at point "1" in the Late Boundary Condition; however, because the syllable at *Late DP (= point "1")* in the Early Boundary Condition was expected to have an F0 reset which might result in a relatively wider F0 range, the F0 range of these two

Boundary Conditions at *Late DP (=point "1")* did not seem to differ. No significant difference was found in F0 slope at any Position.

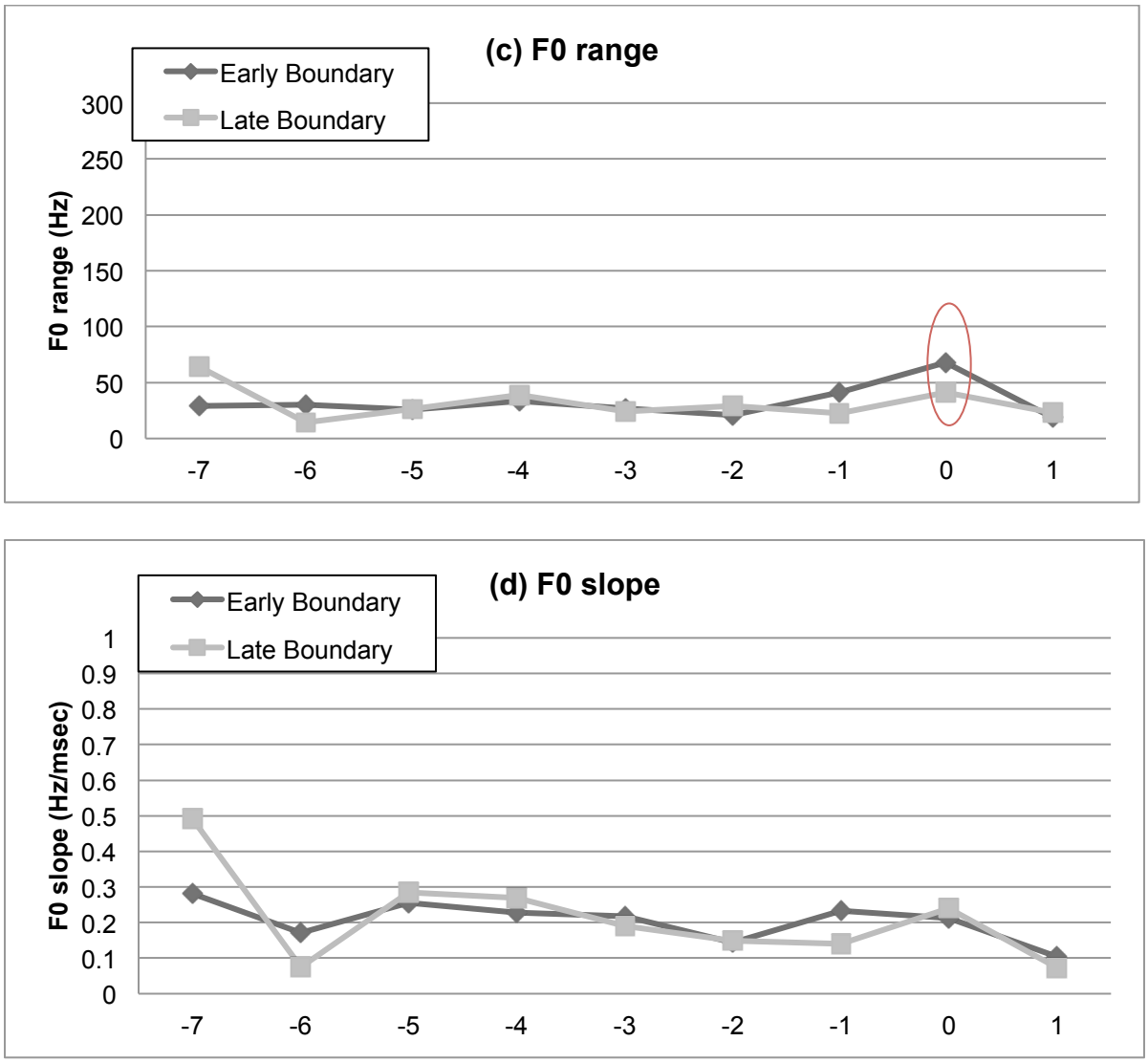


FIGURE 5-7. The averaged values of the two F0 measures: (a) F0 range, and (b) F0 slope. The x-axis denotes the position of the gates: '0' and '1' correspond to the *Early DP* and the *Late DP,* respectively.

Therefore, the results for F0 mean and F0 median confirmed that pitch declines as a sentence continues, and they also provide evidence for pitch reset phrase-initially in a Tone Sandhi Group. The result for F0 range suggests that a wider F0 range tends to appear phrase-finally.

5.5.3 Final glottalization / aperiodicity

Figure 5-8 presents average H1*-H2* through the sentences for the two Boundary Conditions. There is no significant difference in H1*-H2* at any Position between the Early Boundary Condition and the Late Boundary Condition. In other words, creakiness did not vary with phrasing and thus could not have served as a cue for listeners. Interestingly, it is found that the H1*-H2* values increased as the sentences proceeded, which suggests that the speaker in this experiment tended to use a generally breathier voice, instead of a creakier voice, in all sentences.

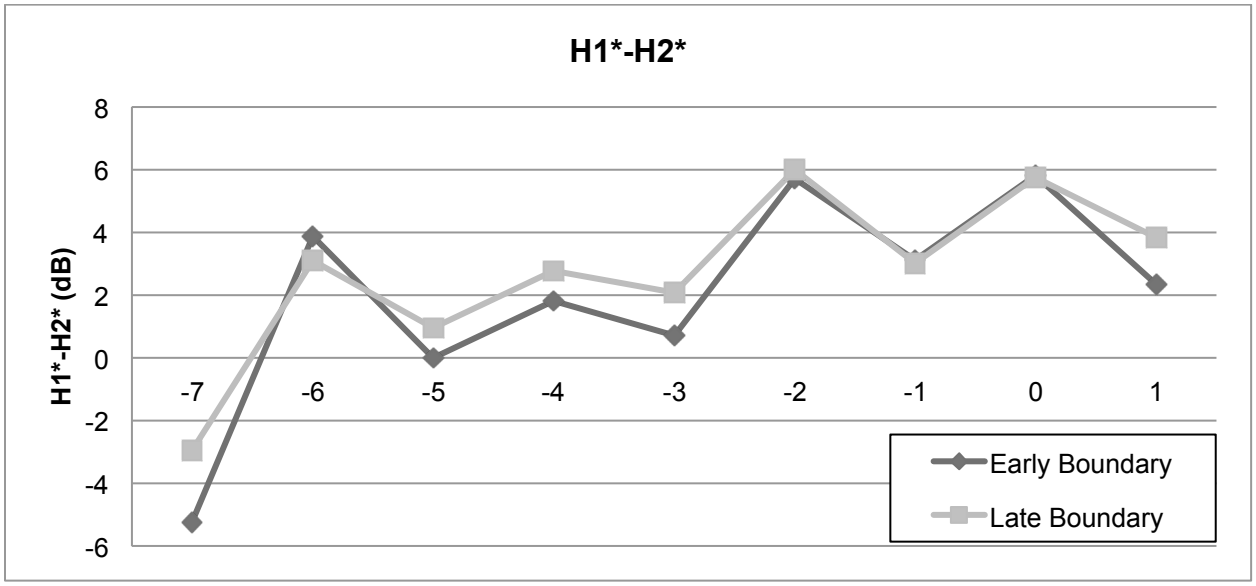


Figure 5-8. Average H1*-H2* across the two Boundary Conditions at the nine Positions. "1" and "0" denote the *Late DP* and the *Early DP* respectively.

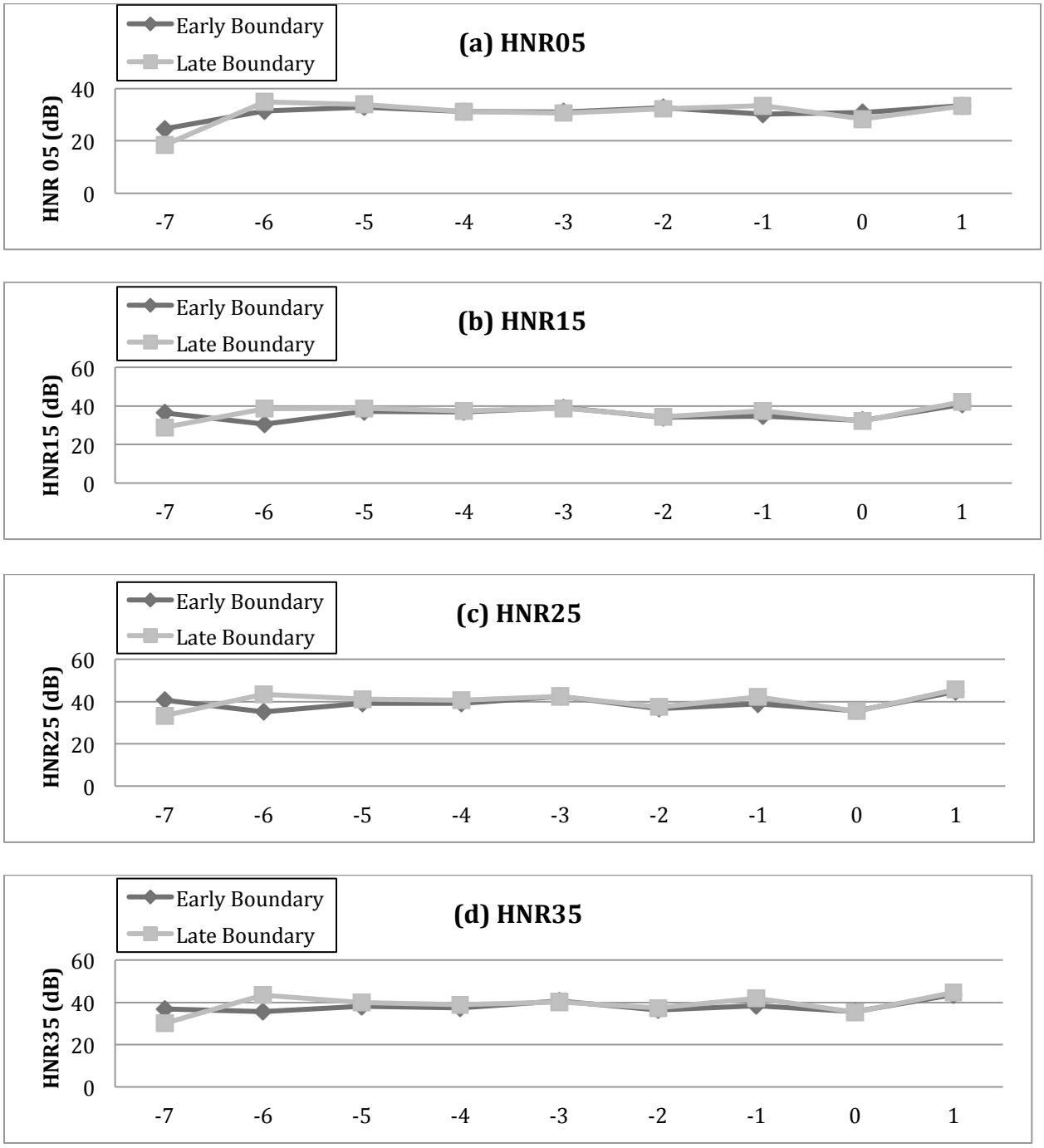


Figure 5-9. Average HNRs across the two Boundary Conditions at the nine positions. (a) HNR05 (b) HNR15 (c) HNR25 (d) HNR35. "1" and "0" denote the *Late DP* and the *Early DP* respectively.

Figure 5-9 presents average HNR values through the sentences for the two Boundary Conditions. The two Boundary Conditions do not differ in HNRs in either Position. This

indicates that listeners could not use aperiodicity or noise to disambiguate the sentences. Figure 5-10 shows the average CPP for the two Boundary Conditions. Like the HNRs results, no significant difference between the two Boundary Conditions was found in CPP at either Position. This result again shows that aperiodicity or breathiness was not a cue that listeners could use to disambiguate the sentences.

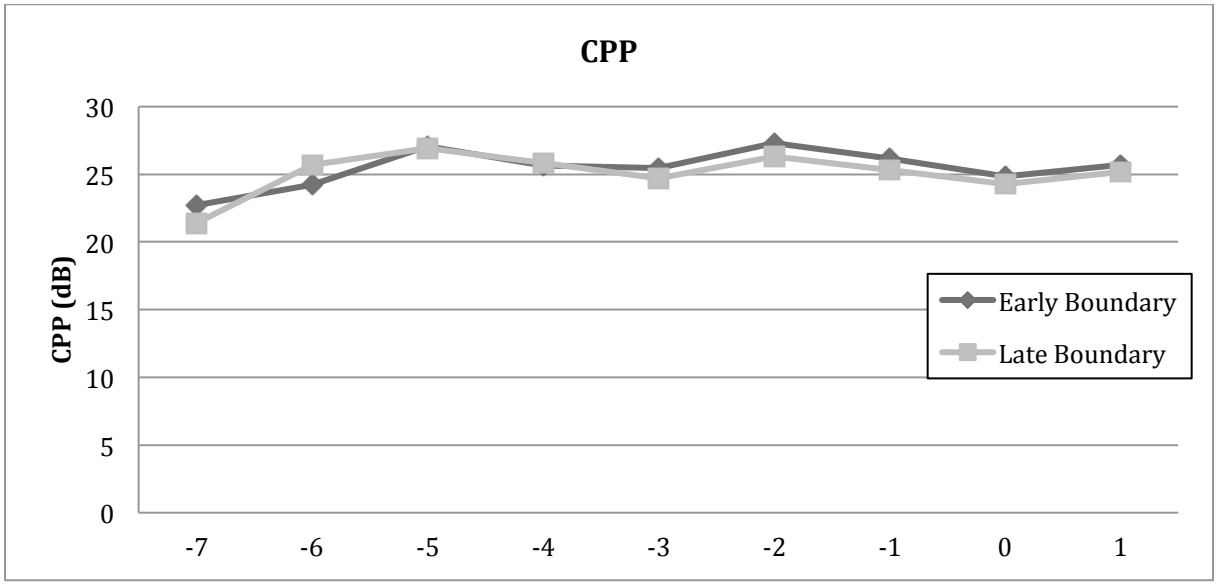


Figure 5-10. Averaged CPP across the two Boundary Conditions at the nine positions. "1" and "0" denote the *Late DP* and the *Early DP* respectively.

5.6 Multiple linear regressions between the scores and the acoustic measures

In order to speculate about the correlations between listener's responses (both choice score and confidence ratings) and acoustic measures at each position, multiple linear regression was carried out in R. Table 5-4 shows which of the acoustic measures contributed significantly to the regression equation and what proportion of the variance could be explained by all the measures together. Separate multiple regression analyses were carried out for Choice Score and Confidence Rating. The proportions of explained variance of Choice Score and of Confidence Rating were 5% and 19% respectively. It seems that the acoustic measures contribute more to confidence rating than choice scoring.

TABLE 5-4. Results of multiple regression analysis for choice scores vs. acoustic measures and for confidence vs. acoustic measures. The ticks indicate the acoustic measures that contributed significantly to the regression equation; the proportion of explained variance (R^2) are also given.

	Choice Score	Confidence Rating
Duration	√	√
F0 range	√	√
F0 slope	√	√
F0 median	√	√
F0 mean		√
H1*-H2*	√	
HNR05	√	√
HNR15		
HNR25	√	
HNR35	√	√
CPP	√	√
R^2	0.05	0.19
F	5.69	27.97
p	< .01	< .01

Previous studies have established that duration and f0 determine which interpretation listeners would assign to a syntactically ambiguous sentence (e.g. Lehiste et al. 1976; Beach 1991). The acoustic analyses here support this claim and show clear differences in duration and f0 for the comparison of the two Boundary Conditions. The results also suggest that sentences were disambiguated only at the disambiguation points, not earlier. In addition, the voice quality in sentences from the two Boundary Conditions did not differ from each other. Listeners therefore could not use voice quality as a cue to disambiguate sentences. However, the multiple regression results revealed that listeners' responses (both choice score and confidence rating) correlate with not only duration and F0 measures but also voice and noise cues.

Chapter 6: General Discussion and Conclusion

This dissertation investigated the Taiwanese Tone Sandhi Group by examining its perception and acoustic correlates. The most important objective in this research was to answer the question discussed at the beginning of the dissertation: *How do listeners perceive and recognize Taiwanese Tone Sandhi Groups?* The dissertation started with the recognition of the checked citation vs. sandhi tones on monosyllables, and explored the acoustic differences between citation and sandhi tones in a corpus. This study was then followed by one on the identification of Tone Sandhi Groups in read speech, and the detection of Tone Sandhi Group boundaries in spontaneous speech. The final study was a sentence disambiguation experiment implemented with a gating paradigm.

6.1 Are Taiwanese sandhi tone and citation tone neutralized?

Taiwanese sandhi tones and citation tones are widely assumed to share the same tonal categories. In other words, pairs of tones are said to have the same tonal output, and thus their pitch contours should show no difference. However, the prosodic positions where the citation tone and the sandhi tone appear are different – citation tones appear in the final position of a Tone Sandhi Group (=TSG) whereas sandhi tones are only found in the non-final positions of a TSG. Across languages, domain-final positions usually involve final lengthening, pitch declination and laryngealization. This suggests that there might be differences in these acoustic measures between citation and sandhi tones due to their positional difference.

The results from the checked tone recognition experiment in Chapter 2 indicate that listeners identify sandhi tones more accurately than citation tones. This accuracy result suggests that the

sandhi tone and the citation tones may be incompletely neutralized. In addition, the Identification Point (IDP) result seems to suggest that sandhi tones are recognized sooner; however, this finding might need to be revisited later in that durational ratio is a more comprehensive measure than raw duration. The corpus study in Chapter 2 further shows that the aforementioned acoustic measures in domain-final position do preserve distinctions between all the tone pairs, at least for the speaker who produced the corpus. In general, citation tones are longer in duration, wider in F0 range and creakier in voice quality, especially in TSG-final position. The findings from the two studies suggest that listeners were expecting these acoustic cues to final position (i.e. duration, f0 range and voice quality) when they were identifying tones. Once they realized that the anticipated cues were not present in a non-final position, it is very likely that they would immediately choose the sandhi tones as the answer. However, this argument was not directly tested in this current work.

The scope of the dissertation then expanded to listeners' identification of the prosodic domains. The results from the prosodic domain experiment in Chapter 3 indicate that sandhi tones are indeed different from citation tones in that listeners are more accurate identifying tones at TSG and IP boundaries (= citation position) than tones at Word boundaries (= sandhi position).

The finding from the prosodic domain experiment seems to contradict the results of the checked tone recognition study. However, this is not the case, in that the stimuli in the two experiments were very different – the stimuli presented in the checked tone experiment were parts of a syllable whereas the stimuli presented in the prosodic domain experiment were low-pass filtered 4-syllable utterances. In other words, listeners were asked identify the tone of a syllable from its components in one experiment (*bottom-up*), yet they were asked to identify the

tone of a syllable from its context (*top-down*) in the other experiment. In bottom-up processing, recognition decisions can not be made before sufficient acoustic/phonetic information is accumulated. Listeners were able to identify sandhi tones and citation tones within 60 msec based on their observation of the presence or absence of some acoustic/phonetic cues (i.e. duration, f0 range, and voice quality). In top-down processing, the decisions were made before sufficient bottom-up information could have accumulated (i.e. before the target syllable was actually presented). Listeners could possibly identify the prosodic domains based on their observations about the signals before the syllable at the prosodic boundary. In the checked tone identification experiment, it seemed that listeners were aware of the absence of some acoustic cues in the sandhi tones so that they identified these tones faster than citation tones. However, this argument awaits further analyses in future work. In the prosodic domain experiment, listeners were presented not only with the syllables of interest but also the preceding syllables, which provided them with the opportunity to compare sandhi tones and citation tones. Besides, the stimuli were low-pass filtered, which means that the available cues were limited to duration, f0 range and voice quality. In this experiment, listeners were better at identifying citation tones because the aforementioned acoustic cues were more prominent in citation tones.

Therefore, with the identification tasks in the two experiments (the checked tone experiment in Chapter 2 and the prosodic domain experiment in Chapter 3), we find that even though sandhi tones and citation tones with the same pitch contours suggest neutralization, these two tones are distinguishable from the listeners' perspective. The acoustic/phonetic cues listeners utilize can be duration, f0 range and voice quality, not only on the syllables of interest, but also on the 'prosodic position' where the syllables appear (i.e. domain final or domain non-final).

The results from the boundary strength rating experiment in Chapter 4 provide data about prosodic discrimination by Taiwanese speakers. The TSG boundary is not distinguishable from the Word boundary. However, IP boundary is distinct from Word boundary. Therefore, the boundary type is an essential factor in tone neutralization in Taiwanese. Table 6-1 provides the summary of the experiments with the Taiwanese listeners in Chapters 2-4.

Table 6-1. Summary of the experiments with the Taiwanese listeners. The data of the Identification Task were collected from the checked tone experiment (Chapter 2) and the prosodic domain experiment (Chapter 3). The values here are percent correct. The data of the Discrimination Task were collected from the boundary strength rating experiment (Chapter 4).

	Identification Task		*Discrimination Task*	
	Normal (Ch.2)	*Filtered* (Ch.3)	*Normal* (Ch. 4)	*Filtered* (Ch. 4)
one-syllable				
Word (sandhi) vs. TSG (citation)	77 vs. 67	---	same	same
Word (sandhi) vs. IP (citation)	---	---	different	different
4-syllable				
Word (sandhi) vs. TSG (citation)	---	69 vs. 82	---	---
Word (sandhi) vs. IP (citation)	---	69 vs. 83	---	---
2-second				
Word (sandhi) vs. TSG (citation)	---	---	different	same
Word (sandhi) vs. IP (citation)	---	---	different	different

6.2 Is the TSG boundary perceivable and distinct from other prosodic boundaries?

Earlier production and articulation studies suggested that the TSG is an independent prosodic domain (Peng 1997; Keating *et al.* 2003; Pan 2003; Pan 2006). This dissertation provides the perception data that supports this claim. The stimuli of the prosodic domain experiment in Chapter 3 were 4-syllable low-pass filtered read speech, and the perception results indicate that syllables from the Word boundary condition are significantly different from syllables from either the TSG boundary or the IP boundary condition. Furthermore, the syllables from the TSG boundary condition are also distinct from the syllable from the IP boundary condition. Thus the

TSG is a discrete prosodic domain in that listeners are able to separate it from the other two domains in read speech.

The stimuli of the boundary strength rating experiment in Chapter 4 were spontaneous speech with Taiwanese listeners and American English listeners. For native speakers (i.e. Taiwanese listeners), a TSG boundary is distinct from Word boundary and IP boundary only in normal (i.e. unfiltered) and long (2-second) fragments. Otherwise, a TSG boundary is not distinguishable from Word boundary. The acoustic analyses on the last syllables in the stimuli reveal that not only duration and F0 measures but also most of the voice quality measures were useful to some extent to the listeners. However, the R^2 values obtained from the multiple linear regressions suggested that these acoustic measures on the very last syllables were far from sufficient to account for listeners' perceptions. Therefore, in future work, it would be interesting to examine the acoustic measures across the entire stimulus fragments given that more information would provide them more cues (as shown in the prosodic domain experiments in Chapter 3 and the boundary strength rating experiment in Chapter 4) which facilitate listeners' judgments.

6.3 How do native speakers disambiguate sentences?

Chapter 5 focused on listeners' responses to ambiguous sentences. The results show that sentences were disambiguated only at the disambiguation points, not earlier. The listeners were presented with two ambiguous sentences in characters on the screen in each trial; in other words, they were aware of the location of the disambiguation point, and maybe they only need these two syllables (the two disambiguation points) around the choice of boundary locations in this task to make their decisions. The multiple linear regression results between their choice score and the acoustic measures at each position across the syllables up to the disambiguation points show that

listeners used not only duration and f0 measures but also voice quality to make their judgments. In future work, the visual presentation of the characters in the ambiguous sentences should be displayed one word after the other and a closer observation of the relations between the choice scores and the acoustic measures should be implemented. In addition, since the disambiguation point is strongly connected with the prosodic position (i.e. the disambiguation points are the syllables at the TSG boundaries), it will be interesting to replicate the disambiguation experiment while the stimuli (both auditory and visual) are presented in reverse order. The design of the proposed experiment is based on the question whether the listeners focus on a discrete point at the prosodic boundary or do they focus on the whole prosodic domain to disambiguate sentences. A backward gating experiment should be performed to determine whether the so-called disambiguation point has provided sufficient information or whether listeners need to seek gather more cues from the other parts of the prosodic domain. The visual-world eye-tracking paradigm would also be a useful way to examine the time course of information uptake in such sentences.

6.4 Conclusion

To conclude, the studies in this dissertation provide insight into the properties of the Taiwanese Tone Sandhi Group in terms of its perception and acoustics as well as its status in the Taiwanese prosodic hierarchy. The Tone Sandhi Group is a discrete prosodic domain whose properties are variable in that sometimes it is like a Word/Syllable and at other times like an Intonation Phrase. However, it remains identifiable and discriminable even in experiments with degraded speech or limited exposure of the stimuli.

Taken into account the results we obtained from the series of experiments, we attempt to answer the question outlined at the beginning of this Chapter: *How do listeners perceive and*

recognize Taiwanese Tone Sandhi Groups? First of all, listeners perceive citation tones and sandhi tones differently, which suggests that the Tone Sandhi Group is an independent prosodic domain so that listeners know whether a tone is at domain-final position or not. Secondly, TSG is determined prosodically because listeners could identify the citation tones from the sandhi tones not only in normal speech but also in filtered speech. Thirdly, duration, f0 range and voice quality are the three acoustic/phonetic cues that contribute to listeners' identification and recognition of the TSG as well as the discrimination of the TSG from the other prosodic domains. Lastly, similar cues are utilized at the disambiguation points in the sentence disambiguation task.

APPENDICES

Appendix A: English instructions for listener in Boundary strength rating experiment

Thank you for participating in this experiment. We are studying how people learn "breaks" between words. In this experiment, you will be presented with the non-English spoken utterance fragments that are either 2-second or one-word long. These fragments could be like this:

(1) The bonds of love and family that we forge are vital sources of meaning and identify.
(2) The bonds of love and family …
(3) The bonds of love and …

We would like you to judge on how strong a break will follow these fragments. For instance, the break after (1) is bigger than the break after (2) because the break in (1) is at a sentence boundary whereas the break in (2) is at a phrase boundary. There will probably be a very small or even no break after (3) because it is in the middle of a phrase.

Your task in this experiment is to express your judgments on a gradient slider. If you think there will be a bigger boundary after the last word, then you pull the slider rightward; if you think there will be smaller boundary after the last word, then you pull the slider leftward. Please make the judgment by your intuition since you can only listen to each fragment once.

There will be two sessions. The fragments in the first session are recordings less speech-like. The fragments in the second session are the original recordings.

Appendix B: Significant interaction effects in Boundary strength rating experiment
In the table below, "native" and "language" refer to "native language" and "stimulus language" in the text, respectively.

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
break × length	2	7.827	3.913	8.651	< .01 *
break × language	2	7.875	3.937	12.052	< .01 *
break × quality	2	18.45	9.227	14.689	< .01 *
native × length	2	3.095	1.547	3.421	< .05 *
break × length × language	2	3.796	1.8981	9.355	< .01 *
native × break × language	2	4.517	2.258	6.913	< .01 *
native × break × length × language	2	1.351	0.6757	3.330	< .05 *
native × break × length × quality	2	3.054	1.5270	4.996	< .01 *
native × break × language × quality	2	3.366	1.6831	7.470	< .01 *
native × break × length × language × quality	2	1.750	0.8752	4.597	< .05 *

Appendix C: Stimuli in the Disambiguation experiment

Eight pairs of the ambiguous sentences and their two intended interpretations. The tones transcribed here are the surface tones.

1. *in33 ui31 i33 ai51 i33 seng55 chap3 ban31 kou55 hou31 i33 be55 saN33 a55 kou31*
(A) She was given ten thousand dollars to buy new clothes to impress the doctor whom she loves.
(B) Knowing she has lost a lot of money due to her love for gambling, I have saved ten thousand dollars for her to buy new clothes.

2. *i33 kong55 i33 ka31 ching55 kah3 tioh3 lau33 hueh3*
(A) He says his fingernail bleeds when he bites his nails.
(B) He says his nails bleed when he teaches.

3. *i33 chin33 gau33 kong55 kou51 su33 lang33 po55 sip3 pan55 long55 beh5 chhiaN31 i0*
(A) He is good at telling stories, so the tutoring center wants to hire him.
(B) He is good at telling folklores, so the private tutoring center wants to hire him.

4. *i33 kong55 chit5 e33 gin55 a51 bo33 chhai51 chiN13 long55 hou33 i33*
(A) He said because this child didn't have any money for food, he gave him everything.
(B) He said that because it is such a pity to see this child unable to reach his potential, he gave him all of his money.

5. *i33 ai51 chit5 e33 tng33 tou33 chhi33 be33 kah5*
(A) He loves this long rectangular-shaped city very much.
(B) He loves the length of this item very much.

6. *i33 tiaN31 tiaN31 khi55 bio55 hoe33 siong33 sim55 e33 sou55 chai33*
(A) He often depicts a scene of sorrow.
(B) He often builds temples; these temples are the heart and soul of the monks.

7. *na31 bo33 chui31 lang13 tioh3 e31 sai55 li31 khui55 a0*
(A) If there is no criminal present, then you all can leave.
(B) If he is not drunk, then he can leave.

8. *tai31 pak3 u31 nng31 e33 tiam51 bin33 tai33 ko55 bo33 mua55 i31*
(A) Mr. Tai is not satisfied with these two stores in Taipei.
(B) The representative is not satisfied with just having two locations in Taipei.

BIBLIOGRAPHY

Abramson, A. S. (**1962**). The Vowels and Tones of Standard Thai: Acoustical Measurement and Experiments. Bloomington, IN, Indiana University Research Center in Anthropology, Folklore, and Linguistics.

Abramson, A. S. (**1972**). Tonal experiments with whispered Thai. In A. Valdman (ed.), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre* , pp. 31-44. The Hague, Mouton.

Abramson, A. S. (**1975**). The tones of Central Thai, Some perceptual experiments. In J. G. Harris and J. R. Chamberlain (eds.), *Studies in Thai Linguistics in Honor of William J. Gedney*, pp. 1-16. Bangkok, Central Institute of English Language.

Albritton, D., McKoon, G. and Ratcliff, R., (**1996**). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology, Learning, Memory and Cognition*, **22**, 714-735.

Barrie, M. (**2006**). Tone circles and contrast preservation. *Linguistic Inquiry*, **37**, 131-141.

Baum, S. R., Pell, M. D., Leonard, C. L. and Gordon, J. K. (**2001**). Using prosody to resolve temporary syntactic ambiguities in speech production, acoustic data on brain-damaged speakers. *Clinical Linguistics & Phonetics*, **15** (6). 441-456.

Beach, C. M. (**1991**). The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations. *Journal of Memory and Language*, **30**, 644-663.

Belotel-Grenié, A. and Grenié, M. (**2004**). The Creaky Voice Phonation and the organization of Chinese discourse. *Proceedings of the International Symposium on Tonal Aspects of Languages, With Emphasis on Tone Languages*.

Berkovits, R. (**1993**). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, **21**, 479-489.

Berkovits, R. (**1994**). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, **37**, 237-250.

Blankenship, B. (**2002**). The timing of nonmodal phonation in vowels. *Journal of Phonetics,* **30**, 163-191.

Boersma, P. and Weenink, D. Praat, Doing phonetics by computer (version 5.2.25) [Computer program]. Retrieved from <http,//www.praat.org>.

Bruce, G., Granström, B., Gustafson, K., and House, D. (**1992**). Aspects of prosodic phrasing in Swedish. *Second International Conference on Spoken Language Processing*, 109-112. Banff, Canada.

Calrson, R. and Swerts, M. (**2003**). Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials. *The proceedings of the International Congress of Phonetic Sciences*, Barcelona, Spain.

Carlson, R. Hirschberg, J. and Swerts, M. (**2005**). Cues to upcoming Swedish prosodic boundaries, Subjective judgment studies and acoustic correlates, *Speech Communication* **46**, 326-333.

Chen, M. Y. (**1987**). The syntax of Xiamen tone sandhi. *Phonology Yearbook*, **4**, 109-150.

Chen, M. Y. (**2000**). Tone Sandhi, Patterns across Chinese Dialects. Cambridge University Press.

Chen, Y. and Yuan J. (**2007**). A corpus study of the 3rd tone sandhi in Standard Chinese. *International Journal of Bilingual Education and Bilingualism.*

Cheng, R. L. (**1968**). Tone sandhi in Taiwanese. *Linguistics*, **41**, 19-42.

Choi, Y-Y., Hasegawa-Johnson, M., and Cole, J. (**2005**). Finding intonational boundaries using acoustic cues related to the voice source, *Journal of Acoustical Society of America*, **118** (4), 2579-2587.

Cooper, W. E., and Sorensen, J. M. (**1977**). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, **62**, 682 - 692.

Du, T-C. (**1988**). Tone and stress in Taiwanese. Doctoral dissertation, University of Illinois, Urbana Champaign.

Geers, A. E. (**1978**). Intonation contour and syntactic structure as predictors of apparent segmentation. *Journal of the Acoustical Society of America*, **4**, 411-458.

Gordon, M. and Ladefoged, P. (**2001**). Phonation types, A cross-linguistic overview. *Journal of Phonetics*, **29**, 383-406.

Gordon, M. and Pamela Munro (**2007**). A Phonetic Study of Final Vowel Lengthening in Chickasaw. *International Journal of American Linguistics*, **73**, 3, 293-330.

Grosjean, F. (**1983**). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics*, **21**, 501-529.

Grosjean, F. and Hirt, C. (**1996**). Using Prosody to predict the end of sentences in English and French, Normal and brain-damaged subjects. *Language and Cognitive Processes*, **11**,107-134.

Groz, B. and Hirscheberg, J. (**1992**). Some intonational characteristics of discourse structure. *Proceedings of the 2nd International Conference on Spoken Language Processing*, 429-432.

Hacker, M. and Ratcliff, R. (**1979**). A revised table of *d'* for M-alternative forced choice. Perception & Psychophysics, **26** (2),168-170.

Hanson, H., Stevens, M., Kuo, J. N., Chen, H.-K. J., and Slifka, J. (**2001**). Towards models of phonation. *Journal of Phonetics*, **29**, 451-480.

Hirschberg, J. and Nakatani, C. H. (**1996**). A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 4th International Congress of Spoken language Processing*, 286-293.

Hsiao, Y. (**1991**). Syntax, rhythm and tone: a triangular relationship. Doctoral dissertation, University of California, San Diego.

Hsieh, F.-F. (**2005**). Tonal chain-shifts as anti-neutralization-induced tone sandhi. In S. Arunachalam, T. Scheffler, S. Sundaresan, and J. Tauberer (eds.), *Proceedings of the 28th Annual Pen Linguistics Colloquium, University of Pennsylvania Working Papers in Linguistics 11.1*, pp. 99-112. Philadelphia, Penn Linguistics Circle.

Hsu, C. and Jun, S. (**1998**). Prosodic strengthening in Taiwanese, syntagmatic or paradigmatic? *UCLA Working Papers in Phonetics* **96**, 69-89.

Ischebeck, A. K., Friederici, A. D. and Alter, Kai (**2007**). Processing prosodic boundaries in natural and hummed speech: An fMRI study. *Cerebral Cortex*, **18**, 541-552.

Jun, S. A., and Oh, M. (**1996**). A prosodic analysis of three types of wh-phrases in Korean. *Language and Speech*, **39**, 37-61.

Keating, P., Cho, T., Fougeron, C. and Hsu, C., (**2003)**. Domain-initial strengthening in four languages. *LabPhon VI*, Cambridge University Press.

King, R. **(1988**). Theoretical problems in the description of Taiwanese tone. Ms., Harvard University.

Klatt, D. H. (**1975**). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, **3**,129-140.

Kreiman, J. (**1982**). Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, **10**, 163-175.
Krivokapi, J. (**2007**). Prosodic planning, Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, **35**, 162-179.
Lee, L., and Nusbaum, H. C. (**1993**). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics,* **53**, 157-165.
Lehiste, I. (**1973**). Phonetic disambiguation of syntactic ambiguity. *Glossa*, **7**,107-122.
Lehiste, I., Olive, J. P., and Streeter, L. (**1976**). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, **60**, 1199-1202.
Lehiste I. (**1979**). Perception of sentence and paragraph boundaries. In B. Lindblom and S. Ohman (eds.), *Frontiers of speech communication research* (pp. 191-201). New York: Academic Press.
Liberman, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (**1985**). Measures of the sentence intonation of read and spontaneous speech in American English. *Journal of Acoustical Society of America, 77, 2,*
Lin, H-B. (**1988**). Contextual Stability of Taiwanese Tones. Ms. University of Connecticut.
Lin, H-B. (**1989**). Cues to the perception of Taiwanese tones. *Language and Speech* **32**, pp. 25-44.
Lin, H-B., and Repp, B. H. (**1989**). Cues to the perception of Taiwanese tones. *Language and Speech*, **32**, 25-44.
Lin, H. Y., and Fon, J. (**2009**). Perception of temporal cues at discourse boundaries. *The Proceedings of Interspeech*, Brighton, UK.
Lin, J.-W. (**1994**). Lexical government and tone group formation in Xiamen Chinese. *Phonology*, **11**, 237-275.
Myers, J. and Tsay, J. (**2008**). Neutralization in Taiwan Southern Min Tone Sandhi. *Interfaces in Chinese Phonology: Festschrift in honor of Matthew Y. Chen on his 70th birthday*, 47-78. Language and Linguistics Monograph Series Number W-8. Taipei, Taiwan: Academia Sinica.
Nepor, M., and Vogel, I. (**1986**). *Prosodic phonology*. Dordrecht, Foris.
Nooteboom, S. G., Brokx, J. P. L. and de Rooij, J. J. (**1978**). Contributions of prosody to speech perception. In W. J. M. Levelt and G. B. Flored d'Arcais (eds.) *Studies in the perception of language* (pp. 75-107). Chichester: John Wiley & Sons.
Pan, H. (**2003**). Prosodic hierarchy and nasalization in Taiwanese. *Proceedings of the 15th ICPhS*, pp. 575-578.
Pan, H. (**2006**). Boundaries and Tonal Articulation in Taiwanese Min. *Proceedings of Speech Prosody***, 51**.
Peng, S.-H. and Beckman, M. (**2003**) Annotation conventions and corpus design in the investigation of spontaneous speech prosody in Taiwanese. *Proceedings of SSPR 2003*, pp. 17-22.
Peng, S.-H. (**1997**). Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics*, **25**, 371-400.
Pijper, J. R. and Sanderman, A. A. (**1994**). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, **96**, 2037-2047.
Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. (**1991**). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America* **90**, 2956-2970.

Ratcliff, R., McKoon, G. and Verwoerd, M. (**1989**) A Bias Interpretation of Facilitation in Perceptual Identification. Journal of Experimental Psychology. **15**, 3, pp. 278-287.
Sanderman, A. (**1996**). Prosodic phrasing. Production, perception, acceptability and comprehension. Doctoral dissertation. Eindhoven University of Technology.
Scott, D. R. (**1982**). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, **71**, 996-1007.
Stanislaw H. and Todorov, N. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers,* **31** (1), 137-149.
Streeter, L. A. (**1978**). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, **64**, 6, 1582-1592.
Shih, C-L. (**1986**). The Prosodic Domain of Tone Sandhi in Chinese. Doctoral dissertation, University of California, San Diego.
Shih, C-L. (**1997**). Mandarin third tone sandhi and prosodic structure. *Studies in Chinese Phonology* (ed.) Jialing Wang and Norval Smith. Walter de Gruyter, 81-123.
Shue, Y.-L. (**2010**). The Voice Source in Speech Production: Data, Analysis and Models. Doctoral dissertation, University of California, Los Angeles.
Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). VoiceSauce, a program for voice analysis. *Proceedings of the 17th ICPhS*, 1846-1849.
Speer, S. , Kjelgaard, M. M., and Dobroth, K. M. (**1996**). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, **25**, 247-268.
Speer, S. R., Shih, C-L., and Slowiaczek, M. K. (**1989**) Prosodic structure in language understanding, Evidence from tone sandhi in mandarin. *Language and Speech*, **32**, 337-354.
Strangert, E., and Heldner, M. (**1995**). The labeling of prominence in Swedish by phonetically experienced transcribers. *The proceedings of 8th ICPhS*, 13-19, Stockholm, Sweden.
Streeter, L. A. (**1978**). Acoustic determinants of phrase boundary location. *Journal of the Acoustical Society of America*, **64**, 1582-1592.
Studdert-Kennedy, M. (**1980**). Speech perception. *Language and Speech,* **23**, 45-66.
Swerts, M. (**1997**). Prosodic features at discourse boundaries of different strengths. *Journal of the Acoustical Society of America*, **101**, 514-521.
Tanner, W.P. and Swets, J. A. (**1954**). A decision-making theory of visual detection. *Psychological Review,* **61**, 401-409.
t'Hart R., Cohen A. (**1990**). *A perceptual study of intonation: an experimental-phonetic approach to speech melody.* Cambridge University Press.
Tsay, J., Myers, J., and Chen, X-J. (**1999**). Tone Sandhi as Evidence for Segmentation in Taiwanese. *The Proceedings of the 30th Annual Child Language Research Forum.*
Tsay, J. and Myers, J. (**2001**). Processes in the Production of Taiwanese Tone Sandhi: an Acoustic Phonetic Study. *The Proceeding of 5th National Conference on Modern Phonetics*, Peking: Tsinghua University, 233-237.
Tseng, C. Y. (**1981**). An Acoustic Phonetic Study of Tones in Mandarin Chinese. Doctoral dissertation, Brown University.
Tseng, C.-C. (**1995**). Taiwanese Prosody, An integrated analysis of acoustic and perceptual data. Doctoral dissertation. University of Hawai'i.
Vance, T. J. (**1977**). Tonal distinctions in Cantonese. *Phonetica*, **34**, 93-107.
Wagner, M. and Watson, D. G. (**2010**). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, **25**, 7, 905-945.

Wang, W. S-Y. (**1967**). Phonological features of tone. *International Journal of American linguistics*, **33**, 93-105.
Wang, S. and Fon, J. (**2012**). Durational Cues at Discourse Boundaries in Taiwan Southern Min. Proceedings of 6th International. Conference on Speech Prosody, 599–602
Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M and Price, P. (**1992**). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America,* **91**(3), 1707-1717.
Wright, M. S. (**1983**). A Metrical Approach to Tone Sandhi in Chinese Dialects. Doctoral dissertation. UMass Amherst.
Yip, M. (**1980**). The Tonal Phonology of Chinese. Doctoral dissertation. Massachusetts Institute of Technology.
Yoon, T-J. Cole, J. and Hasegawa-Johnson, M. (**2007**). On the edge: acoustic cues to layered prosodic domains. *ICPhS proceedings*, 1017-1020.
Oliva, Aude, 9.63/F09, Laboratory in Visual Cognition, Fall 2009. (Massachusetts Institute of Technology, MIT OpenCourseWare), http,//ocw.mit.edu (accessed May 4th, 2013). License, Creative Commons BY-NC-SA.