A Weighted-Constraint Model of F0 Movements

by

Hyesun Cho

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author .....
Department of Linguistics and Philosophy
September 15, 2010

Certified by .....
Edward Flemming
Associate Professor of Linguistics
Thesis Supervisor

Certified by .....
Michael Kenstowicz
Professor of Linguistics
Thesis Supervisor

Accepted by .....
Irene Heim
Professor of Linguistics, Department Head

A Weighted-Constraint Model of F0 Movements

by

Hyesun Cho

Submitted to the Department of Linguistics and Philosophy
on September 15, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation develops a grammar of phonetic implementation of phonologically significant F0 (pitch) events, which is applicable across languages. Through production studies of various languages, we show that phonetic universals exist which govern phonetic realization of the phonological representations of tones. In the previous literature, there have been two conflicting views concerning tonal timing: tones are aligned with respect to segments (the Segmental Anchoring Hypothesis) or tones occur at a fixed interval from other tones (the Constant Duration Hypothesis). In this dissertation, the two hypotheses are tested in languages with various tonal phonologies: Seoul Korean (phrasal boundary tone), Tokyo Japanese (lexical pitch accent), Mandarin (lexical tone), and English (intonational pitch accent). In all languages, both tendencies to maintain segmental alignment and a target duration for pitch rises are simultaneously observed. We thus adopt a weighted-constraint model (Flemming, 2001) where segmental alignment and target duration are interpreted as weighted constraints. In this model, timing of tones is determined to minimize the summed cost of violation of these conflicting constraints. Mixed-effects models were fitted to the data to obtain the actual weights in each language. Relative weights of the constraints reflect cross-linguistic differences in the alignment of tones. The relative weights of constraints in the phonetic realization grammar are not random but systematic, reflecting the phonological nature of tones in each language.

The experimental studies in this dissertation show that tonal alignment patterns depend on phonological status and context of tones. Lexically-contrastive tones (Japanese accented words, Mandarin lexical tone) or prominence-lending tones (English pitch accents) are more strictly aligned with respect to their anchoring points than phrasal boundary tones (Seoul Korean, Japanese unaccented words), if other conditions are equal. Tones show different alignment patterns depending on phonological context: tones are more strictly aligned in word-final context than in word-medial context in Japanese accented words, and in lexical-tone context than in neutral-tone context in Mandarin. In addition, languages show different phonetic realization patterns depending on whether contour tones are contrastive in the language (Mandarin and English) or not (Korean and Japanese). These results point to the fact that details of phonetic realization of tones are determined by language-specific phonetic realization grammar, rather than by default universal rules.

Thesis Supervisor: Edward Flemming
Title: Associate Professor of Linguistics

Thesis Supervisor: Michael Kenstowicz
Title: Professor of Linguistics

Acknowledgments

This dissertation would not have existed without help of many people all of whom I am very blessed to know, work with and learn from. First of all, I am very grateful to my respectful professor Edward Flemming. He has been an insightful, devoted, understanding, and supportive advisor throughout all my years at MIT. He was very helpful at various stages and aspects of my dissertation writing. He led me to develop the experimental results I had for my first generals paper with his weighted-constraint model framework, from which the whole idea of this dissertation originated. He helped me with experimental as well as theoretical aspects of this dissertation. I have learned statistical methods and modeling insights from him. We had to test many models to arrive at the best models, the ones presented in this dissertation are only the part of many models that we tried. In doing so, Professor Flemming gave me various ingenious ideas of the models worth trying. He was always available whenever I needed his advice during the dissertation writing and modeling. Whenever I was stuck with problems, he always gave me brilliant solutions which I would never be able to figure out myself. Words are not enough to describe how much grateful I am to him.

I am also very grateful to my advisor Michael Kenstowicz, who has always been supportive from the very beginning of my life at MIT. Thanks to him, I became interested in tones in languages. As a specialist in tonal phenomena, he introduced to me very important papers that have become essential in shaping of the initial ideas for my first generals paper which is eventually developed into this dissertation. He helped me with the experimental designs, read my dissertation very carefully, and raised important questions regarding phonological implications of this research.

I am also grateful to my committee member Adam Albright who taught me the importance of 'ruminating' or 'deep thinking' while I was writing my second generals paper, from which I learned how to write and develop arguments. I thank him for carefully reading my dissertation and giving me valuable advice, in particular, concerning the phonological representations that are compatible with the model proposed in this dissertation. I thank Donca Steriade, who always delightfully listened to me with her pleasing manner and encouraged me whenever I talked with her about my work.

I also thank Stefanie Shattuck-Hufnagel and Jonathan Barnes for their helpful advice and interest in my research. I thank Nanette Veilleux for teaching me how to use the elbow script and giving me other useful tips. I thank the audience at the 12th Conference on Laboratory Phonology for their helpful comments on this research.

I thank professor Jae-Young Lee, my advisor at the English Department at Seoul National University, who encouraged me to study phonology and have always been supportive. I thank professor Jongho Jun at the Linguistics Department at Seoul National University, who has always kindly supported me, especially when I was writing my second generals paper. I am also grateful to professor Minhwa Chung at Seoul National University for providing me with an invaluable opportunity to continue research after MIT. I thank Ho-Young Lee from whom I first learned experimental phonetics at Seoul National University. I thank professors at Ewha Womans University, especially the late professor Seunghwan

Lee from whom I first learned a course in phonetics.

I thank Maria Giavazzi, my best friend whom I was very lucky to have for the past five years. I thank her for her unfailing friendship, her warm heart and kindness throughout all these years. Including her, I thank my classmates, 'ling05': Patrick Jones, Gillian Ghallager, Jonah Katz, Jessica Coon for their friendship and help. I have been blessed to have such kind, considerate, and friendly classmates.

I thank my dear friends Giorgio Magri and Lola De Hevia for their warmest friendship and nice Italian dinners. I thank all my friends at the Linguistics Department at MIT, especially, my dear friends Samer Al Khatib and Guillaume Thomas who always made me laugh during my dissertation writing, Youngah Do and Daeyoung Sohn who always cheered me. I thank Yasutada Sudo for his kind help with my Japanese experiment as a consultant as well as a subject, checking through the speech materials and finding the words that I needed for the experiment. I also thank him for his essential help with the bibliography in Latex. I also thank Junya Nomura for helping me with the Japanese speech materials. I thank Michael Yoshitaka "Mitcho" Erlewine for his kind help with my Chinese experiment and for his helpful Latex workshop. I thank Claire Halpert, Bronwyn Bjorkman, Jeremy Hartman, and Patrick Jones for willingly participating in my English experiment.

I am thankful to Samer Al Khatib, Yasutada Sudo, Guillaume Thomas, Sasha Podobryaev, Natasha Ivlieva, Rafael Nonato, Maria Giavazzi, Jonah Katz, and Mitcho Erlewine for showering me with many drinks and dinners to celebrate my becoming a doctor.

I am very grateful to Youngah Do and Jinyoung Kim for providing me with a comfortable place to stay in my last ten days in Cambridge when I was revising this dissertation.

My thanks extend to all who kindly participated in my experiments. In particular, I thank Korean speakers Young-Sook Jung, Jae-Young Song, Sung-Min Sohn, Woo-Geun Chung, Ki-Min Jun, Han-Bi Na, Jin-Young Kim, Mr. Baek Yong-Ho and Mrs. Lee Eun-Kyung, Japanese speakers Megumi Matsudani, Ayaka Sugawara, and Yuki Sakurai sensei, and a Mandarin speaker professor Yuncheng Zhou. I also thank my dear friend Tingting Mao for her kind help in finding Mandarin Chinese speakers, and Ikue Shingu sensei for helping me find Tokyo Japanese speakers.

Last but not least, I thank my dear parents, Jongkwan Cho and Inhee Lee for their infinite and warmest love and invaluable wisdom that have strongly supported me throughout the program. I thank my little brother Hongjin Cho and my sister-in-law Kongju Seo, my dear sisters Heesook Cho, Heebok Cho, and Sunhwa Cho for their warmest love and care.

I thank God for allowing me the opportunity to meet with all these invaluable people in my life and for guiding me throughout the five year's study at the MIT Linguistics department.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation aims to develop a grammar of phonetic implementation of phonologically significant F0 (pitch) events, which is applicable across languages. Through production studies of various languages, we show that phonetic universals exist which govern phonetic realization of phonological representations of tones. Cross-linguistic similarities are explained by a set of constraints that are common across languages, so tendencies to satisfy these constraints are observable in all languages. At the same time, we demonstrate that the extent to which these tendencies are manifested differs from language to language. This means that not all the constraints are of the same importance in different languages. Cross-linguistic differences are explicable by assigning different weights to constraints where the weights reflect relative importance of the constraints. We show that the relative weights of constraints in the phonetic realization grammar are not random but systematic, reflecting the phonological nature of tones in each language.

More specifically, the main theme of this dissertation is the timing of pitch targets (L, H) in a segmental string. The research question is how the timing of tones is determined: whether the tones occur with respect to segments, or at a fixed interval from the other tones. The first approach is known as the "Segmental Anchoring Hypothesis (SAH)" (Ladd et al., 1999). According to this hypothesis, pitch movements are analyzed into L and H level targets and these level targets are aligned ("anchored") with respect to certain segmental landmarks ("anchors"). The timing of the tones follows the timing of their respective segmental anchors, so the L and H tones comprising a rising movement are independent of each other. It has been claimed that the segmental alignment of a tone remains stable regardless of changes in speech rate, syllable structure, or prosodic context (Prieto et al., 1995; Arvaniti et al., 1998; Xu, 1998; Ladd et al., 1999; Igarashi, 2004; Dilley et al., 2005; Ishihara, 2006). This dissertation argues against the basic assumption (the independence of pitch level targets) and the claim (the stability of tonal alignment) of the SAH. That is, we will show that the L and H tones comprising a rising F0 movement are not independent of each other and that the segmental alignment is systematically and gradually affected by speech rate. The second approach to tonal timing is that tones may be timed with respect to each other, regardless of the segments. For example, in the English L*+H- bitonal pitch accent, the H- occurs a fixed time interval after L*, regardless of the segmental/syllable characteristics of the materials following the accented syllable (Pierrehumbert, 1980: 80).

We probe these two different approaches in this dissertation. We show that both factors (segmental anchoring and fixed rise duration) simultaneously affect tonal timing in languages we examined. Furthermore, we ask whether the timing of the tones differs depending on the phonological status (lexical or boundary) and context by examining languages that have varying phonological statuses of tones. From the experimental studies of these languages, we will show that detailed phonetic realization patterns of tonal timing differ depending on phonological status and context of the tones. In this chapter, we review the segmental anchoring approach and the fixed interval approach in more detail, show the conflicting predictions of these approaches, and outline the hypotheses that will be tested throughout the dissertation.

1.1 The models of F0 movements

This section reviews two conflicting hypotheses concerning F0 movements: the Segmental Anchoring Hypothesis and the Constant Duration Hypothesis. The difference lies what is stable in the realization of F0 movements: segmental alignment of the tones or fixed duration of a pitch movement, in other words, whether tones are timed to occur at specific segmental positions or to occur at a fixed interval from other tones.

1.1.1 The Segmental Anchoring Hypothesis

According to Bruce (1977), F0 contours are most appropriately analyzed as a sequence of local minima and maxima: "reaching a certain pitch level at a particular point in time is the important thing, not the movement (rise or fall) itself" (Bruce, 1977: 132). That is, phonologically significant F0 events such as F0 minima and F0 maxima are aligned with regard to segmental landmarks such as the beginning and the end of the accented vowel. This view is held by the segmental alignment literature originating from Arvaniti et al. (1998). Strong effect of the segmental alignment of the tones was first found in Greek by Arvaniti et al. (1998). They showed that the H peaks of the prenuclear rising pitch accent in Greek were always found right after the onset of the post-accentual vowel. Despite the varying duration of the accented syllables (e.g. long: [paˈɾemvasi], [ˈvjenume] versus short: [likˈlemona]), L troughs were consistently aligned at the onset of the accented syllable and H peaks were consistently aligned on average $17ms$ after the onset of the first postaccentual vowel. This means that the L and H tones are aligned with respect to segmental landmarks, independently of each other.

The "Segmental Anchoring Hypothesis (SAH)" (Ladd et al., 1999) states that the beginning and end of pitch rises and falls are aligned to segmental landmarks. Ladd et al. (1999) examined the stability of the alignment of F0 minima and maxima in Standard Southern British English by manipulating speech rate: fast, normal, and slow. A prediction of strict segmental anchoring is that the duration of the rise will decrease as speech rate increases and thus the anchors of the L and H tones get closer to each other (for illustration, see Figure 1-1a). So, the duration of a rise is expected to change depending on speech rate, while segmental alignment remains relatively stable.

As predicted by the SAH, Ladd et al. (1999) found that the duration of the rising movement decreased as speech rate increased. In addition, they tested the effects of speech rate on the difference between the timing of F0 extrema and several segmental landmarks. The distance between L (F0 minimum) and the beginning of the onset consonant of stressed syllables was not significantly affected by speech rate, which means that L is anchored with respect to the onset of the stressed syllable. For the alignment of H, several anchoring positions were tested: alignment of H relative to C1 (offset of vowel of stressed syllable), alignment of H relative to V1 (onset of unstressed vowel), and alignment of H as a proportion of the interval from C1 to V1 (H-C1/V1-C1). Among these tested points, the proportional point in the stressed syllable was least affected by speech rate. The proportion was speaker-dependent, ranging from 0.42 to 1.84 in normal speech. The effect of speech rate on the proportion was not significant for four out of six speakers, which means that the alignment of the tone with respect to the proportional point was stable regardless of speech rate for those speakers. Based on these results, Ladd et al. (1999) proposes the Segmental Anchoring Hypothesis stating that the beginning and the end of a pitch movement are stably aligned ("anchored") with respect to segmental landmarks ("anchors").

The SAH has received extensive support in the literature from a variety of languages: Mandarin Chinese (Xu, 1998), British English (Ladd et al., 1999), Russian (Igarashi, 2004), Northern and Southern German (Atterer and Ladd, 2004), English (Dilley et al., 2005), and Tokyo Japanese (Ishihara, 2006). These studies show that the alignment is not affected by changes in speech rate or syllable structure; instead, shape properties such as the duration of a rise change significantly. In Mandarin, the Rising tone maintains consistent alignment to the associated syllables (Xu, 1998). For example, the F0 peak of the Rising tone always occurs near the offset of the syllable that carries the tone, and the onset of the F0 rise always occurs near the center of the syllable, regardless of syllable structure (whether the syllable has a coda or not) or speech rate. Ishihara (2006) varied syllable structure and speech rate to examine stability of tonal alignment of pitch accents of initial-accented words in Tokyo Japanese. The structure of the initial accented syllables was varied (open or closed, long or short vowel). F0 peaks were consistently aligned with regard to the end of the first mora or the beginning of the second mora. Russian also supports the SAH. Igarashi (2004) showed that the duration of the F0 rise in Russian prenuclear rising pitch accents decreased as speech rate increased. The alignment of L and H was stable regardless of speech rate: that is, the interval between L and the onset of the accented syllable was not significantly affected by speech rate (fast, normal, slow), and the difference between H and the onset of the vowel in the post-accentual syllable was not significantly affected by speech rate.

A crucial claim of the SAH is that the pitch level targets are independent of each other. The SAH states that L and H tones are aligned to their respective anchors, that is, the timing of a tone follows the timing of its anchor. This implies that L and H tones comprising a rising or falling pitch movement are timed independently.

1.1.2 The Constant Duration Hypothesis

The intonation researchers at the Institute for Perception Research in Netherlands ('IPO' or 'The Dutch School of intonation, 1965-1995) held the view that global F0 curves are com-

binations of line segments, rather than pitch level targets ('t Hart and Cohen, 1973; 't Hart and Collier, 1975; 't Hart et al., 1990). They argued that "there are no pitch 'levels'", and "the speaker *intends* to produce audibly gradual pitch-transitions." ('t Hart et al., 1990: 75).

In the IPO model, properties of F0 movement shape, such as slope, duration, and magnitude, are considered as primary descriptive units. If the properties of F0 shapes are the primary units, the shape properties are not supposed to vary considerably due to factors such as speech rate or segmental structure. Such approaches are known as the "constant slope" and "constant duration" hypotheses (Ladd et al., 1999). In this dissertation, we refer to this kind of approach as the "Constant Shape Hypothesis (CSH)". In particular, the hypothesis that the duration of a rise is constant will be referred to as the "Constant Duration Hypothesis (CDH)". The Constant Duration Hypothesis states that a rise has a fixed duration. The CDH is at odds with the SAH, because the SAH predicts that the rise duration will change depending on speech rate, whereas the CDH predicts that the duration of a rise will be relatively stable.

In the IPO model, pitch movements are specified for their duration. 't Hart et al. (1990: 73) decomposed pitch movements into several features: direction (rise, fall), timing with regard to syllable boundaries (early, late, and very late), rate of change (fast, slow), and size (full, half). Five types of rises are represented with numbers (1,2,3,4,5), falls are represented with letters (A,B,C,D,E). Each rise and fall is specified for timing, rate of change, and size. The categories are not impressionistic labels, but have a precise phonetic content, which is derived from the stylization of real utterances. For example, a 'fast, early, full rise' (Type 1 Rise) has a rate of 50 ST/s and a duration of 120 ms (corresponding to a rise of 6 semitones), and its peak is timed 50ms after the onset of the vocalic nucleus of the syllable. A 'half rise' (Type 5 Rise) has a duration of 60ms (a rise of 3 semitones).

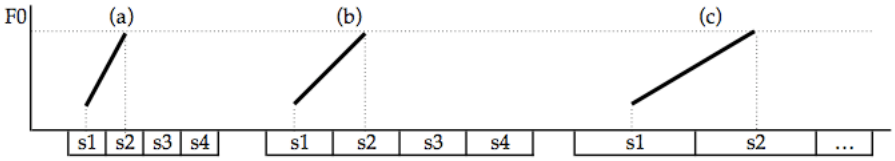
Empirical evidence for a relatively constant rise duration has been found in Northern Finnish (Suomi, 2009). In Northern Finnish, segmental duration is adjusted to accommodate a constant rise duration. The duration of the accentual F0 movement is stable, but the duration of associated morae varies. The loci of the tones are postulated in the first and second mora of the word, and the duration between the loci remains constant across different word structure.

The stability of F0 features other than rise duration, such as slope of pitch movements, has been found in a number of languages. Thorsen (1984) reported that in Danish, the F0 contour is invariant and segments and syllables are superimposed on the F0 contour; as a results, short vowels have a falling tune and long vowels have a falling-rising tune. In Dutch, the slope of the fall remained relatively stable for a given speaker under variation in speech rate (Caspers and van Heuven, 1993: 169). In French, speakers reduced pitch levels when speaking fast, which means that slope of pitch rises was relatively stable (Fougeron and Jun, 1998). A similar pitch reduction pattern was observed in Russian (Igarashi, 2004). In Mandarin (Xu, 1998), not only both the onset and F0 peak of the rising tone are strictly aligned, but also the slope of the rise does not vary systematically with either syllable structure or speaking rate.

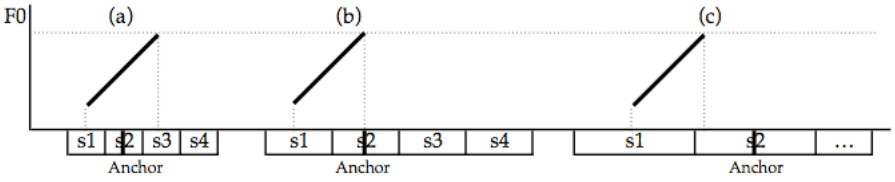
1.1.3 Conflicting predictions

The Segmental Anchoring Hypothesis and the Constant Duration Hypothesis are conflicting. If tonal targets are anchored to segments, the duration of pitch rise has to vary depending on changes in segmental duration due to changes in speech rate or inherent segment duration, e.g. the faster the rate, the shorter the rise duration. This means that constant duration cannot be maintained. On the other hand, if the duration of pitch rise remains stable, segmental anchoring cannot be maintained when the duration between the anchors changes due to syllable structure or speech rate.

Under the SAH, the tones are considered to be independent of each other. The shape properties such as slope and duration are derivable from interpolation between these targets. The duration of a pitch rise is determined by the distance between the anchoring points of the tones. In addition, in the SAH literature, it is also argued that the magnitude of a rise is relatively stable across speech rates (Ladd et al., 1999), or at least, the duration does not affect the magnitude of the rise (Ladd, 2004). If the magnitude of the rise is constant, the slope of the rise gets steeper at faster speech rates. Thus, the following is predicted: (i) the duration of the rise decreases as speech rate gets faster and the anchors get closer to each other, and (ii) the slope of the rise increases at fast speech rates, as illustrated in Figure 1-1a. These predictions are observed in prenuclear rising pitch accents in Standard British English (Ladd et al., 1999).



(a) The Segmental Anchoring Hypothesis



(b) The Constant Shape Hypothesis

Figure 1-1: Conflicting predictions on tonal timing: (a) Prediction of the Segmental Anchoring Hypothesis, (b) Prediction of the Constant Shape Hypothesis

On the other hand, if the shape of the pitch rise remains stable, it is predicted that the peak of the rise will be found later than the anchoring point when speaking fast. This is illustrated in Figure 1-1b. The rise peak will be found later than the anchor in fast speech (Figure 1-1b(a)), at the anchor in normal speech (Figure 1-1b(b)), and before the anchor in slow speech (Figure 1-1b(c)). It may not seem plausible that the shape of the pitch movement remains completely unchanged, but we may expect to observe some tendency similar to this. In fact, this kind of effect is observed in Spanish (Prieto and Torreira, 2007). Prieto

and Torreira (2007) found that the timing of the prenuclear peaks in Spanish varies depending on speech rate and segmental structure (whether the syllable is open or closed). The prenuclear peaks occurred later than a segmental landmark (the end of accented vowel, the end of accented syllable) when speech rate is fast or when the syllable is open. However, they did not interpret such results as an effect of a target duration. They attributed such variations to differences in the coordination of the articulatory gestures at the beginning and the end of a syllable (Prieto and Torreira, 2007: 493 ff.). That is, the coordination between tonal gestures and supraglottal gestures is less tight later in the syllable than at the beginning of the syllable. So, the H is less strictly aligned than the L. In our experiments, we found that in many languages, the alignment of L also varies systematically, in fact, varies more than H does, so the gestural account will not be sufficient, if not inadequate.

Studies have found that the duration of a rise changes depending on experimental manipulations of speech rate (Ladd et al., 1999). The change of duration depending on speech rate was considered as evidence for segmental anchoring, and at the same time, as evidence against constant duration (Ladd et al., 1999; Igarashi, 2004; Dilley et al., 2005). Dilley et al. (2005) tested two competing hypotheses (the SAH and what they called *constant interval hypothesis*). They manipulated the timing of L in English L+H* pitch accents as in e.g. *Norma Nelson* ("early boundary") and *Norman Elson* ("late boundary"). Prior to this study, Ladd and Schepman (2003) has shown that in these phrases, the F0 valley occurred earlier when the boundary was earlier, and later when the boundary was later. Using the same speech materials, Dilley et al. (2005) tested the hypothesis that if L+H* has a constant duration, the timing of H* will be affected by the manipulation of L; on the other hand, if the two targets in L+H* are segmentally anchored, timing of H* will not be affected by the manipulation of the timing of L. If the SAH is correct, the time interval between L and H* will be longer if L occurs earlier. They found that the duration from L to H* is significantly greater when L is early than when it is late. So they concluded that their results support the SAH; and the constant interval hypothesis cannot be supported.

It has been assumed that either the CDH or the SAH is universally correct because they make conflicting predictions, but it is plausible that both factors may simultaneously affect the timing of tones. This dissertation provides evidence showing that both factors are present in a language. The reason why the Constant Shape/Duration Hypothesis has not been paid much attention to may be because in many languages, such as English and Dutch, segmental anchoring effects were substantially large so that even if there had been a tendency to the constant duration, it might not have been readily observable. That is, both factors are present, but because of the magnitude of segmental anchoring effects, it is easy to overlook CDH effects.

1.2 Hypotheses in this dissertation

Segmental alignment has been considered stable in a range of languages: Greek (Arvaniti et al., 1998), English (Ladd et al., 1999; Dilley et al., 2005), Dutch (Ladd et al., 2000), Northern and Southern German (Atterer and Ladd, 2004), Mandarin (Xu, 1998), Japanese (Ishihara, 2006), Russian (Igarashi, 2004). We were able to find only a little evidence for

constant duration: Spanish (Prieto and Torreira, 2007: based on our interpretation, not the authors') and Finnish (Suomi, 2009). However, the tones in the segmental anchoring languages are lexically contrastive (Japanese, Mandarin) or associated with prominent syllables (English, Greek). A strong tendency for segmental alignment might have been due to the phonological status of tones in these languages.

Thus, it may be possible to observe effects of constant duration if we examine tones that are not contrastive or prominence-lending. The Seoul dialect of Korean is a good language to test the conflicting hypotheses, because it does not have lexical tone or stress, and only phrasal tones exist. Alignment is thus expected to be less strict, so it is easier to observe the effects of constant duration, if they exist. For this reason, in Chapter 2, the timing of four tones in the LHLH intonational pattern in Seoul Korean is examined. In the experimental results, the timing of the phrase-initial L and H tones shows tendencies to segmental anchoring as well as constant duration. In addition, the timing and scaling of the tones are thoroughly analyzed. In the phrase-initial L and H, the effects of segmental anchoring are observed. The effects of a target duration are also observable, as expected for a phrasal tone with less strong segmental anchoring. That is, the H peaks are found later than the anchoring point in fast speech, as predicted in Figure 1-1b. In addition, we found that the L trough tends to occur earlier than their anchors. This means that the rise starts earlier and terminates later than the expected locations when less time is available to execute the rise due to a fast speech rate. These results are interpreted as the effects of a target duration.

This means that the timing of L and H tones is determined by segmental anchoring as well as target duration, and so we propose a model with alignment and duration constraints. That is, both L and H tones are aligned to their respective anchors: Align(L) and Align(H). At the same time, the Duration constraint requires the duration between L and H to be constant. These are formalized as the interaction of weighted constraints for scalar representations (Flemming, 2001). The actual timing is determined as the values that minimize the cost of violation of these constraints. The model is developed in Chapter 3. Because this model consists of the alignment and duration constraints, it is referred to as the *Alignment-Duration* (AD) model. What differentiates this model from the Segmental Anchoring Hypothesis is that in the AD model, the L and H tones comprising a rising movement are considered to be dependent on each other. We compare the AD model with a model with independently-aligned tones (the Independent-Alignment (IA) model). It turns out that the AD model is significantly better than the IA model in predicting the timing of L and H tones in Seoul Korean. This means that the tones comprising a rising movement are not independent of each other.

In Chapter 4, the proposed model is applied to languages with varying phonological status of tones: Japanese (lexical pitch accent), Mandarin (lexical tone), and English (intonational pitch accent). These languages are chosen because the tones in these languages have phonological status different from Seoul Korean and from one another. Two hypotheses are tested: first, that both the alignment and duration constraints exist across languages; second, that the tonal alignment patterns vary depending on phonological status and context of tones, and the cross-linguistic differences are reflected in the relative weights of the constraints. More specifically, we test the hypothesis that lexically specified tones are more

strictly aligned than phrasal tones. It is also expected that the alignment pattern will vary depending on phonological context: the word-medial or word-final context of the Japanese accentual peak, and the lexical tone or neutral tone context of the Mandarin Rising tone. It is predicted that in Japanese, word-final pitch peaks will be more strictly aligned due to the upcoming word boundary, and that in Mandarin, the tone alignment will be less strict in the toneless (neutral tone) context.

1.3 The weighted-constraint model of phonetic implementation

The literature on segmental anchoring versus constant shape has assumed that one is universally correct, but given mixed evidence from many languages, this is untenable. Welby and Lœvenbruck (2005: 2371-2371) found that none of their results overwhelmingly support either hypothesis. There was a considerable variation in the alignment pattern of L2 (the second L) in the French LHLH accentual phrase; so segmental anchoring was not supported. Yet they did not find support for the constant shape hypothesis either; the rise duration was not constant. We also observed variations in the alignment of rise peaks in Seoul Korean, but we propose that the variations are systematic, and in fact the variations are the combined effects of both segmental anchoring and constant duration. With appropriate methods, the effects of both factors can be modeled in a quantitatively precise way. We propose a constraint-based approach using weighted constraints for phonetic implementation (Flemming, 2001). Given the observed variations in tonal alignment and rise duration, we interpret segmental anchoring and constant duration as violable constraints, rather than as inviolable principles. The premise of our constraint-based approach is that differences in phonetic realization patterns in languages are differences in the phonetic implementation grammar in which segmental anchoring and constant duration are violable constraints. The weights of the constraints reflect the relative importance of the targets.

We found that both alignment and shape targets exist in the languages we examined, but the constraint weights differ depending on the nature of the tones. For example, the effects of both segmental anchoring and constant duration are observed in the phrase-initial tones in Seoul Korean and the Rising tone in Mandarin, but the effect of segmental anchoring is much stronger in Mandarin than in Seoul Korean. These results suggests that phonetic universals exist which govern phonetic realization of tones, but the degree to which the constraints are realized are language-specific, so the language-specific phonetic details cannot be provided by universal rules by default. The weighted-constraint model provides a framework for the language-specific phonetic implementation of tones using constraints common across languages, with differences reflected in the constraint weights.

In brief, the organization of this dissertation is as follows. In Chapter 2, the timing and scaling of each part of the LHLH intonational pattern of Seoul Korean are thoroughly examined. Based on the results of Seoul Korean, in Chapter 3, the Alignment-Duration model is developed. In Chapter 4, the proposed model is applied to other languages with varying phonological status and context of tones: Japanese, Chinese, and English. Chapter 5 is the conclusion.

Chapter 2

The Seoul Korean Accentual Phrase

2.1 The Seoul dialect of Korean

In Chapter 1, we have discussed conflicting evidence that supports either the Segmental Anchoring Hypothesis (British English (Ladd et al., 1999), Southern and Northern German (Atterer and Ladd, 2004), Tokyo Japanese (Ishihara, 2006), and Mandarin (Xu, 1998)) or the Constant Duration Hypothesis (Northern Finnish (Suomi, 2009)). We suggested that strong segmental anchoring found in these languages may be due to the fact that the tones in these languages are lexically contrastive or prominence-lending. Thus, even if there had been any tendency to a constant duration, the effects might have been obscured. On the grounds that the rise duration changed depending on speech rate or segmental structure, the CDH has been rejected (Arvaniti et al., 1998; Ladd et al., 1999; Dilley et al., 2005). However, the previous results might just mean that the effects of segmental anchoring are easier to observe than the effects of constant duration. If we examine tones that are not lexically contrastive or prominence-lending, it may be easier to observe the effects of constant duration, because segmental anchoring of such tones is expected to be less strong.

For this reason, we examine the tones in the Seoul dialect of Korean in this chapter. Alignment is expected to be less strict in Seoul Korean than in other languages, because Seoul Korean does not have lexically specified tones, pitch accents or stress; it has only phrasal level tones, which are, crucially, not associated with prominence within the phrase. Previous studies have shown that the phrase-initial rise peak in Seoul Korean can be variably timed, with the second syllable or with the third syllable in the phrase when the phrase is more than four syllables long (Jun, 1996; Lee and Kim, 1997), which supports variability in segmental alignment of the peak. In other languages, different locations of peaks can change the meaning of words e.g. pitch accent languages such as Japanese, or different locations of stress will make English words ungrammatical or unintelligible. On the other hand, the locations of H peaks are not contrastive in Korean words and phrases. Therefore, we conjecture that alignment is less important in Seoul Korean than in other languages and thus there will be more room for temporal variation. If alignment is less strict, it may be easier to observe the effects of shape constraints that might have been obscured by the strong effects of alignment in other languages.

According to K-ToBI (Korean Tones and Break Indices, Jun (2000)), the intonational

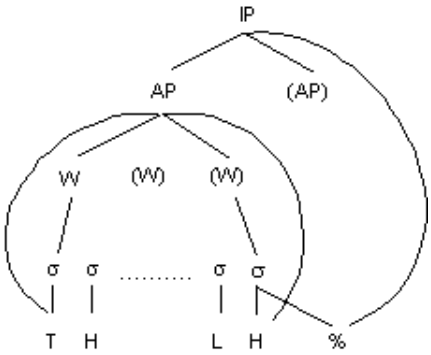


Figure 2-1: Intonational structure of Seoul Korean (Jun, 2000)

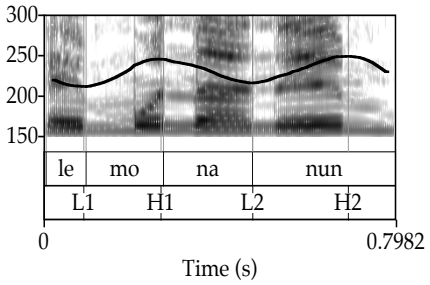


Figure 2-2: Example of the Seoul Korean LHLH Accentual Phrase

units of Seoul Korean are IP (Intonational Phrase) and AP (Accentual Phrase). An AP is a prosodic level smaller than an IP and larger than a phonological word, as illustrated in Figure 2-1. An AP longer than three syllables is marked by a rise at the beginning and another rise at the end, and tones in the syllables in between are underspecified, so it has the [TH..LH] tonal pattern (T=H if the AP initial segment is aspirated or tense, T=L otherwise). Thus, an AP with four syllables has the LHLH tonal pattern if the AP-initial segment is lax or sonorant. An AP is marked by a boundary tone at the beginning and the end of the AP (in the figure, they are associated with the AP node). The AP-final tone can also be an IP boundary tone if there is only one AP in the IP (associated with the IP node).

Figure 2-2 shows an example of a four-syllable AP, *lemonanun* [1][lemona-nin], consisting of a three-syllable content word [lemona] 'Lemona (a name of a product)' followed by a one-syllable topic marker [nin]. In this thesis, the four tones in a 4-syllable Accentual Phrase are referred to as L1, H1, L2, and H2. Following (Jun, 2000), each tone is considered to be phonologically associated with each of the four successive syllables.

In this chapter, the timing and scaling of each tone in the four-syllable AP are examined. For each tone, we test the conflicting hypotheses: segmental anchoring and constant duration. As expected, we observe tendencies to constant duration as well as segmental anchoring. Furthermore, which F0 features (alignment, duration, scaling) are more important is different in each tone. For example, for the initial H peak, there was more variability

[1]Transcription following the Yale Korean romanization (Martin, 1992)

of alignment than pitch level; for the second L, it was the scaling that changes more than the alignment. This chapter presents the experimental results of how the LHLH intonational pattern is phonetically realized depending speech rate, which reflects effects of segmental anchoring and constraint shape/duration.

2.2 Experiment

2.2.1 Hypotheses

The primary purpose of examining tones in Seoul Korean is to observe the effects of constant duration more easily, if they are present. Throughout this chapter, the two conflicting hypotheses, the SAH and the CDH, are tested with respect to each tone of the Seoul Korean AP. In this section, we review the predictions of these two hypotheses with the timing of the initial H peak (H1) in the Seoul Korean AP.

According to the SAH, it is expected that H peaks will be found aligned with regard to a specific segmental anchor point regardless of speech rate. Figure 2-3 illustrates what would be expected under the SAH for the timing of the first H peak in Korean 4-syllable APs at fast (a), normal (b), and slow (c) speech rates. In the figure, the labels 'S1', 'S2', 'S3', and 'S4' indicate each syllable. 'H' indicates the timing of the H peak. 'A_H' indicates the timing of the anchor point for H. The second syllable in a four-syllable AP has been described as the canonical location for the initial H (Jun, 1996: 40). In the figure, we suppose that the anchor of the first H peak is located somewhere in the second syllable and H occurs right at the anchor in normal speech. According to the SAH, H alignment is stable and this is not affected by speech rate. Thus, 'H' always equals 'A' (H=A) for all speech rates. It follows that duration of the rise (the duration from L to H) will increase as speaking rate decreases. If Seoul Korean follows the SAH, it should show a linear relation between H and A_H, with a slope of 1 in a linear regression model. In addition, in the SAH literature, it is assumed that the magnitude of the rise is not affected by changes in rise duration, so it is hypothesized to be constant (Ladd et al., 1999; Ladd, 2004). As a result, the slope of a rise increases as speech rate increases, as can be seen in Figure 2-3.

On the other hand, the predictions made by the CDH are at odds with those made by the SAH. Figure 2-4 illustrates the predictions of the CDH. Under this hypothesis, the duration, slope and F0 level of the rise remain the same at various speech rates, while the alignment between the H peak and its associated syllable (S2) is not necessarily maintained. While the shape of the rise remains invariant, syllable duration changes when speech rate changes. Thus the peak will be found in different positions relative to the presumed anchor. The alignment will be achieved only under a normal speech rate. The H peak will appear after the anchor at a fast rate, as shown in Figure 2-4a (H>A). The H peak falls at the anchor at a normal rate (Figure 2-4b, H=A). The H peak will appear before the anchor at a slow rate (Figure 2-4c, H<A). It might seem unlikely that pitch movements could remain completely invariant under changes in speech rate that affect the duration of segments, but even a tendency to maintain a shape target would result in systematic deviation from the anchor as a function of speech rate. We expect a tendency that relative to the anchor, H occurs

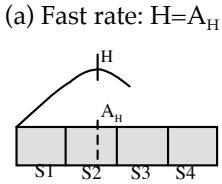


Figure 2-3: Predictions of the SAH

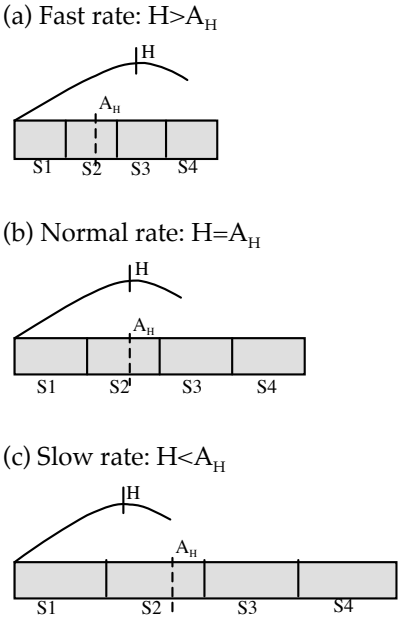


Figure 2-4: Predictions of the CSH

earlier the slower the speech rate, and later the faster the speech rate. If such a pattern is found in Seoul Korean, that will support the CDH.

Thus, by manipulating the duration of segments, we can test which hypothesis is correct in predicting tonal timing of Seoul Korean. The duration of segments can be manipulated by varying speech rate or inherent segment duration. In our experiment, inherent segment duration will be balanced out in our speech materials, and speech rate will be manipulated. Furthermore, the timing and scaling of each tone will be examined aiming at a complete description of the interaction of the timing and scaling of tones in the Accentual Phrase.

2.2.2 Experimental methods

The speech materials consisted of 4-syllable AP's, embedded in carrier phrases. An AP consisted of a three-syllable content word followed by a particle. There were 36 target phrases. The content words were collected so that height of vowels and different manner of articulation of consonants (obstruent or sonorant) were roughly balanced in all syllable positions. The inherent vowel duration is affected by vowel height, that is, high vowels tend to be shorter than low vowels. In order to control the segment duration, in each syllable position in the target words, the number of different height of vowels (high, mid, low) was balanced. There were two particles that followed the content words: [nɨn] (a topic marker 'as for'; 31 phrases) and [man] ('only', 5 phrases). The particle [man] was used when the last syllable of the preceding word ended with a coda, because in that context [nɨn] cannot surface, because its allomorph [ɨn] surfaces instead, and the coda of the last syllable is re-syllabified as the onset of the topic marker. On the other hand, [man] can attach to closed syllables, so the coda of the third syllable stays as the coda. To have both closed and open third syllables, we used [man] in five target phrases. There were 25 fillers with

different lengths (3, 5, or 6 syllables in a phrase) to break up the monotony. The order of the sentences was randomized and adjusted so that no more than two 4-syllable AP's came in succession.

The subjects were ten native Seoul Korean speakers, five females (A1, A2, A3, A4, A5) and five males (B1, B2, B3, B4, B5). Eight of them were college or graduate students in their 20's and 30's. One male (B1) and one female (A3) were in their 40's. Speech materials were presented on a sheet of paper in Korean, and the speakers were first asked to read the sentences twice naturally without any instructions on speech rate in order to elicit a natural normal speech rate. After that, they were asked to read the same sentences at a fast speech rate twice, and then at a slow speech rate twice. However, there were large speaker variations in eliciting slow speech utterances. Some slow speech rates are closer to normal speech, but some speakers spoke very slowly, so they may not necessarily sound natural.

Recording took place in a sound-attenuated recording booth in the phonetics lab in the MIT Linguistics department. Speakers wore a head-mounted microphone (Shure SM10A), which was connected to a USBPre sound input device for digitization. The digitized speech signals were directly recorded using Amadeus II (version 3.8.4, by HairerSoft) installed on an iMac, at a sampling rate of 44.100k Hz with 16 bits per sample.

2.2.3 Measurement: locating F0 events

L and H are the representations of level tones widely used in autosegmental and intonational phonology. However, very often, accurately locating L and H tones in the actual speech signal is not clear-cut. A first approximation of L and H tones is to take the local F0 minimum (L) or F0 maximum (H). The local minima and maxima have been commonly used as the locations of L and H tones for rising pitch movement in many studies, e.g. Greek (Arvaniti et al., 1998), English (Ladd et al., 1999), Dutch (Ladd et al., 2000), Japanese (Pierrehumbert and Beckman, 1988; Ishihara, 2006).

Although in many languages the F0 minima and maxima were reliably used as the indicator to the L and H tones, studies have found that locating the L tone with an F0 minimum is not appropriate in some languages. In particular, when there is a long plateau before a rising movement, the F0 minimum is often far from the actual rise, so the minimum measured on a plateau does not seem to be a reliable location of the L tone comprising the rising movement. For example, in the Mandarin Rising tone, the portion between an F0 minimum and a fastest acceleration point is often just flat (Xu, 1998). Moreover, this portion varies considerably in length and slope, from a complete plateau to a shallow rise. Thus, instead of taking the F0 minimum as the L tone of the Rising tone, Xu used the maximum acceleration point (the maximum of the second derivative of the F0 curves) as the location of the L tone.

For this reason, F0 maxima and F0 minima alone are not sufficient to accurately describe the shape of pitch rises or falls across languages, especially when the shape of rises can be decomposed into several parts. It has been recognized that a rising pitch movement consists of three components: acceleration, fast glide, and deceleration, rather than a steady linear increase from the F0 minimum to the F0 maximum (Sundberg, 1979; Xu and Sun, 2001). In this dissertation, we identify five pitch events as illustrated in Figure 2-5. The five

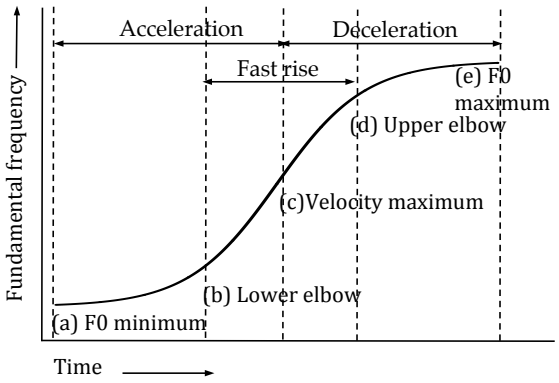


Figure 2-5: Schematic illustration of a sigmoid rise and its measurement points

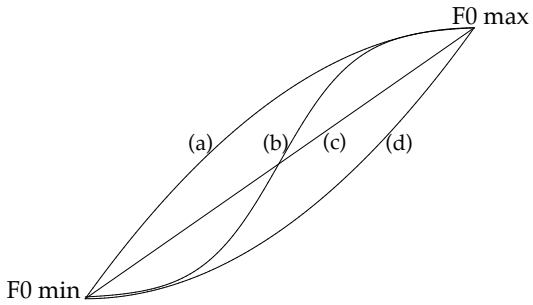


Figure 2-6: Schematic illustrations of shape of rises: (a) dome (b) sigmoid (c) straight line (d) scoop (adapted from Barnes et al. (2010))

pitch events are (a) F0 minimum, (b) lower elbow, (c) maximum velocity, (d) upper elbow, and (e) F0 maximum. The maximum acceleration point (the lower elbow (b)) corresponds to what is referred to as an "elbow" in previous studies (Beckman and Welby, 2006; Welby and Lœvenbruck, 2006; Barnes et al., 2008). In this dissertation, the elbows indicate the points where pitch contours are inflected, so we will locate two elbows (lower and upper) in a sigmoid rise, instead of one. The elbows will also be referred to as "inflection points" in this dissertation. For a sigmoid rise, there is a *fast rise* between the lower and the upper elbows, so the maximum velocity point is found between the elbows. Most of the pitch change is executed between the two elbows in a sigmoid shape.

Besides sigmoid, there are more than one possible trajectory from a minimum to a maximum. Barnes et al. (2010) describe three different shapes of rises: domed, linear, and scooped rises. Adding a sigmoid shape to these, the possible trajectories from L and H are dome, sigmoid, straight line, and scoop, as shown in Figure 2-6.

Figure 2-7 schematically illustrates the rising contours of a dome and a scoop shape with the five F0 events labeled. In the scooped rise, the fast rise occurs later in the contour, so the two elbows precede the maximum velocity point. In the domed rise, the fast rise occurs early in the contour, so the maximum velocity point precedes the two elbows.

Locating inflection points is indispensable for a complete description of rise shapes across languages, e.g. in Mandarin, omitting the information about the fastest acceleration

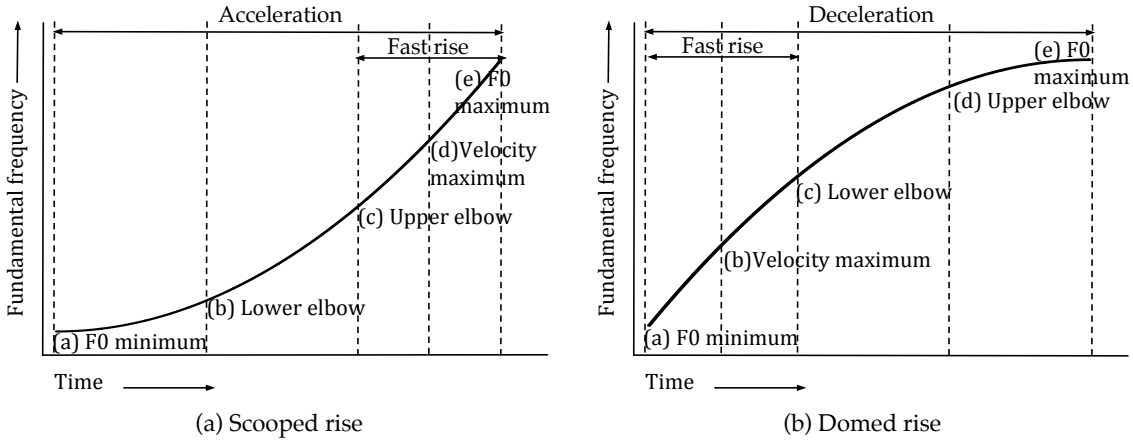


Figure 2-7: (a) Scooped rise, (b) Domed rise

points will result in an incomplete description of the rise shape. We sought a measurement technique that is applicable across languages, because we aimed at a cross-linguistically applicable analysis in this dissertation. In addition, we did not know in advance which are the better indicators of L and H tones, elbows or F0 extrema, in a given language. Thus, we aimed at locating all five points in pitch rises for every language, so we sought a technique that locates elbow points in any given F0 rise.

Thus, in our analysis, we located the five points for each rising movement: F0 minimum, lower elbow, maximum velocity point, upper elbow, F0 maximum. F0 extrema were located manually. Elbows and maximum velocity points were located using automatic methods that will be described shortly. After these points were identified and labeled in all speakers and languages, pitch and time values of each point were collected using a Praat script (modified from Lennes 2003). The collected data were rearranged using Python scripts and fed into R for statistical analyses and modeling.

F0 maxima and minima

F0 maxima and F0 minima were identified manually using Praat. A region where it seems an extremum occurred was manually selected, and the precise location of the extremum was found by a Praat command.

Inflection points

A number of automatic elbow-finding algorithms have been suggested because of the inability of humans to locate pitch elbows with accuracy and consistency (Giudice et al. 2007; Welby and Lœvenbruck, 2005: 21; Xu, 1998: 196-197). Xu (1998) used the maximum acceleration point (the maximum of the second derivative of the F0 curves) to locate the inflection point. Li (2003: 45) found that the beginning of the rise obtained in this way is impressionistically too late in many cases in various dialects of Chinese (Mandarin, Zhenhai, Shanghai, etc). Li thus took the mid point between the F0 minima and the moment of greatest ac-

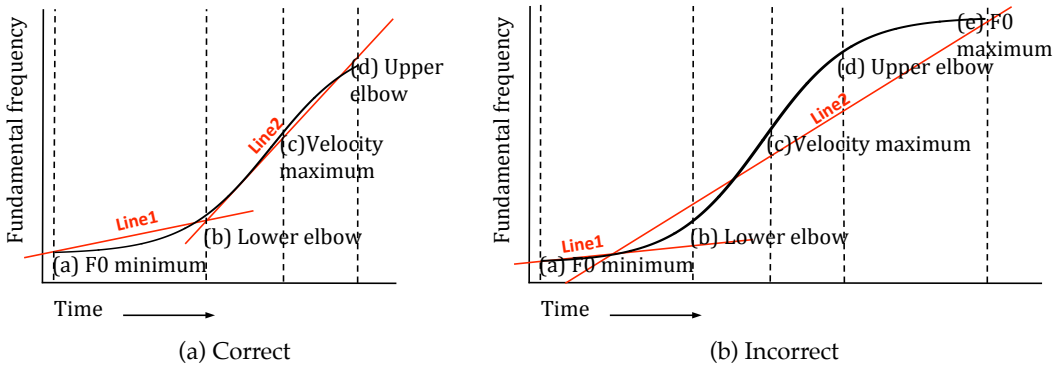


Figure 2-8: The problem of two-piece linear regression: (a) Correct, (b) Including the upper shallower region: the lower elbow is located too early.

celeration as the L tone. del Giudice et al. (2007) compared several methods with human intuition (six human labelers), and found that the least-squares fitting algorithm (Beckman and Welby, 2006) is the closest to human intuition.

The least-squares fitting algorithm searches for the best fitting lines that minimize the summed squares of the distance between the lines and the data points. This is essentially the same procedure as a usual linear regression, except that the elbow-finding algorithm in Beckman and Welby (2006) is a break-point regression, i.e. two lines are searched instead of one line. The intersection point of the two linear regression lines is the elbow. Figure 2-8a shows a schematic illustration of the two-piece break-point regression, with the analysis window from the F0 minimum to the upper elbow.

However, a problem with this algorithm is that the resulting location of the elbow is easily affected by the selection of the analysis window. The algorithm by Beckman and Welby (2006) was designed to locate the elbow point that corresponds to what we call the lower elbow, by fitting only two regression lines. To find a lower elbow, a window of analysis has to be selected first. A window can be selected from an F0 minimum to somewhere midway through a rise or up to the point just before the upper elbow, so that the shape of the contour in the window is appropriate to be fitted with two lines. An example is shown in Figure 2-8a. In this example, the window is selected from the F0 minimum to the point that we defined as the upper elbow. However, if the analysis window includes any shallow or plateau part in the upper region, as in Figure 2-8b, two lines are not enough, because the location of the (lower) elbow is affected by the plateau area near the F0 peak. As a result, the algorithm finds a point somewhere earlier than where it seems the correct elbow should be.

The two-piece linear regression is very sensitive to where one selects by eye as an analysis window, so it may not be very reliable. To locate a lower elbow precisely and consistently, knowing the precise location of the upper elbows is necessary; but without knowing the precise location of the lower elbow, the upper elbow cannot be located, for the same reason. Either way, the two-piece regression is subject to inconsistency and imprecision for a complete description of a rising movement.

Thus, we propose a three-piece linear regression, instead of two-piece. Figure 2-9 shows

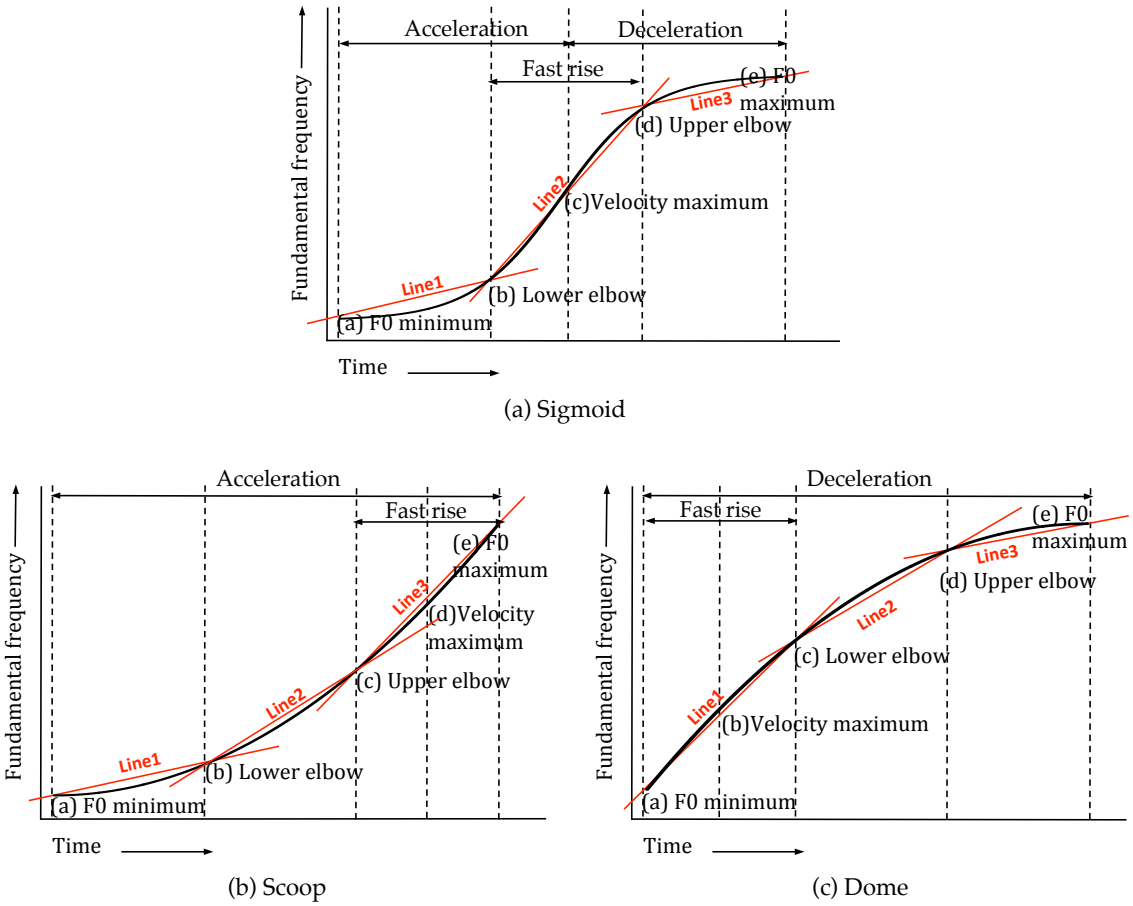


Figure 2-9: (a) Sigmoid, (b) Scoop, (c) Dome: Fitting three types of F0 rising contours with three-piece linear regression lines.

the schematic illustration of the proposed method, which is an extension from the original algorithm by Beckman and Welby (2006). For each shape of rise, the analysis window is the portion from the F0 minimum to the maximum. The algorithm finds the three regression lines that best fit the given F0 data points. The intersection point of the first and second lines is the lower elbow; the intersection point of the second and third lines is the upper elbow.

There are several advantages of using this three-piece linear regression. The analysis window does not have to be manually selected, so we can consistently use the portion from the F0 minimum to the maximum as the fitting window. Thus, the result is less affected by subjective human decisions. Secondly, the three-piece regression locates the lower and upper elbows at the same time.

Furthermore, using the slopes of the three regression lines, we can define and automatically classify shapes of rises, quantitatively rather than impressionistically. As can be seen in Figure 2-9a, for a sigmoid rise, the second line is the fastest and the first and third lines are shallower than the second line. For a scoop rise in Figure 2-9b, the third line is the fastest, and the slopes increase towards the F0 maximum. For a dome rise in Figure 2-9c,

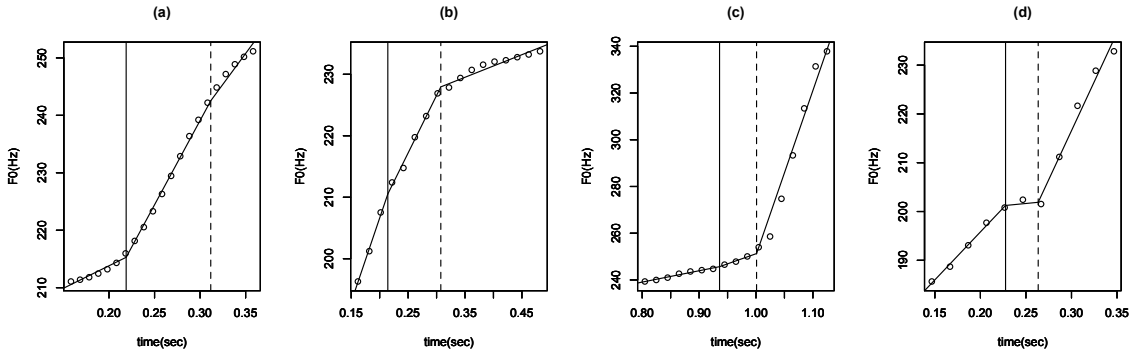


Figure 2-10: Examples of locating inflection points by three-piece linear regression. The circled points are raw F0 data. (a) Sigmoid (Korean speaker A1), (b) dome (Korean speaker A3), (c) scoop (Chinese speaker C2), (d) other (Korean speaker A2)

the first line is the fastest, and the slopes decrease towards the F0 maximum. These relations are summarized in (1), where 'slope1' is the slope of the first regression lines ('Line1'), 'slope2' is the slope of the second regression line ('Line2'), and 'slope3' is the slope of the third regression line ('Line3').

(1) a. Sigmoid: slope1, slope3 < slope2
b. Scoop: slope1< slope2 < slope3
c. Dome: slope1 > slope2 > slope3
d. Linear: slope1 = slope2 = slope3

Another possibility is that the first and third slopes are steeper than the second slope, but in most cases, such patterns reflect a dip due to local segmental perturbation, so they are classified as 'other'. Figure 2-10 shows actual examples of the three-piece regression for each shape. In each panel, the beginning and end of the F0 curves correspond to the absolute F0 minima and maxima. The F0 curves are divided into three parts by three linear regression lines that best fit the F0 data points. The solid vertical line is the location of the first elbow; the dashed vertical line is the location of the second elbow. In practice, linear shape is impossible because that means three lines have exactly the same slope. Thus, we will have only three kinds of shape (dome, sigmoid, scoop), and 'other' which does not belong to any of these categories.

Velocity maxima

In addition to the inflection points between F0 minima and maxima, the locations and values of the velocity maxima are obtained. To locate the maximum velocity point, we fit the F0 contours with cubic smoothing splines (Penalized Smoothing Splines; Ripley (2009)). A spline is a function defined piecewise by polynomials. That is, the raw F0 data is fitted with the smoothing function with a piecewise polynomial of degree three. The velocity maximum can be found by taking the first derivatives of the smoothed curves from spline smoothing. Preliminary examinations of our data revealed that smoothing splines are bet-

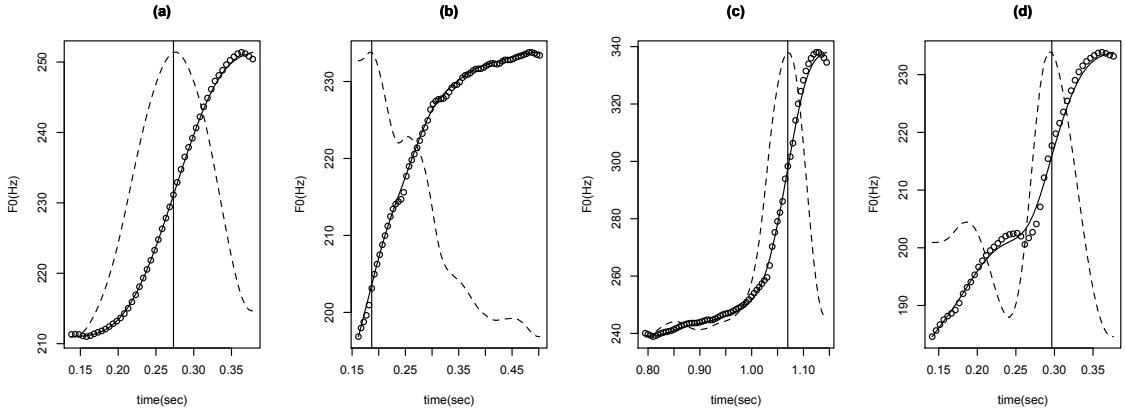


Figure 2-11: Examples of locating maximum velocity points using spline smoothing. The tokens are the same as those in Figure 2-10. The circles are raw F0 points. The curves in solid lines over the raw points are the fitted splines. The dashed curves are the first derivatives of the fitted splines. The solid vertical lines are the locations of velocity maxima (=the maximum of the function in the dashed line).

ter than a global polynomial (cubic) fitting, because splines are piecewise. A cubic function can globally fit F0 contours that are close to a sigmoid shape. However, if the shape is different from a normal sigmoid, then global fitting with a cubic function would be poor. For example, for scoop or dome shapes (Figure 2-7a, 2-7b), global fitting results in a quadratic function, rather than cubic, because scoop or dome shapes are closer to a quadratic function than a cubic function. If the smoothed curve is a quadratic function, its derivative is a straight line, so the derivative maximum is not available. On the other hand, spline fitting is versatile, because it fits the curve piecewise instead of globally fitting the F0 curve in the selected window. The velocity maximum is located by finding the location of the maximum of the derivative of the spline smoothing function. The maximum velocity points obtained by this procedure were examined by the author, and most of them seemed to correspond to the author's judgments.

One problem of splines fitting is that it may find velocity maxima at local blips caused by segmental perturbation. This source of errors can be minimized by appropriately adjusting parameter values, although it cannot be totally avoided. The following parameter settings seemed most appropriate for our data when using the "smooth.Pspline" function in the pspline package (Ripley, 2009)): method=1 (the method for controlling the amount of smoothing: 1 uses the value supplied for 'spar'), spar=0.00001 (the coefficient of the integrated squared derivative of order 'norder'), norder=2 (the order of spline=cubic). The actual examples of finding maximum velocity locations by splines are presented in Figure 2-11, which show that splines locate the maximum velocity point regardless of shape.

Segment boundaries

Segment boundaries were all manually labeled in Praat (Boersma and Weenink, 2009), using the labels as shown in Figure 2-12. Time and F0 values at each segmental boundary

were collected using a Praat script (modified from Lennes (2003)).

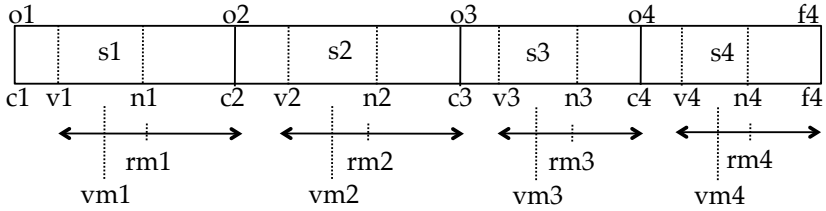


Figure 2-12: Segment labels: Each label indicates the beginning of segments. The numbers indicate the syllable (s1: the first syllable). 'c' is an onset consonant', 'v' is a vowel, 'n' indicates a coda, 'o' indicates the onset of a syllable, 'rm' is the middle of a rime, 'vm' is the middle of a vowel. For example, c1: the beginning of the onset consonant in the first syllable, v1: the beginning of the vowel in the first syllable, n1: the beginning of the coda in the first syllable. 'o1' the beginning of the first syllable, 'vm1' the middle of the first vowel, 'rm1' the middle of the first rime. 'f4' is the end (*final*) of the fourth syllable. 'o', 'f', 'rm', and 'vm' points were not directly labeled, but computed from the segmental boundary information.

Summary of measurements

In summary, five F0 events were located with automatic or manual methods: F0 minima, F0 maxima, lower and upper elbows, maximum velocity points. F0 minima and maxima were located manually. Lower and upper elbows were located by fitting three-piece linear regression. Maximum velocity points were located by using the derivative of fitted smoothing splines functions. These procedures were conducted for all speakers and languages studied in this dissertation. The example of the final outcomes are shown in Figure 2-13. The upper tier is the segmental tier, and the lower tier is the tonal tier. For all the labeled points, the timing and F0 values were collected using Praat scripts adapted from Lennes (2003). In the Korean analysis in this chapter, F0 minima and F0 maxima were taken as L and H tones.

2.2.4 Statistical method: linear mixed-effects models

A mixed-effects model is a model with both fixed effects and random effects. *Fixed effects* are parameters that are associated with an entire population, and *random effects* are associated with individual experimental units drawn at random from a population (Pinheiro and Bates, 2000; Baayen, 2008; Baayen et al., 2008). For example, if according to a hypothesized model, the timing of H peak is predicted by the timing of the end of the vowel, the dependent variable is the timing of H peak, and the predictor variable (the timing of the end of the vowel) is the fixed effect. However, depending on speakers, the H peak may occur earlier than the end of the vowel, or later than the vowel. A mixed-effects model finds the grand mean, allowing for by-speaker adjustments. That is, the timing of the H peak is predicted by the fixed effect (the timing of the end of the vowel) and random effects of speaker; to estimate speaker-specific adjustments to the grand mean, a speaker-specific

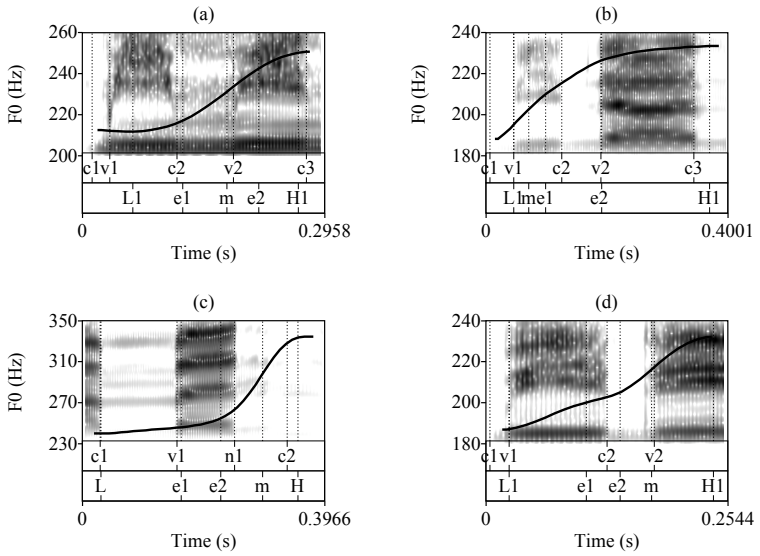


Figure 2-13: Measurement examples for each shape: (a) sigmoid (b) dome (c) scoop (d) other. These are the same tokens as those used in Figure 2-10 and Figure 2-11. The segments are labeled in the upper tier: 'c1' the beginning of the first consonant, 'v1' the beginning of the first vowel, 'n1' the beginning of the coda in the first syllable, 'c2' the beginning of the second consonant, 'v2' the beginning of the second vowel. In the lower tier: 'L' F0 minimum, 'H' F0 maximum, 'e1' the lower elbow, 'e2' the upper elbow, 'm' maximum velocity point.

intercept ("by-speaker random intercept") can be added to the grand mean. In addition, the timing of the end of the vowel may be significantly different across speakers. In that case, speaker-specific adjustment of the timing of the end of the vowel ("by-speaker random slopes for the timing of the vowel") can be added to the coefficient for the timing of the end of the vowel to find the predicted value of the H peak timing for a particular speaker.

By likelihood ratio tests, we can test whether adding a parameter to a model significantly contributes to the goodness of fit, if the compared models are nested. For example, to test whether adding by-speaker random slopes for the timing of the vowel can be justified, the model with and without the parameter can be tested by a likelihood ratio test (Baayen et al., 2008: 395). The likelihood ratio test statistic is $2log(L_g/L_s)$, where the likelihood of the more general model (with the additional parameter) is L_g, the likelihood of the more specific model (without the parameter) is L_s. The likelihood ratio test statistic follows a chi-squared distribution with $(g - s)$ degrees of freedom, where g is the number of the parameters in the more general model, and s is the number of the parameters in the more specific model. The probability (p) of obtaining the chi-squared statistic value shows the probability that the null hypothesis that the more specific model (without the parameter) is sufficient is true. If the p value is high (> 0.05), it means that adding the factor (for our example, by-speaker random slopes for the end of the vowel) is not justified, and the model without the parameter is a better model.

When comparing models, maximum likelihood estimation (ML) is used for model fit-

ting. ML aims to find the parameter values that make the model's predicted values most similar to the observed values. Once the best-fitting model is determined by likelihood ratio tests, the coefficient values of the model are re-estimated using restricted maximum likelihood estimation (REML), which is more precise for mixed-effects models (Baayen et al., 2008: 394). In this dissertation, ML is used for model comparison and REML is used for coefficient estimation.

If the compared models are not nested, the deviance values of the two models can be used as the measure of goodness of fit. The deviance of a model is essentially a log-likelihood ratio test (LRT) between the model of interest and a 'saturated' model with a parameter for every observation so that the data are fit exactly. In this sense, deviance is similar to residual variance, the lower the deviance the better the model, if the compared models have the same number of parameters. For example, we compare deviance values when we test which segmental landmark is the best predictor of the H peak. We can compare whether the timing of H peak is better predicted by the timing of the end of the vowel or the middle of the rime. If the model predicting the timing of H peak by the timing of the middle of the rime has the lower deviance than the model with the end of the vowel, we consider the middle of the rime as the better anchoring point for H.

Mixed-effects models are appropriate when pooling across speakers, because it allows by-speaker adjustments to the grand mean and by-speaker adjustments to the coefficients of each fixed-effect. In a linear multiple regression which is not a mixed-effects model, speakers and other factors are treated as the same, i.e. as fixed effects. For our example, both the timing of the end of the vowel and each speaker's mean would be treated equally as predictor variables, without distinguishing the parameters that apply to the entire population and individual speakers randomly drawn from the population. Thus, we use mixed-effects models in most cases in this dissertation.

2.3 Overall Shape

In this section, we describe the overall shape of the Seoul Korean LHLH Accentual Phrase. Figure 2-14 is a schematic illustration of the LHLH intonation pattern realized on a four-syllable Accentual Phrase. We refer to the F0 turning points as L1, H1, L2, and H2. Following the previous literature (Jun, 2000; Lee and Kim, 1997), it is supposed that the tones are associated with each syllable (s1, s2, s3, s4). We also refer to the F0 movements between these points as 'Rise1' (from L1 to H1), 'Fall' (from H1 to L2), and 'Rise2' (from L2 to H2). In a normal speech rate, Rise1 spans s1 and s2, Fall spans s2 and s3, and Rise2 spans s3 and s4. We calculate 'local speech rate' for each movement based on the syllables they span. That is, for Rise1, the local speech rate is the inverse of the duration of the first two syllables: $1/(s1 + s2)$. For Fall, it is $1/(s2 + s3)$, and for Rise2, $1/(s3 + s4)$.

In this dissertation, the fast, normal, slow speech conditions will be referred to as the 'categorical speech rate', as opposed to 'local speech rate' (the inverse of the syllable duration).

In Figure 2-15, we illustrate some typical examples of the same AP produced at fast, normal, and slow speech rates by speaker A1. In normal speech (b), the LHLH tune is

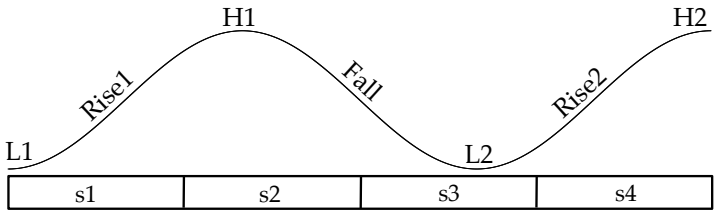


Figure 2-14: Schematic illustration of the Seoul Korean 4-syllable LHLH Accentual Phrase

produced with four syllables (*malummonun* [mari̟mmoni̟n] "A lozenge is" (lozenge+topic marker)). In fast speech (a), only one rise is produced. When the segmental duration of AP is extremely short at the fast rate, the tones in the LHLH tunes are not fully realized, and only one rise is produced over the four syllables. On the other hand, in slow speech (c), there is a plateau throughout the first syllable, and the actual rise starts from the end of the first syllable.

Figure 2-16 illustrates the averaged pitch contours by speech rate for each speaker. We can see that in most speakers (A1, A2, A4, A5, B2, B4, B5), the shape of the first rise and fall in normal speech roughly corresponds to the shape of the only rise and fall in the fast speech. In the fast speech condition, not all four tones are fully realized; instead there is only one rise and fall. In some speakers, this rise and fall in fast speech has a smooth sigmoid shape which is fairly similar to the one in normal speech (speakers A1, A4, B5). Speaker B2 also has one rise in fast speech, but it is very different from the normal speech rise: it is a sharp rise and fall, and the pitch range is much higher. In other speakers, there is a plateau between the first inflection point to the start of the fall (A2, A5, B2). No speaker fully produced the two rises in the fast speech. H2 is particularly high in normal speech of speakers A4 and B4; they also often put a short pause after the high H2, in which case the AP can be considered as an IP (Jun, 1996, 2000).

Compression and truncation

The question arises as to how we view the tunes with only one rise in fast speech. We suggest that some of them are due compression of four tones, and others are due to truncation of one of the two rises. When there is tonal or temporal pressure, intonation tunes undergo *compression* or *truncation* (Ladd, 2008: 182-184). In English, the tonal sequence L*+H_L_H% (rise-fall-rise tune) can be spread over a long phrase, or the whole contour can be realized with a monosyllabic word (e.g. "Sue?"), which involves compression. On the other hand, Hungarian has a restriction that only two tones can occur in one syllable. Thus, when L*_H_L% is realized on one syllable, L% is not realized, so it becomes L*_H. This is truncation of tones. (Ladd, 2008: 182-183) suggests that the difference between compression and truncation is that compression involves adjustment of phonetic values of existing tones, e.g. undershoot of L between H peaks is a common process (e.g. Mexican Spanish (Prieto, 1998), English (Pierrehumbert, 1980), Japanese (Kubozono, 1991, 1993)), whereas truncation involves a change from one type of intonational pattern to another tone sequence which can be more easily realized (Catalan (Prieto, 2005)).

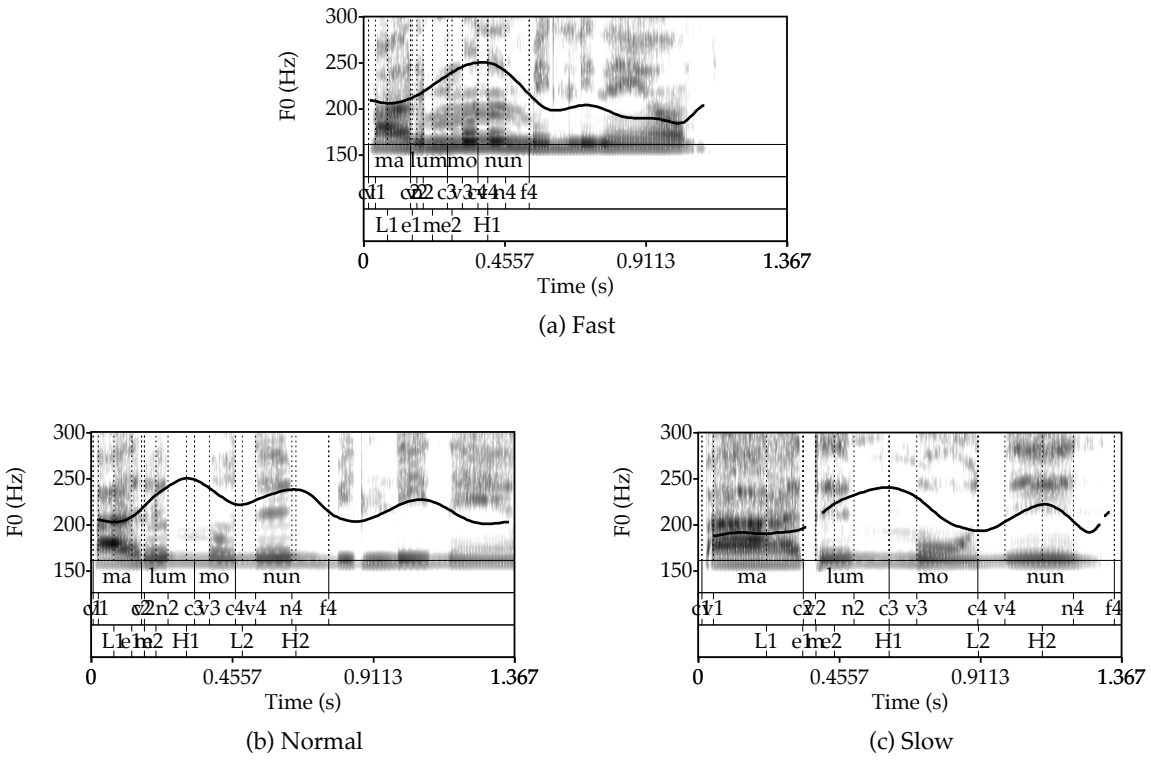
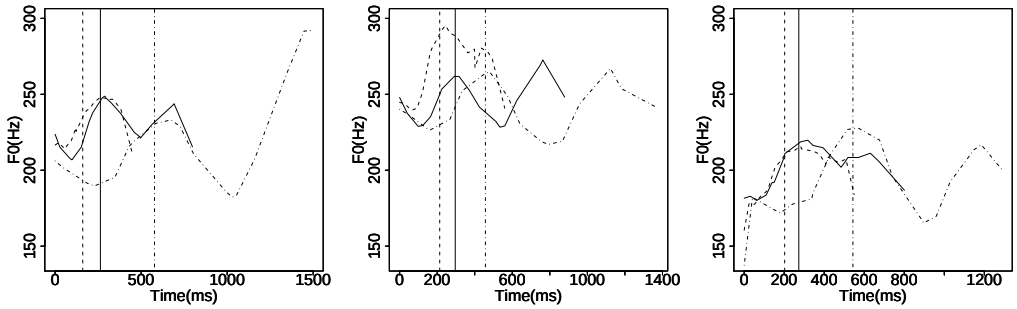
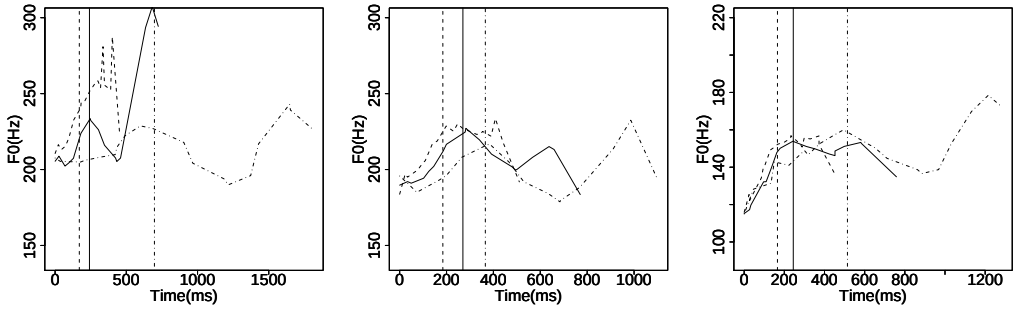


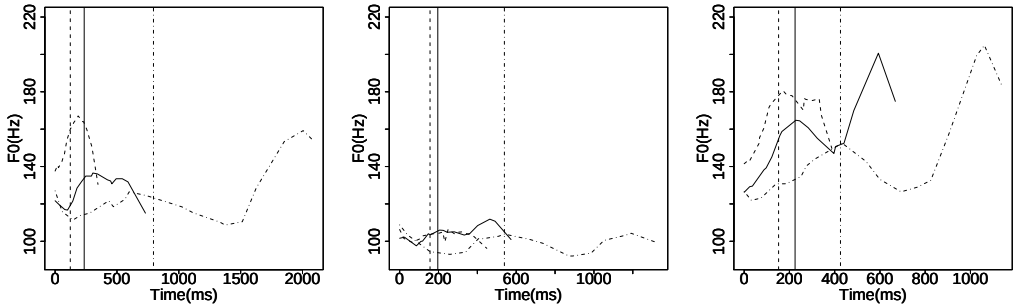
Figure 2-15: Examples of Korean AP *malummonun*[mariːmmoniːn] "A lozenge is" (lozenge+topic marker) produced at different speech rates by the same speaker (A1). (a) Fast rate, (b) Normal rate, (c) Slow rate. 'f4' is the end of the fourth syllable (the end of the AP). The time and F0 scales are the same across all.



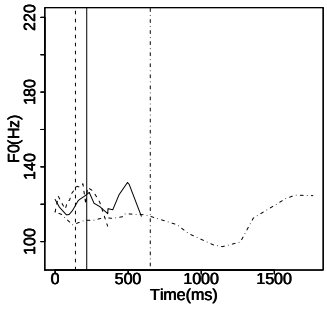
(a) A1 (b) A2 (c) A3



(d) A4 (e) A5 (f) B1



(g) B2 (h) B3 (i) B4



(j) B5

Figure 2-16: Averaged contours of the Korean LHLH intonation by speaker and by speech rate (Solid: normal, dashed: fast, dot-dashed: slow). The vertical lines are the location of the middle of the second syllable.

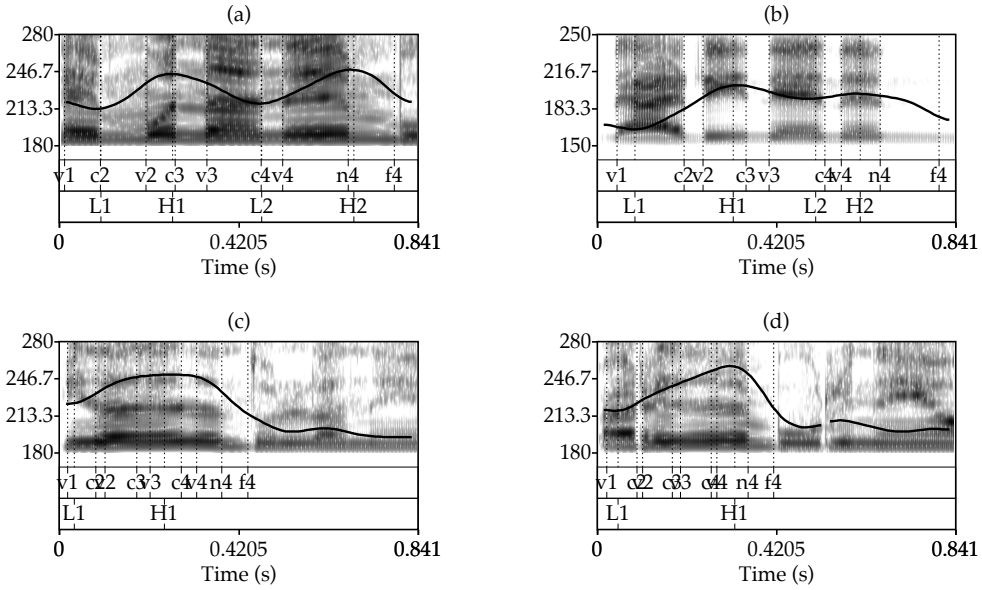


Figure 2-17: Types of phonetic realization of the LHLH AP: (a) Full scaling (Speaker A1, normal rate), (b) Undershot of L2 (Speaker A3, normal rate), (c) Compression (Speaker A1, fast rate), (d) Truncation (Speaker A1, fast rate). The label 'f4' is the end of the AP.

We suggest that the rise with a plateau shape in fast speech in Seoul Korean is due to compression, and the sharp rise in the same condition reflects truncation of one of the two rises. Plateau shapes are observed in some speakers' fast speech (e.g. A2, A5, B2). A possible interpretation of these plateaus is that the L2 between H1 and H2 is not realized (extreme undershoot) because of time pressure, and the F0 inflection point just before the fall may be the reflex of H2. This can be further supported by the fact that in some speakers, a similar plateau-like shape is observed even in normal speech (e.g. A3, B2), which suggests that realization of L2 is not strictly required in general, so L2 can be undershot even when there is no time pressure. Thus, it is most likely to be skipped under time pressure. On the other hand, a sharp rise in fast speech (A1, A4, B2, B5) indicates truncation of one of the rises. That is, the underlying LHLH tune is changed to LH when there is time pressure, so in the surface contour, there is no trace of the second LH. This can be supported by the existence of utterances that have a smooth rise without a plateau in fast speech (A1, A4, B2, B5). In particular, in speaker B2, it is hard to justify the claim that the speaker still plans to produce four tones when speaking fast.

Figure 2-17 shows the varying realization of the LHLH Accentual Phrase: (a) Full scaling, (b) undershot, (c) plateau, and (d) truncation. In many of the slow and normal speech rate utterances, full scaling patterns (a) are commonly observed. Undershot patterns (b) are more likely to be found in normal speech than slow speech. There are also speaker variations in undershooting behaviors. Some speakers are more monotonous than others, so more likely to undershoot L2. For example, in the previous Figure 2-16, the averaged contours of normal speech of speakers A3 and B2 have higher L2 than speakers A1 and A2's normal speech realizations. However, within the same speaker, the observation that

Table 2.1: Classification of shapes of Rise1 in Seoul Korean by speech rate

	sigmoid	dome	scoop	other	N/A
All	1304 (61%)	272 (12%)	154 (7%)	343 (15%)	80 (4%)
Fast	471 (66%)	127 (18%)	19 (3%)	80 (11%)	18 (3%)
Normal	504 (70%)	72 (10%)	57 (8%)	62 (9%)	25 (3%)
Slow	329 (46%)	73 (10%)	78 (11%)	201(28%)	37(5%)

L2 is undershot when speech rate is fast still holds, e.g. for speaker A3, L2 is fully realized in slow speech while it is considerably undershot in normal speech.

When speech gets faster, the realization of L2 can apparently be skipped, producing a plateau-like shape between H1 and H2, as shown in Figure 2-17(c). This pattern is often found in categorical fast speech and sometimes in normal speech. Furthermore, at extremely fast rates, the whole Accentual Phrase can be realized with just one rise, as shown in (d). In such cases, it no longer seems reasonable to say that underlyingly there are two rises. We suggest that at this stage the LHLH intonational pattern is truncated into one LH. In sum, the changes from full scaling to truncation largely depend on speech rate.

In Section 2.5.2, we will show that phonetic pressure (i.e. the absolute time available for the fall from H1 to L2) has a gradual effect on L undershoot. This suggests that when time pressure is below a certain threshold, the tones are compressed, but above a certain threshold, the tune itself is changed to a different tune (i.e. LHLH becomes LH).

Shape depending on speech rate

As mentioned in Section 2.2.3, we have classified the shapes of pitch rises into sigmoid, dome, and scoop based on the slopes of three regression lines fitted to the rising F0 movement, according to the criteria shown in (1). Table 2.1 shows the result of shape classification of the initial rise (Rise1) in Seoul Korean by speech rate. The 'other' category is the shape that does not belong to any of the shape categories (that is, the first and third regression lines are steeper than the second regression line). The N/A's are the data where F0 values cannot be measured due to perturbations, and were not included in later analyses. The majority of rises have a sigmoid shape, 70% of the rises in normal speech were sigmoid. Domes are found more in fast speech than in normal or slow speech. Considering that domes start with a fast rise, this may mean that when there is time pressure due to fast speech rate, a rise may start relatively earlier than when there is less time pressure. On the other hand, scoops are found mostly in slow speech, so it may mean that when there is less time pressure, the fast portion of rise starts later.

The 'other' category accounts for a high proportion in slow speech, this may reflect difficulties in measurement in some of the overly slow speech in Seoul Korean. Often, the rising movement in slow speech was elicited with plateaus in the first and second syllables, which makes measurement of F0 minima and maxima unreliable. The proportion of the 'other' category becomes much lower in other languages in Chapter 4, when we elicited slow speech closer to normal speech (more natural and less slow).

2.4 Rise1

From this section to Section 2.6, the experimental results of the timing and scaling of each part (Rise1, Fall, and Rise2) of the Korean LHLH Accentual Phrase are presented. The two hypotheses are tested for each tone (L1, H1, L2, H2), which was described in Section 2.2.1. That is, the effects of segmental anchoring and constant duration are examined for each tone. We will see that the timing of the tones showed tendencies that support both the SAH and the CSH.

2.4.1 Alignment of L1 and H1

Segmental anchoring

According to the SAH, it is expected that H1 and its anchor A_{H1} will have a high positive correlation. To estimate the best anchoring point, we tested several possible segmental landmarks in and near the second syllable, listed in (2), and illustrated in Figure 2-18.

(2) Various segmental landmarks to test for the best anchoring point for H1
   a. 'c2' the beginning of the consonant in the second syllable
   b. 'v2' the beginning of the vowel in the second syllable
   c. 'vm2' the middle of the vowel in the second syllable
   d. 'rm2' the middle of the rime in the second syllable
   e. 'c3' the beginning of the consonant in the third syllable
   f. 'v3' the beginning of the consonant in the third syllable

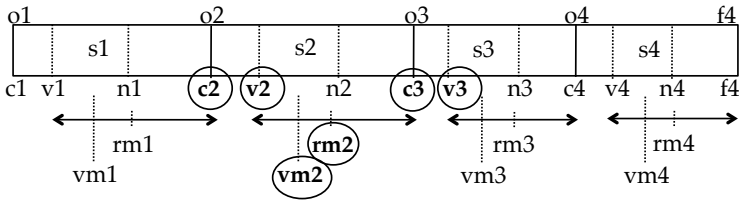
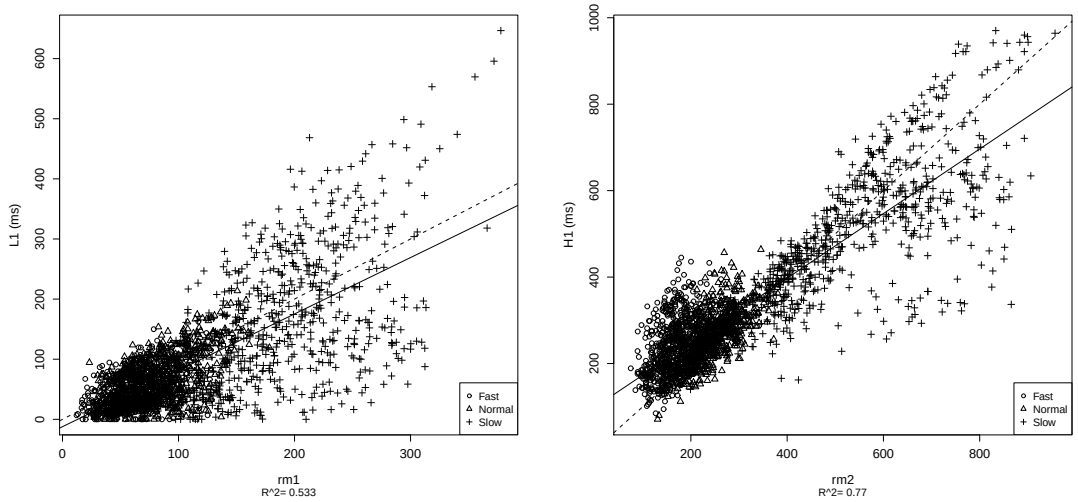


Figure 2-18: The circled positions were the candidate anchoring points for H1.

As a first approximation, a linear regression model was fitted to the data with the timing of H1 as a dependent variable, timing of a candidate segmental landmark, speaker, and their interaction as predictor variables. The highest R^2 was found with the segmental landmark 'rm2', the middle of the rime in the second syllable (R^2=0.846). In Figure 2-19b, the timing of H1 is plotted against timing of 'rm2', pooling across all Korean speakers. The solid line is fitting H1 against the anchor (R^2=0.77) (The difference in the R^2 value is because the fitted line in the figure ignores the speaker variable). The dashed line is the $y = x$ line. A positive correlation is expected given the SAH, so we can say that there is a tendency to segmental anchoring.

However, mixed-effects models are more appropriate when pooling data across speakers, for the reasons mentioned in Section 2.2.4. Thus, we tested the segmental landmarks again using mixed-effects models in order to examine whether 'rm2' is still the best anchor-



(a) Alignment of L1 (b) Alignment of H1

Figure 2-19: (a) L1 against A_{L1}, (b) H1 against A_{H1}, the dashed line is $y = x$

Table 2.2: Mixed model comparisons for the model predicting H1 from A_{H1}. The LRT is with the immediately preceding model.

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H1	none	speaker	none	
(b)	H1	rm2	speaker	none	$\chi^2(3) = 3663.4, p < 0.001$
(c)	H1	rm2	speaker	rm2	$\chi^2(3) = 355.05, p < 0.001$

ing point when speakers are treated as random effects. Several mixed-effects models are tested to find the best model specification, i.e. whether a certain fixed or random effect is necessary or not. The three models that were compared were shown in Table 2.2. In model (a), the dependent variable is the timing of H1, and there were by-speaker random intercepts and no random slopes for 'rm2'. In model (b), the timing of 'rm2' is added to model (a) as a fixed effect. In model (c), the timing of H1 was the dependent variable, the timing of the anchor ('rm2') was the fixed effect, and there were by-speaker random intercepts and by-speaker random slopes for 'rm2'.

According to the likelihood ratio tests, model (c) was significantly better than model (b) (without by-speaker random slopes for the anchor) [χ^2(2)=355.05, p<0.001], and model (a) (by-speaker random intercepts only and no fixed effects) [χ^2(3)=3663.4, p<0.001]. The result that the models with a fixed effect of 'rm2' were significantly better than the model without the fixed effect of the anchor means that the effect of a segmental landmark is significant in predicting the timing of H1, so this is a tendency to segmental anchoring. Mixed-effects models were fitted to test other segmental landmarks, and it turned out that 'rm2' remains as the best, yielding the lowest deviance (23283). The coefficient (=the slope of the linear regression) for the anchor 'rm2' was 0.772.

Although there was a tendency to segmental anchoring, if peak timing is determined by the SAH only, the regression line in Figure 2-19b should be almost parallel to the $y = x$ line, and the slope of 'rm2' in model (c) must be close to 1. A linear regression shows that the slope is significantly different from 1 [Slope=0.75, $t(2080) = 28, p < 0.001$, the slope compared with 1]. However, statistical significance tests on the regression lines in this kind of plots are not very informative, because the data are pooled across speakers in the plot, so it is not reflecting random effects of speakers, and furthermore, not meaningful for the model we develop in Chapter 3. These plots are to give us the idea of a broad pattern of the data.

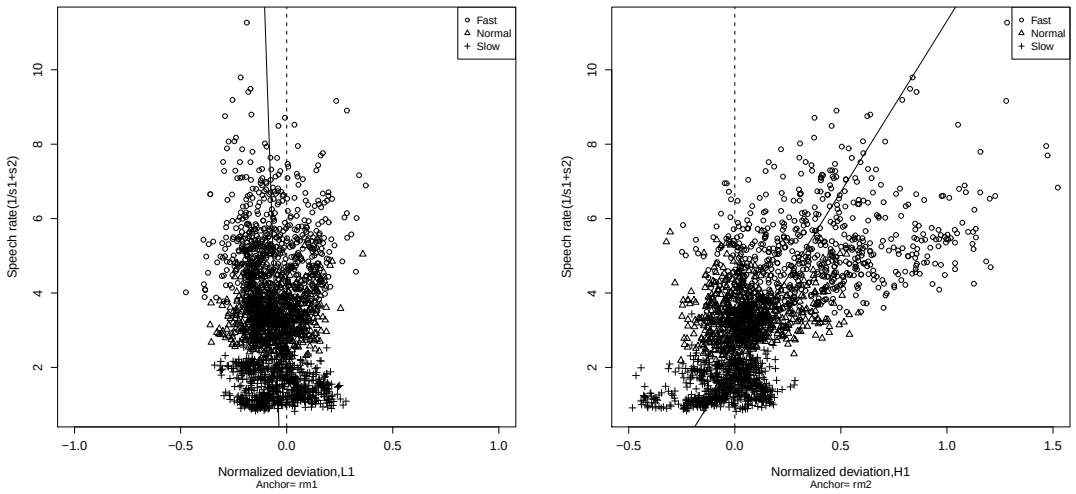
Instead of the slope of 1 which is expected according to the SAH, the plot shows a pattern that when the anchor occurs earlier, the peak occurs later than the anchor; when the anchor is later, the peak is earlier than the anchor. This can be seen in the plot. In the first condition, the solid line is above the dashed line in the lower left region of the plot while in the second condition the solid line is below the dashed line in the upper right region of the plot. This suggests that there should be another factor in addition to segmental anchoring that systematically affects the timing of the F0 peaks.

The same analysis was carried out for L and A_L. We tested several possible alignment points ('v1','vm1','rm1','o2','c2','v2') by mixed-effects modeling. For each landmark, mixed-effects models were fitted with timing of L1 as the dependent variable, timing of a segmental landmark as a fixed effect, with by-speaker random intercepts and slopes for the timing of the anchor. The model with the lowest deviance was found with 'rm1', the middle of the rime in the first syllable (21928). In Figure 2-19a, timing of L1 is plotted against timing of this best anchoring point 'rm1', pooling across all Korean speakers. The solid line is fitting L1 against the anchor and the dashed line is the $y = x$ line. The slope is close to 1 (almost parallel to $y = x$), which suggests a tendency to segmental anchoring, stronger than in the case of H1. We trimmed down the model to test the significance of the parameters. Removal of by-speaker random slopes for anchor was significant [χ^2(2)=591.78, p<0.001], and removal of by-speaker random slopes for anchor and the fixed effect of anchor was also significant [χ^2(3)=2312.8, p<0.001]. Thus, the best-fitting model for the L1 is the model with the fixed effect of anchor, by-speaker random intercepts and by-speaker random slopes for anchor.

The anchor estimates in this section are the best anchors by the criteria of the SAH. More precise estimation of the anchor location will be carried out based on the model proposed in Chapter 3. For now, the current estimation methods will be used in the following sections.

Evidence for target duration

We further examined the locations of L1 and H1 relative to their anchors. We found that there is a systematic variation in the timing of L1 and H1, relative to the segmental anchor depending on speech rate. Figure 2-20 illustrates the results. In the plot, the x-axis is 'normalized deviation', and the y-axis is local speech rate. H1 deviation means the difference between the anchor and the H1 peak ($H1 - A_{H1}$). This is normalized by local speech rate, i.e. normalized deviation $= (H1 - A_{H1})/(s1 + s2)$, where $(s1 + s2)$ is the duration of the first two syllables; its inverse is the measure of *local speech rate*. The shorter the duration of the



(a) L1 deviation from A_{L1}

(b) H1 deviation from A_{H1}

Figure 2-20: (a): L1 deviation from A_{L1}, (b): H1 deviation from A_{H1}, the dashed line is the position of the anchor

first two syllables, the greater the local speech rate. In the figure, the normalized deviation of L1 and H1 is plotted against local speech rate ($1/s1 + s2$). The dashed line is the position of the anchor. We used the best anchor estimate from the previous section ('rm2' for H1, 'rm1' for L1).

In Figure 2-20b, we can see that H1 peaks occur later relative to the anchor at a faster speech rate and earlier at a slower speech rate. (Note that the predictor variable (local speech rate) and the dependent variable (normalized deviation) are switched in the plot, in order to better visualize relative deviations from the anchor). This is a tendency reflecting the prediction of the CSH. As for L, we observe that the direction of the CSH effect is reversed, as shown in Figure 2-20a. That is, L occurs earlier as speech rate gets faster and later as rate gets slower.

Combining the results of Figure 2-20a and 2-20b, when speaking faster, L tends to occur earlier, and H tends to occur later relative to their respective anchors. That is, the rise starts earlier and terminates later than the anchor to accommodate the target duration when speaking fast. We interpret this result to mean that there is a preferred duration for a rise, or *target duration*. The relation between a target duration and relative locations of tones is illustrated in Figure 2-21. Suppose that in normal speech, the L and H occur at their respective anchors and the duration from L to H is the target duration. As speaking rate increases, the duration of segments decreases and the anchors get closer to each other, and thus the time available to produce a rise is reduced. If the duration of the rise remains constant across speech rates, in fast speech the L occurs earlier than A_L, and later than A_L in slow speech. For H, H occurs after A_H in fast speech, and earlier than A_H in slow speech. These are the kind of results that we found in the deviations of L1 and H1 in Seoul Korean.

Under the segmental anchoring hypothesis, it is assumed that tonal targets are indepen-

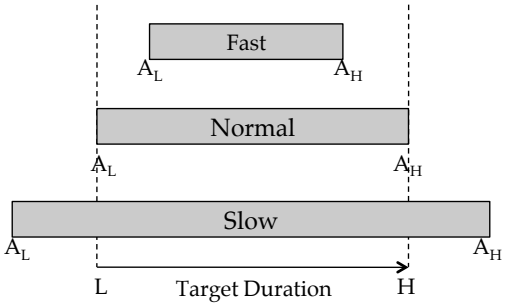


Figure 2-21: The effect of a target duration

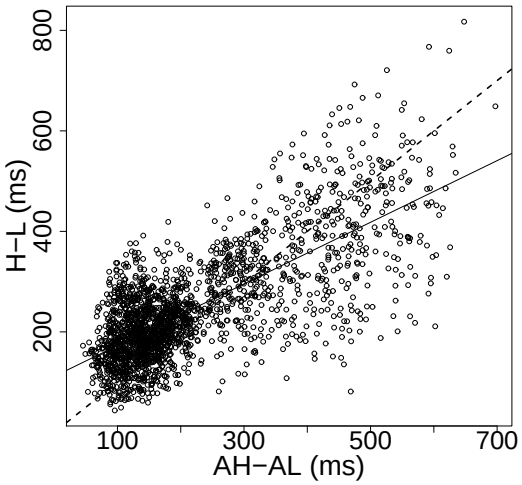


Figure 2-22: $(H - L)$ against $(A_H - A_L)$. The solid line is the regression line, the dashed line is the $y = x$ line.

dent of each other. It has been argued that L and H tones are independently aligned with regard to the anchoring points, thus the duration between L and H is entirely determined by the duration between the anchors (Arvaniti et al., 1998). However, the deviation patterns of L1 and H1 shown above leads us to believe that L1 and H1 are related to each other, while also aligned with regard to segments. If the timing of tones are predicted straightforwardly from their anchoring points only, the duration of the rise $(H - L)$ and the distance between the anchors $(A_H - A_L)$ must have a linear relation with a slope of 1. However, this is not what is observed. In Figure 2-22, $(H - L)$ is plotted against $(A_H - A_L)$. The plot shows that $(H - L)$ and $(A_H - A_L)$ has a positive correlation, but $(H - L)$ does not change as much as $(A_H - A_L)$ changes. This means that there is a tendency to maintain a target duration, as well as segmental anchoring. The model predicting tonal timing thus requires two factors: segmental alignment and target duration. This kind of a model is formalized in Chapter 3.

2.4.2 Scaling

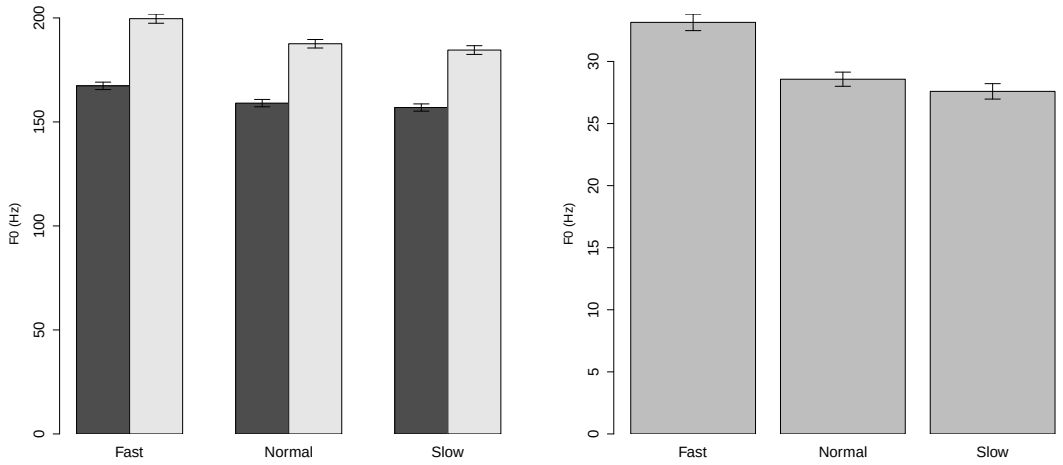
Speech rate effects

In the previous literature, two patterns have been predicted regarding the scaling of a pitch rise: the rise magnitude is relatively consistent, and slope gets shorter as speaking gets faster; or the slope is constant, and the magnitude of the rise gets smaller as speaking gets faster. In Korean, the rise magnitude increases as speaking gets faster, which is not predicted by any of these models. The rise magnitude is relatively stable only within the same speech condition (fast, normal, slow).

The SAH literature has claimed that the duration of a rise is derivable from the location of the segmental anchors. In Greek, there was no correlation between the pitch difference between L and H and the interval between L and H (Arvaniti et al., 1998: 17). Pitch scaling was not affected by duration. The authors interpreted this result as counter-evidence to a constant shape, because if the slope had been constant, there should have been a systematic relation between magnitude of rises and segmental duration, e.g. magnitude of rises decreases as segmental duration increases. In English, the magnitude of rises was not systematically affected by speech rate in most speakers (Ladd et al., 1999): four out of six speakers maintained the magnitude of rises almost constant across speech rates. Based on these findings, Ladd (2004) claimed that the slope of a rise is not constant, but both slope and duration are almost entirely determined by segmental anchoring of tones. That is, slope gets steeper and duration gets shorter as speaking gets faster. On the other hand, in other languages, the size of a pitch rise is reduced in fast speech. In Russian, segmental anchoring is maintained, two out of seven speakers had a smaller excursion size at fast rate (Igarashi, 2004). There was a significant negative correlation between excursion size and rise duration (i.e. as speech rate increases, excursion size decreases). French showed the same trend (Fougeron and Jun, 1998). Speakers reduced magnitude of rises in fast speech.

Seoul Korean shows two patterns that have not been observed in the studies mentioned above. First, overall pitch range increases in faster speech condition ('raising') in the terminology of Ladd (2008: 198). This is shown in Figure 2-23a. The pitch range is higher in fast speech than in normal or slow speech. Second, the magnitude of the rise increases in faster speech ('widening') (Ladd, 2008: 198). This is shown in Figure 2-23b: the magnitude of rises is wider in fast speech than in normal speech. Ladd (2008) discusses the relation between the two dimensions of pitch scaling: pitch level (the absolute pitch height) and pitch span (the difference between F0 minima and maxima). The relation between pitch level and span is not straightforward. Pitch raising often accompanies widening of pitch span, but the two dimensions are not necessarily dependent, i.e. one may speak in a narrow pitch span in a high pitch range, or in a wide pitch span in a low pitch range. Yet, they are at least partially dependent because they tend to co-vary, i.e. there is a tendency that pitch raising increases pitch span (Ladd, 2008: 197-8). In Seoul Korean, pitch levels of both L and H tones tend to increase in the fast speech condition. We suspect that when speaking fast, speech becomes more effortful, which stiffens, rather than relaxes, speech organs including vocal folds. Stiff vocal folds facilitates producing higher pitch ('t Hart et al., 1990).

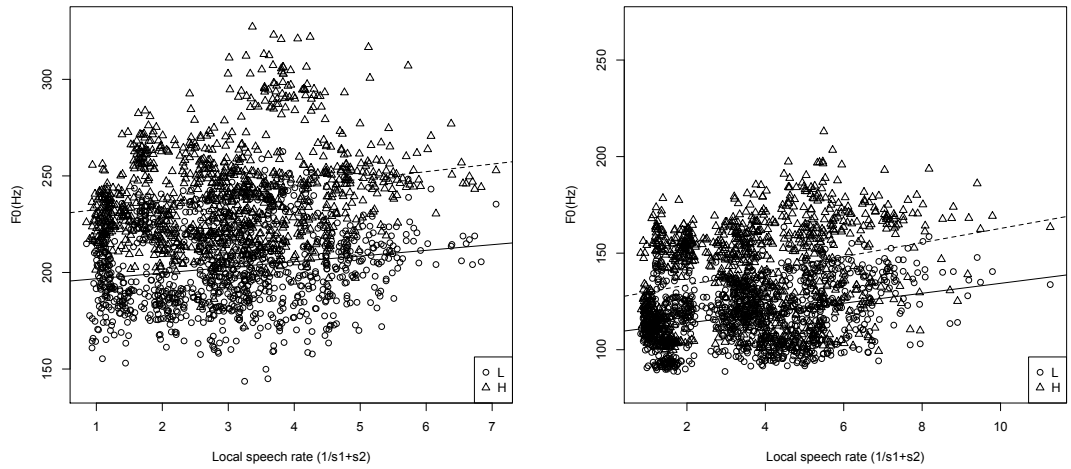
Figure 2-24 shows L and H levels against local speech rate. Since female and male



(a) L1 and H1 Levels

(b) Magnitude of rises

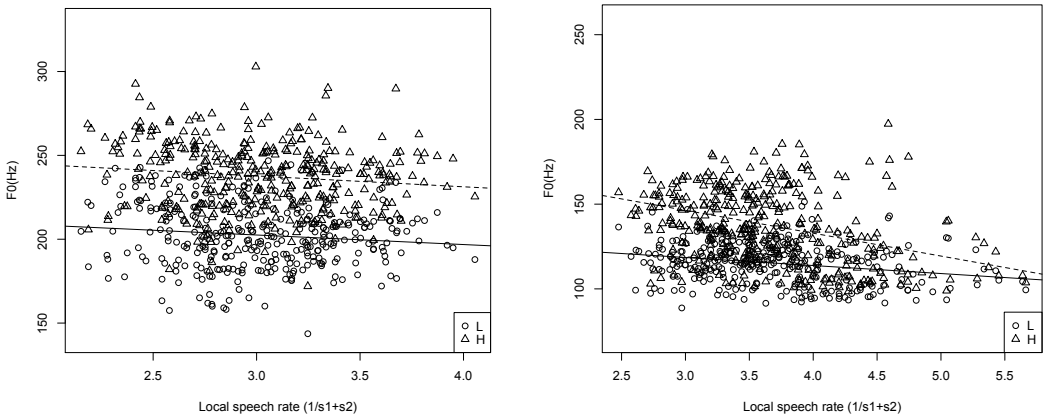
Figure 2-23: (a) L1 and H1 levels for fast, normal, and slow speech. Dark bars are L1. Light bars are H1. (b) Magnitude of Rise1 (H1-L1) for each speech rate category. The error bars are the mean ± 1 standard error. Pooling across all speakers.



(a) Females

(b) Males

Figure 2-24: L1 and H1 levels for all speech conditions (fast, normal, slow): (a) Females, (b)Males. The solid line is fitting L1 levels, the dashed line is fitting H1 levels.



(a) Females (b) Males

Figure 2-25: L1 and H1 levels for normal speech only: (a) Females, (b)Males. The solid line is fitting L1 levels, the dashed line is fitting H1 levels.

speakers have distinct pitch ranges, they are plotted separately. In both females and males, pitch range increases as local speech rate increases. However, it seems that categorical speech rate and local speech rate have different effects. Whereas slow speech has a smaller magnitude of rise than the other two speech rates, within a given categorical speech rate, excursion size tends to be constant, which means that the magnitude is independent of local speech rate or segmental makeup. In Figure 2-25, the levels of L1 and H1 in normal speech were plotted against local speech rate. In female speakers, both L and H levels are slightly decreasing with increasing speech rate, but the difference between the levels is relatively constant. In male speakers, H1 level is substantially reduced at faster rates (shorter segment duration), apparently the familiar pitch reduction effect at fast rates that has been found in other languages. However, this reduction effect is just due to speaker variability. As will be shown soon, local speech rate does not significantly affect the level of H, if speaker variability and segmental effects are taken into account. This can be examined by mixed-effects modeling.

Summarizing the observations, categorical speech rate has an effect on pitch range: pitch range is higher in fast speech than slower speech. Within a given category, the magnitude of rises is constant, local speech rate has little effect on the magnitude of rises. By manipulating speech rate, we may have ended up manipulating other unknown factors (which might be linguistic or non-linguistic). So, speakers change not just categorical speech rate, but speech style is changed, e.g. speech becomes more effortful or excited in fast speech. Nevertheless, within the same speech style, the magnitude of rise is maintained fairly constant.

Mixed-effects models were fitted to the data to statistically test these observations. We tested the level of L1 and the level of H1 separately. For the level of L1, mixed models were compared as shown in Table 2.3. According to the LRT, model (e) turns out to be the best,

Table 2.3: Mixed models for L1 scaling. 'cat.rate': categorical speech rate (fast, normal, slow), 'loc.rate': local speech rate (1/s1+s2). Each LRT is with the immediately preceding model. (f) is significantly better than (d) [$\chi^2(4) = 16.73, p < 0.01$]. The best one is (e).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	L1	none	speaker	none	
(b)	L1	cat.rate	speaker	none	$\chi^2(2) = 405.08, p < 0.001$
(c)	L1	cat.rate	speaker	cat.rate	$\chi^2(5) = 485.27, p < 0.001$
(d)	L1	cat.rate, loc.rate	speaker	cat.rate	$\chi^2(1) = 7.30, p < 0.01$
(e)	L1	cat.rate	speaker	cat.rate, loc.rate	$\chi^2(3) = 13.90, p < 0.01$
(f)	L1	cat.rate, loc.rate	speaker	cat.rate, loc.rate	$\chi^2(1) = 2.83, p = 0.09$

Table 2.4: Mixed models for H1 scaling. Each LRT is with the immediately preceding model. (f) is significantly better than (d) [$\chi^2(4) = 14.06, p < 0.01$]. The best one is (e).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H1	none	speaker	none	
(b)	H1	cat.rate	speaker	none	$\chi^2(2) = 620.48, p < 0.001$
(c)	H1	cat.rate	speaker	cat.rate	$\chi^2(5) = 825.09, p < 0.001$
(d)	H1	cat.rate, loc.rate	speaker	cat.rate	$\chi^2(1) = 2.02, p = 0.16$
(e)	H1	cat.rate	speaker	cat.rate, loc.rate	$\chi^2(3) = 13.27, p < 0.01$
(f)	H1	cat.rate, loc.rate	speaker	cat.rate, loc.rate	$\chi^2(1) = 0.79, p = 0.37$

where the L1 level is the dependent variable, categorical speech rate is the only fixed effect, and random effects are by-speaker random intercepts and slopes for categorical and local speech rates. According to the best model, the difference between normal and fast speech was significant [Coefficient=-8.16 Hz, $t(2046) = -3.91, p < 0.001$] [2], the difference between slow and fast speech was significant [Coefficient=-9.06 Hz, $t(2046) = -2.99, p < 0.01$]. In this kind of linear modeling in general, one of the levels in the fixed effect (fast speech in this case) becomes the base mean, and the coefficients of the other levels (normal and slow speech) are added to find the mean of each level. Thus, the best model result means that L1 level is lower in normal speech than in fast speech by about 8 Hz, and it is lower in slow speech than in fast speech by about 9 Hz. Considering these coefficients, the level of L1 is significantly higher in the order of fast, normal, and slow. Assuming the L1 level is the baseline pitch of speaker's pitch range, we can interpret this result to mean that pitch range increases in the faster speech condition. The effect of local speech rate was insignificant, as shown in the comparison of (f) and (e) in Table 2.3, corresponding to the observations described above.

For the level of H1, the models shown in Table 2.4 were compared. The best model is

[2]There is no agreed method of estimating degrees of freedom and p-values in mixed-effects models, because of the complications due to random effects (Baayen et al., 2008: 396) (See also the discussion by Douglas Bates 'lmer, p-values and all that' at https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html). In this dissertation, degrees of freedom (if they appear) were approximated with the number of observations. This is for the rather presentational purpose, given that once the degrees of freedom are more than 30, they make practically no differences on the p-values (Aiken and West, 1991).

(e), where the L1 level is the dependent variable, categorical speech rate is the only fixed effect, and random effects are by-speaker random intercepts and slopes for categorical and local speech rates. This means that H1 level is significantly different in each speech condition, but is not significantly affected by local speech rate once in the same speech style. According to the best model, the difference between normal and fast speech was significant [Coefficient=-12.16 Hz, $t(2086) = -2.96, p < 0.01$], the difference between slow and fast speech was significant [Coefficient=-17.91 Hz, $t(2086) = -2.99, p < 0.001$]. This means that H1 level was higher in the order of fast, normal, slow, after taking into account variations caused by the speaker random effect.

In summary, both L1 and H1 levels were significantly affected by categorical speech rate, but local speech rate (or segment duration indicated by local speech rate) did not have significant effects on the levels of L1 and H1. Pitch levels were higher in the order of fast, normal, and slow speech. Increasing the magnitude of the rise in faster speech is not predicted by any of the previous models.

Segmental effects

The L1 and H1 levels are not much affected by segmental duration, but it does not mean that the levels are constant. There are still a lot of unexplained variations in pitch levels. Further examinations reveal that some of the variations can be factored out if segmental effects are considered. Vowels have intrinsic pitch; if other conditions are equal, high vowels tend to have higher pitch than mid or low vowels (Ewan, 1975; Stevens, 1998). Also, vowels tend to have higher pitch after obstruents than after sonorants in Seoul Korean (Jun, 2000) and Kyungsang Korean (Kenstowicz and Park, 2006).

Several mixed models were fitted to the data and the likelihood ratio tests were carried out to test the segmental effects on the magnitude of the rise (the pitch difference between L1 and H1 levels). The segmental effects include vowel height (high, mid, low) and consonant manner (obstruent, sonorant). The best model was found with the fixed effects of vowel height and consonant manner (V1Height, V2Height, C1, and C2), and by-speaker random intercepts and slopes for V1Height, V2Height, C1, C2, and categorical speech rate. This means that in predicting the size of pitch rise, local speech rate is not necessary, and categorical speech rate is significant only as by-speaker random adjustments, but not as a fixed effect.

The effects of vowel height and consonant manner were in the direction we expected. The coefficients for the fixed effects were shown in Table 2.5. In linear models, the coefficients are added to the intercept to find the mean of a certain condition: e.g. the rise magnitude for a low V1 is $30 + 7.6 = 37.6$. Thus, a positive coefficient corresponds to an increase in the magnitude, a negative coefficient corresponds to a decrease in the magnitude due to a certain effect. A low or mid vowel in the first syllable lowers L1 level because of their low intrinsic pitch, and thus increases the rise magnitude. V1Height=Low and V1Height=Mid had positive coefficients [7.6 and 4.4]. Similarly, a low or mid vowel in the second syllable decreases H1 level and thus reduces the rise magnitude, because of their low intrinsic pitch. V2Height=Low and V2Height=Mid had negative coefficients [-3.2 and -2.0].

Table 2.5: Segmental effects on the magnitude of a rise (unit: Hz)

		Coefficient	Standard error	
Intercept		30	4.0	$t(2031) = 7.5, p < 0.001$
V1Height	Low	7.6	1.3	$t(2031) = 5.8, p < 0.001$
	Mid	4.4	0.8	$t(2031) = 5.9, p < 0.001$
V2Height	Low	-3.2	0.5	$t(2031) = -6.7, p < 0.001$
	Mid	-2.0	0.6	$t(2031) = -3.4, p < 0.001$
C1	Obstruent	-5.3	1.1	$t(2031) = -4.9, p < 0.001$
	Sonorant	1.1	0.8	$t(2031) = 1.3, p = 0.19$
C2	Sonorant	-3.2	1.1	$t(2031) = -2.9, p < 0.01$

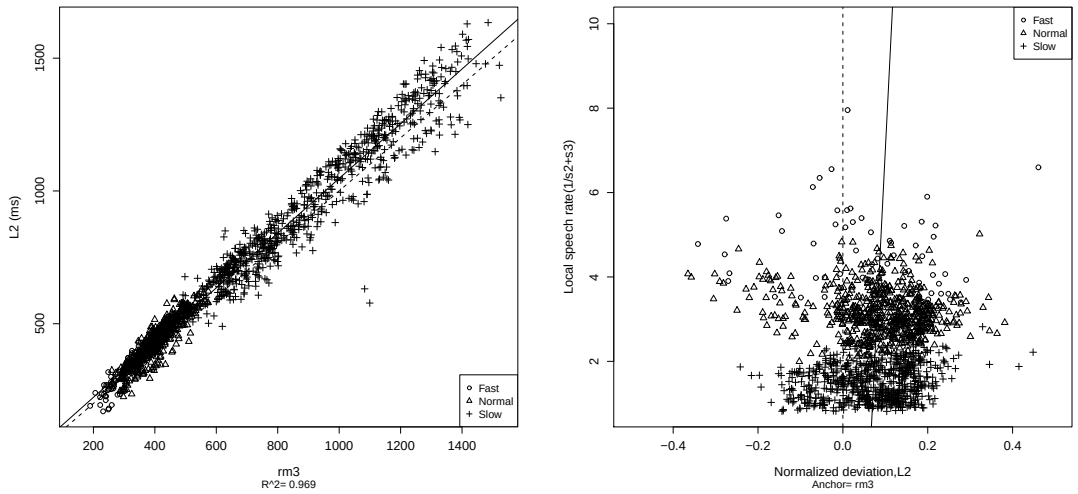
The manner of the onset consonant exhibited similar effects. Vowels after obstruents have higher pitch than those after sonorants, so obstruents in the first syllable raise L1 level, and sonorants in the second syllable reduce H1 the excursion size. An obstruent in the first syllable raises L1 level, and thus decreases the rise magnitude. The coefficients were negative when C1 is obstruent [-5.3]. On the other hand, a sonorant in the first syllable lowers L1, and thus increases the rise magnitude; although the coefficient is not significantly different from zero [t(2031)=1.3, p=0.19], the confidence interval is mostly above zero (-0.5 to 2.7), so there is a small tendency that we expected. A sonorant in the second syllable lowers H1 level, and thus decreases the magnitude.

In summary, Seoul Korean showed both pitch raising and widening in fast speech, which is not predicted by previous models of F0 movements. Within the same speech condition, temporal duration does not directly affect the magnitude of the rise. The observation that segmental duration does not affect the rise magnitude does not mean that the rise magnitude is constant: the rise magnitude varies due to other factors. The rise magnitude is significantly affected by orthogonal factors such as speech style (fast, normal, slow condition) and intrinsic vowel pitch.

2.5 Fall

2.5.1 Alignment of L2

The alignment of L2 is relatively stable compared to H1. We tested various segmental landmarks ('vm2','rm2','o3','c3','v3','rm3','o4','rm4'; see 2-12 for the definitions.) to find the one that has the best correlation with L2. Mixed-effects models were fit for each segmental point with timing of L2 as the dependent variable, timing of the segmental point as a fixed effect, with by-speaker random intercepts and by-speaker random slopes for segmental position. The point with the lowest deviance (15164) among all candidate points was 'rm3', the middle of the rime in the third syllable. In Figure 2-26a, the timing of L2 is plotted against this anchoring point. Note that it is a similar alignment point (middle of the rime in the associated syllable) that emerges as best fit for L1, H1, and L2. However, this estimate is based on the model predicting tonal timing with its anchor only; the precise anchor estimates are different in the model proposed in Chapter 3.



(a) Alignment of L2

(b) Deviation of L2

Figure 2-26: (a): L2 against A_{L2} (rm3), the dashed line is $y = x$, (b): Deviation of L2 from A_{L2}

The best fitting model predicting L2 by its anchor was the model with the anchor as a fixed effect and by-speaker random intercepts and by-speaker random slopes for the anchor. The coefficient of the anchor in the best-fitting model was 1.01, which was not significantly different from 1 [$t(1399) = 0.02, p = 0.98$]. This means that segmental anchoring is very strict in L2.

We also examined the deviation of L2 from its anchor. The normalized deviation plot is shown in Figure 2-26b. In the case of L1, there was a negative trend; L1 was anticipated in fast speech and delayed in slow speech. On the other hand, L2 is slightly delayed at faster speech rates. Mixed-effects modeling shows that the effect of local speech rate on the L2 deviation was significant [$\chi^2(3) = 83.94, p < 0.001$]. The coefficient of the speech rate was 0.02, thus, the deviation of L2 in the positive direction was small but significant.

The two outliers in Figure 2-26a turned out to be due to different phrasing in slow speech. One of the two outliers is shown in Figure 2-27. The initial rise is realized on the first syllable. There are two possible interpretations of this F0 contour: it may be a LHLH pattern, stretched between two rises, with the shape of the initial rise maintained. Or, it may be different phrasing: [LHLH] becomes [LH][LLH] due to slow rate. The second analysis is justifiable because the morphemic structure of the content word [nonara] is no+nara ('No+Country'), so it is easy to separate into two phrases.

From the result of L2 deviation, we can hypothesize that the delay of L2 at faster rates is due to the delay of the preceding tone, H1. We tested whether the timing of H1 affects the timing of L2. Mixed-effects modeling showed that H1 was significant in predicting L2 [$\chi^2(4) = 608.21, p < 0.001$]. This means that L2 is significantly affected by the timing of H1. The later the H1, the later the L2. However, the effect of H1 is very small: L2 is not delayed as much as H1. L2 shows a very strict segmental alignment compared to H1.

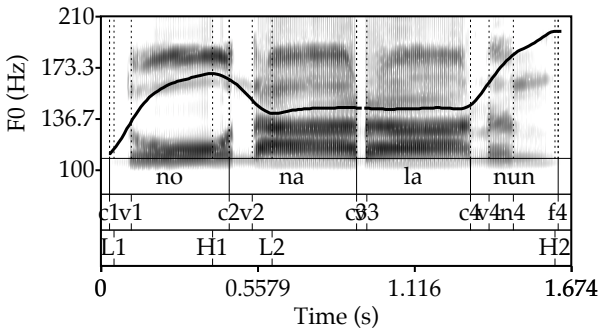


Figure 2-27: *nonalanun* [nonaranin] 'The No Dynasty-TOP' read with a rise on the first syllable

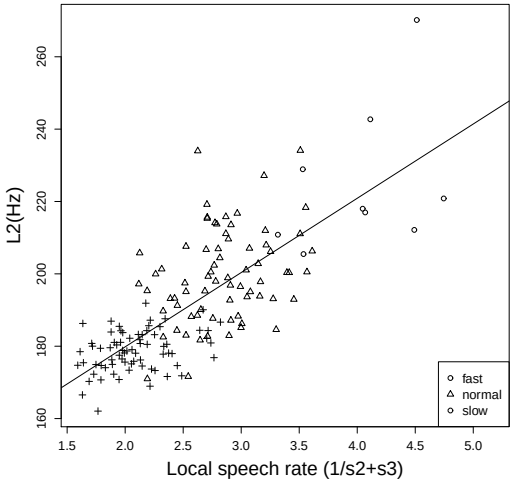


Figure 2-28: L2 level against duration of s2+s3 (Speaker A5)

2.5.2 Scaling of L2

L2 Level

In Section 2.3, we showed that L2 can be variably realized from full scaling to undershoot to truncation, arguing that truncation is the final state of extreme compression, or accumulation of gradual phonetic changes. We will now reaffirm this claim in a quantitative way by presenting the analyses of the level of L2, magnitude of the Fall, and the relation between the level of L2 and the level of H1.

The level of L2 tends to increase as speech rate increases, which means that L2 is 'undershot' in faster speech rates. Figure 2-28 shows L2 level plotted against local speech rate (1/(s2+s3)). The result of only one speaker (A5) is presented. The regression line fits the L2 level against local speech rate. A similar positive correlation between L2 level and speech rate is observed in all speakers.

To examine the effect of local speech rate on the level of L2 across speakers, several

Table 2.6: L2 scaling and speech rate. The best model is (e). Each LRT from (b) to (e) is with the immediately preceding model. The LRT in (f) and (g) are compared with (e).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	L2	none	speaker	none	
(b)	L2	loc.rate	speaker	none	$\chi^2(3) = 1517.8, p < 0.001$
(c)	L2	loc.rate	speaker	loc.rate	$\chi^2(2) = 613.2, p < 0.001$
(d)	L2	loc.rate, cat.rate	speaker	loc.rate	$\chi^2(2) = 111.52, p < 0.001$
(e)	L2	loc.rate, cat.rate	speaker	loc.rate, cat.rate	$\chi^2(7) = 165.15, p < 0.001$
(f)	L2	cat.rate	speaker	loc.rate, cat.rate	$\chi^2(1) = 8.81, p < 0.01$
(g)	L2	loc.rate	speaker	loc.rate, cat.rate	$\chi^2(2) = 12.95, p < 0.01$

mixed-effects models were compared. The results are shown in Table 2.6. The best model was (e), where both local and categorical speech rates were fixed effects. Removal of any of the fixed effects was significant ((e) vs. (f),(g)). In the best model, the coefficient of local speech rate was 8.3, implying a positive correlation between local speech rate and L2 level, that is, the faster the rate, the higher L2. The difference between slow and fast speech was significant [Coefficient=-22 Hz], which means that L2 level is lower in slow speech than in fast speech rate. The difference between normal and fast speech was not significant [Coefficient=-11, t(1398)=-1.6, p=0.12].

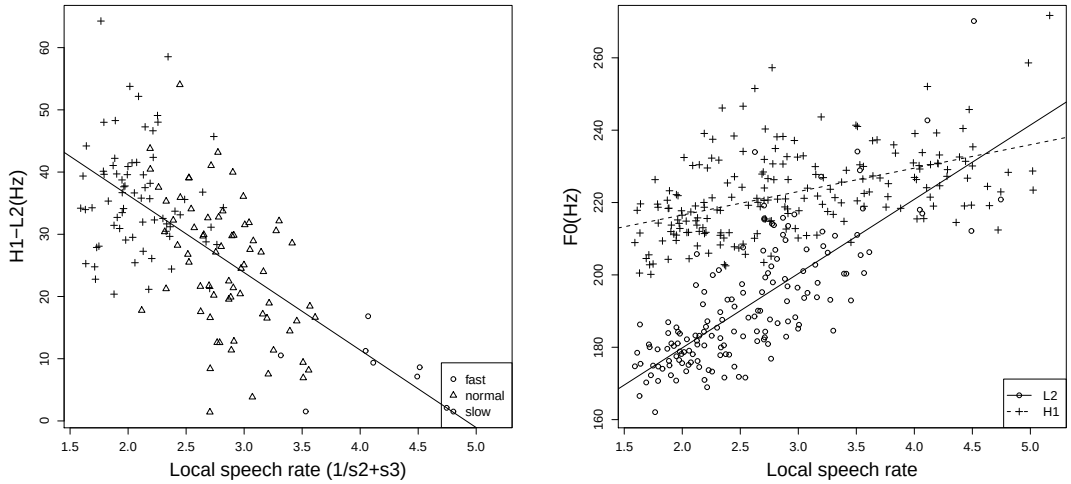
The magnitude of the Fall

The magnitude of the Fall (the difference between H1 and L2 levels) is gradually reduced as speech rate gets faster. This trend is shown in Figure 2-29a. Falls are less likely to be fully realized under time pressure. Whereas the magnitude of the Rise1 was relatively stable within the same speech rate category, the magnitude of the Fall varies systematically depending on time pressure even within the same speech rate category.

Mixed models were compared to examine the effects of speech rate on the magnitude of the Fall, the results are shown in Table 2.7. The best-fitting model is (f), with the magnitude of the Fall as the dependent variable, local speech rate as a fixed effect, with by-speaker random intercepts and slopes for both local and categorical speech rate. The coefficient for local speech rate in the best-fitting model was -6.3 Hz, which means that there was a significant negative correlation between local speech rate and magnitude of Fall. I.e. the faster the rate, the shallower the Fall.

Also notice that categorical speech rate is not significant once local speech rate is a fixed effect: the addition of local speech rate is not significant ((e) vs. (f)). This is an interesting difference from the magnitude of Rise1. We have shown in Section 2.4.2 that in predicting the size of Rise1, local speech rate is not a significant fixed effect. The magnitude of Rise1 was affected by categorical speech rate only. On the other hand, the magnitude of the Fall is significantly affected by local speech rate, but categorical speech rate was not significant (significant only as a by-speaker random effect). This suggests that changes in the magnitude of the Fall are gradual, rather than categorical, phenomena.

In Figure 2-29b, H1 levels and L2 levels are plotted together against local speech rate.



(a) Magnitude of Fall

(b) H1 and L2 levels

Figure 2-29: (a):Magnitude of Fall (H1-L2) against local speech rate (1/(s2+s3)), (b): Convergence of H1 and L2 levels. Both plots are from one speaker (A5).

Table 2.7: Magnitude of the Fall and speech rate. The best model is (f). Each LRT from (b) to (e) is with the immediately preceding model. The LRT in (f) and (g) are compared with (e).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H1-L2	none	speaker	none	
(b)	H1-L2	loc.rate	speaker	none	$\chi^2(1) = 488, p < 0.001$
(c)	H1-L2	loc.rate	speaker	loc.rate	$\chi^2(2) = 375.41, p < 0.001$
(d)	H1-L2	loc.rate, cat.rate	speaker	loc.rate	$\chi^2(2) = 54.22, p < 0.001$
(e)	H1-L2	loc.rate, cat.rate	speaker	loc.rate, cat.rate	$\chi^2(7) = 108.74, p < 0.001$
(f)	H1-L2	loc.rate	speaker	loc.rate, cat.rate	$\chi^2(2) = 3.24, p = 0.20$
(g)	H1-L2	cat.rate	speaker	loc.rate, cat.rate	$\chi^2(1) = 9.64, p < 0.01$

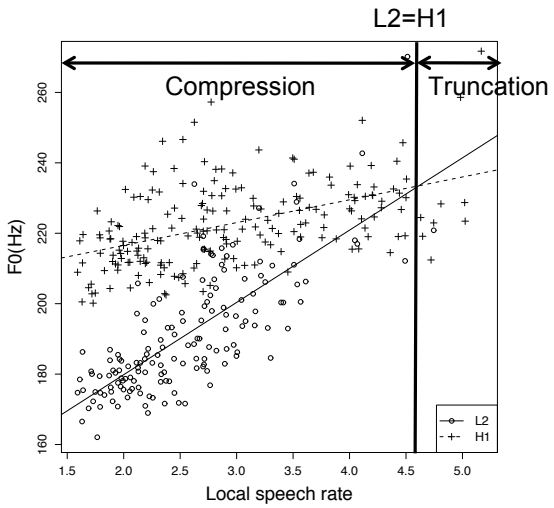


Figure 2-30: The intersection point of H1 and L2 levels: the LHLH sequence is compressed before the intersection point and truncated to LH after the intersection point.

Table 2.8: H1 level with local speech rate as the fixed effect. The best model is (c). Each LRT is with the immediately preceding model.

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H1	loc.rate	speaker	none	
(b)	H1	loc.rate	speaker	loc.rate	$\chi^2(2) = 453.54, p < 0.001$
(c)	H1	loc.rate	speaker	cat.rate	$\chi^2(3) = 363.61, p < 0.001$
(d)	H1	loc.rate	speaker	loc.rate, cat.rate	$\chi^2(4) = 6.92, p = 0.14$

The solid line fits L2 levels, the dashed line fits H1 levels. The dashed line shows that H1 level slightly increases at faster speech rate. As we have already said, this increase of H1 is categorical shift, rather than gradual local changes. At the same time, L1 level increases, but unlike the case of H1, this is a gradual change as a function of local speech rate. The slope of the regression line is steeper for L2 level than for H1 level. Because of this, the regression lines for H1 and L2 intersect with each other at some point (in the plot, at about 4.6 local speech rate), as illustrated in Figure 2-30.

As shown in Figure 2-30, we argue that the LHLH tone sequence is compressed by undershooting the L2 before the intersection point, and after the intersection point, the LHLH pattern is truncated to the LH pattern. That is, in connection with the phonetic realization types of the LHLH AP discussed in Section 2.3, we may interpret these results as the following. When there is enough time (slow rates), the Fall is fully realized (Figure 2-17(a)). The level of L2 gradually rises as speech gets faster (the undershoot effect shown in Figure 2-17(b)). At the intersection point of the two fitting lines, the level of L2 reaches to the same level as H1, resulting in a plateau shape (Figure 2-17(c)). After this point, the LHLH pattern is truncated into a LH pattern (Figure 2-17(d)).

To precisely estimate the intersection point across speakers, linear mixed-models were

Table 2.9: L2 level with local speech rate as the fixed effect. The best model is (d). Each LRT is with the immediately preceding model.

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H1	loc.rate	speaker	none	
(b)	H1	loc.rate	speaker	loc.rate	$\chi^2(2) = 613.2, p < 0.001$
(c)	H1	loc.rate	speaker	cat.rate	$\chi^2(3) = 186.97, p < 0.001$
(d)	H1	loc.rate	speaker	loc.rate, cat.rate	$\chi^2(4) = 76.76, p < 0.001$

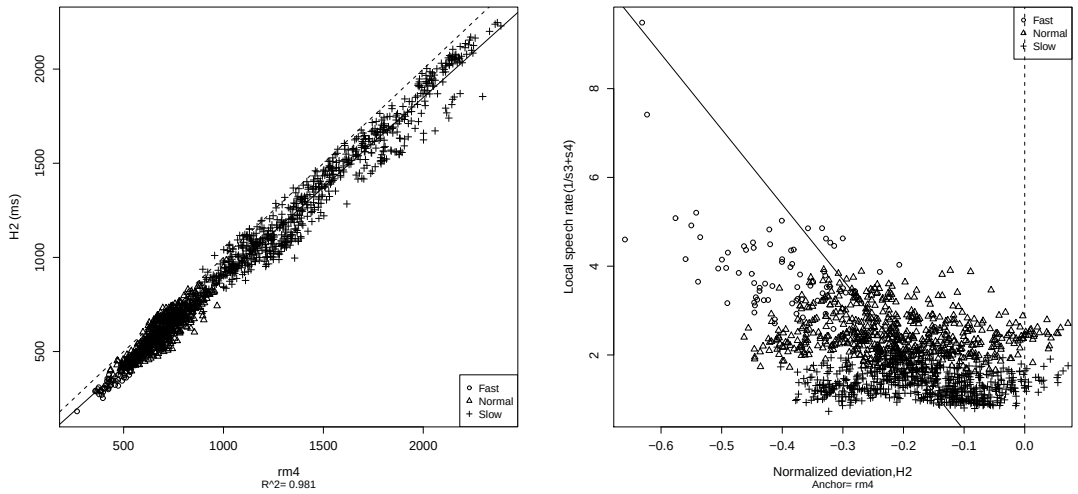
fitted to the data for H1 level and L2 level respectively, with local speech rate as the fixed effect. Only the random effects were compared, as shown in Table 2.8 and Table 2.9. In both H1 and L2, local rate is the inverse of the second and third syllable duration ($1/(s2 + s3)$) where the Fall spans. The best model for H1 level is (c) and the best model for L2 level is (d). According to the best models, the level of H1 and L2 are expressed as in (3), where $P(H1)$ is the pitch level of H1, $P(L2)$ is the pitch level of L2, r is local speech rate ($= 1/(s2 + s3)$).

(3) a. $P(H1) = 1.48 \cdot r + 173$
b. $P(L2) = 9.05 \cdot r + 131$

The intersection point of the two straight lines in (3) can be found by setting the left terms equal: $P(H1) = P(L2)$. By solving this equation, the local speech rate (r) of the intersect point is 5.5. This means that after taking speaker variability into account, there is a point where the pitch level of L2 and the pitch level of H1 equalize. The difference between the two levels is a function of local speech rate, and when local speech rate reaches at about 5.5 (that is, when the duration of the first and second syllable reaches at 180 ms ($1/5.5$), the level of H1 and L2 equalizes, yielding a plateau shape. At this stage, L2 realization is skipped.

As mentioned in Section 2.3, L2 undershoot is a very common compression strategy found in many languages. Languages also show changes in tonal sequence, e.g. the truncation of the final L in monosyllables Hungarian. Ladd (2008: 182) suggests that compression versus truncation strategies are a matter of cross-linguistic preferences, that is, some languages prefer compression (English (Grabe et al., 2000) and Greek (Arvaniti et al., 2006)) whereas other languages prefer truncation: Hungarian (Gósy and Terken, 1994) and Palermo Italian (Grice, 995a).

In Seoul Korean, in fast speech, we found both compression (L2 undershoot) and truncation (one sharp rise in fast speech). The LHLH pattern is realized with a plateau or as a sharp rise and fall in fast speech. The plateau pattern results from extreme undershoot of L2, thus we argue that in that case, underlyingly L2 is present. We suggest that compression and truncation are not separate strategies; truncation (LH) is the next stage after extreme undershoot (a plateau pattern). We showed that speech rate has a gradual effect on L2 level. After speech rate exceeds the point where L2 level and H1 level are the same, L2 is truncated. In this sense, the accumulated effects of compression of LHLH tones lead to truncation. This shows a close connection between phonetics and phonology: an accumulation of gradual phonetic changes may result in a phonological categorical change. The



(a) Alignment of H2

(b) Deviation of H2

Figure 2-31: (a) H2 against A_{H2} (rm4), the dashed line is $y = x$, (b) Deviation of L2 from A_{H2}, the dashed line is the time of the anchor

'truncation languages' (Hungarian and Palermo Italian) may exhibit gradual compression effects if appropriately examined. This is a task for future research.

Furthermore, it is noteworthy that rises fit in the same pattern shown in Section 2.4.1, regardless of whether they are from a full or a truncated pattern. This may support our claim that LHLH and LH patterns are quantitatively closely related to each other.

2.6 Rise2

2.6.1 Alignment of H2

As with other tones, we tested several segmental points as possible anchoring points. The segmental points were 'o3' (the onset of the third syllable), 'rm3' (the middle of the third rime), 'o4' (the onset of the fourth syllable), 'rm4', (the middle of the fourth rime), 'f4' (the end of the fourth syllable). To take speakers' random effects into account, mixed-effects models were fitted for each segmental point with timing of H2 as a dependent variable, timing of a segmental point as a fixed effect, by-speaker random intercepts, and by-speaker random slopes for timing of the segmental landmark. The point with the lowest deviance was 'rm4' ($G^2 = 15161$). The coefficient of the anchor in the best-fitting model was 0.959, reflecting strict segmental anchoring. In Figure 2-31a, timing of H2 is plotted against this anchoring point.

The deviation of H2 shows the opposite trend from the deviation of H1: it is negatively correlated with speaking rate (Figure 2-31b). That is, the faster the rate, the earlier the H2 peak. Mixed-modeling was carried out to take the random effects of speaker into account, as shown in Table 2.10. The coefficient of local speech rate in the best fitting model was

Table 2.10: Mixed model comparisons for the model predicting H2 from A_{H2}. The LRT is with the immediately preceding model. The best model is (c).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H2	none	speaker	none	
(b)	H2	rm4	speaker	none	$\chi^2(1) = 233.14, p < 0.001$
(c)	H2	rm4	speaker	rm4	$\chi^2(2) = 381.76, p < 0.001$

-0.07, suggesting a significant negative correlation between local speech rate and deviation of H2 from the anchor.

This result is in a sense unexpected, because from the deviation pattern of H1, one might also expect a similar direction of H2 deviation, given that the alignment of L2 was also highly stable. We hypothesize a possible explanation: H2 is retracted in order to be at a fixed distance from the following F0 minimum. Accentual Phrases are always immediately followed by a local F0 minimum, regardless of speech rate. This can be seen in examples such as Figure 2-17. We cannot determine whether the F0 minimum after the AP is a boundary L% tone of the preceding AP, or the initial L tone in the carrier phrase, but a most likely interpretation is that the F0 minimum is the boundary L% and the initial L in the next phrase on the top of each other because there is only one F0 minimum following the H2. Regardless of this point, target APs are followed by F0 minima. We tentatively refer to this tone as a L% boundary tone. We may hypothesize that the retraction of H2 is because there is a duration target of fall from H2 to this L%. As in the case of Rise1, if there is a target duration of this fall, we might expect that L% should deviate from its anchor in the opposite direction of H2. That is, as H2 is retracted at a faster rate, L% is delayed at a faster rate. This is just the mirror image of Rise1. So, we preliminarily tested this hypothesis: L% will be delayed with regard to an anchor in the faster rate.

Since most of AP's were produced as IP's in slow speech, we measured fast and normal speech only in order to avoid complications arising from different phrasing. We measured the F0 minima that followed H2, for speaker A1 only. We compared timing of L% with timing of the end of the phrase ('f4'). Figure 2-32 shows the deviation pattern of L%. The direction of L% deviation is the opposite of H2 deviation. L% occurs later than the phrase offset when speech rate is faster, and occurs earlier than the phrase offset when speech rate is slower. Because we measured only one speaker, instead of a mixed-effects model, we fit a linear regression model with normalized deviation as a dependent variable, local speech rate as a predictor. The effect of local speech rate on the deviation of L% was significant [$F(1, 139) = 93.65, p < 0.001$]

The relation between H2 and L% is the mirror image of the one between L1 and H1. At the beginning of the AP, as speech rate increases, L1 is retracted and H1 is delayed with regard to the anchors. At the end of the AP, as speech rate increases, H2 is retracted and L% is delayed with regard to the anchors. We interpreted the result of Rise1 as the presence of a target for rise duration. Likewise, we can also interpret the result in this section to mean that there is a target duration for a fall from H2 to L%. This way, the systematic deviation of H2 can be seamlessly explained.

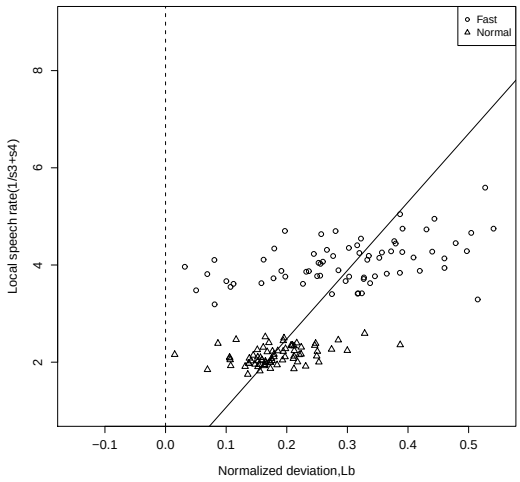


Figure 2-32: Deviation of L% phrasal boundary from the end of the phrase. The dashed line is the end of the phrase.

Table 2.11: Mixed model comparisons for H2 level. The best model is (a). The LRT's are all with (a).

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H2	none	speaker	cat.rate, loc.rate	
(b)	H2	cat.rate	speaker	cat.rate, loc.rate	$[\chi^2(2) = 2.14, p = 0.34]$
(c)	H2	loc.rate	speaker	cat.rate, loc.rate	$[\chi^2(1) = 0.08, p = 0.78]$
(d)	H2	none	speaker	cat.rate	$[\chi^2(4) = 23.43, p < 0.001]$
(e)	H2	none	speaker	loc.rate	$[\chi^2(7) = 198.87, p < 0.001]$

2.6.2 Scaling of H2

Neither categorical speech rate nor local rate significantly affected the H2 level. In Table 2.11, (b) shows that adding categorical rate to (a) as a fixed effect does not significantly improve the fit, and (c) shows that adding local speech to (a) as a fixed effect does not significantly improve the fit. By-speaker random slopes for both local and categorical rates were significant [(a) vs. (d) and (e)]. This means that there was no common pattern of H2 level across speakers.

On the other hand, the magnitude of the Rise2 (H2 level - L2 level) was significantly affected by categorical speech rate and rise duration (the time from L2 to H2), but not by local speech rate. The best-fitting model is model (d), the one with fixed effects of categorical speech rate and rise duration, by-speaker random intercepts and slopes for local speech rate, categorical speech rate, and rise duration. According to the coefficients in the best model, the magnitude of the Rise2 is significantly different between fast speech and normal speech by about 17 Hz, also between fast speech and slow speech by 33 Hz. The effect of the rise duration on the magnitude of the rise was small (0.07 Hz) but significant $[t(1386) = 2.604, p < 0.01]$. The results mean that magnitude of the Rise2 increases when

Table 2.12: Mixed model comparisons for Rise2 scaling. The best model is (d). 'cat' means categorical speech rate, 'loc' is local speech rate, and 'r.dur' is rise duration.

	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	H2-L2	cat	speaker	cat, loc	
(b)	H2-L2	cat, loc	speaker	cat, loc	$[\chi^2(1) = 0, p = 1]$ (with (a))
(c)	H2-L2	cat, r.dur	speaker	cat, loc	$[\chi^2(1) = 112.26, p < 0.001]$ (with (a))
(d)	H2-L2	cat, r.dur	speaker	cat, loc, r.dur	$[\chi^2(1) = 238.55, p < 0.001]$
(e)	H2-L2	cat, r.dur	speaker	cat, r.dur	$[\chi^2(5) = 45.18, p < 0.001]$ (with (d))
(f)	H2-L2	cat, r.dur	speaker	loc, r.dur	$[\chi^2(9) = 297.88, p < 0.001]$ (with (d))

speech gets slower, and rise duration has a tiny effect of increasing the magnitude when speaking slowly. The increase of magnitude in slow speech is categorical, rather than gradual, as local speech rate has no effect.

Comparison of H1 and H2 levels

According to Jun (1996), H1 is in general lower than H2. Lee and Kim (1997) reports that H1 was higher than H2 in about 30% of two speakers' utterances. Figure 2-33 shows our result regarding the difference between H1 and H2. The y-axis shows the difference between H2 and H1. The difference between H2 and H1 in each token was normalized by the pitch range (F0 maximum - F0 minimum) in each token, i.e. $(H2 - H1)/(F0max - F0min)$. The negative values mean that H1 was higher than H2. The positive values mean that H1 was lower than H2. Thus, H2 is generally lower than H1 in fast speech, higher than H1 in normal and slow speech. Thus, in normal and slow speech, the relation between H1 and H2 corresponds to what the previous studies reported. Also, H2 is higher in slow speech than in normal speech, so the difference between H1 and H2 increase when there is more time available for Rise2.

2.7 Summary

In this chapter, we have carefully examined in considerable detail phonetic realization of the LHLH Accentual Phrase contour in Seoul Korean. We have demonstrated how the shape and alignment of each component of the intonational pattern are affected by various factors, such as categorical speech rate, local speech rate, and intrinsic properties of segments. In Rise1, we found partial evidence for both segmental anchoring and target duration. L1 and H1 have a high positive correlation with a segmental anchor. At the same time, there was a systematic deviation of the tone from its anchor. L tends to occur earlier than the anchor in faster speech, and H tends to occur later than the anchor in faster speech. This means that a rise starts earlier and terminates later when speaking fast. Furthermore, the duration from L to H changes less than the duration between the anchors. These results point to the existence of a rise *duration target*. As for scaling, there was a tendency of pitch range raising and pitch span widening in faster speech. Within the same speech rate

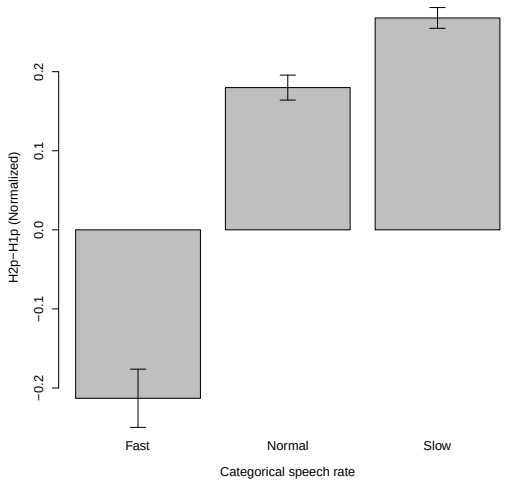
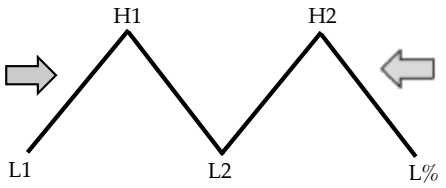


Figure 2-33: The difference between pitch level of H1 and H2. Normalized by the pitch range (F0 maximum-F0minimum) of each token and then averaged across speakers. That is, the bar is the H2-H1 level difference as a proportion in the pitch range. The error bars show mean $\pm$ standard error.

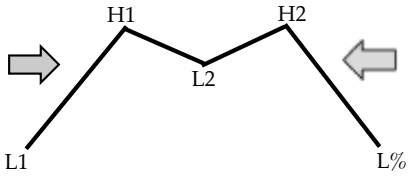
category, the magnitude of Rise1 was not significantly affected by local speech rate, but it was affected by segmental makeup such as intrinsic vowel duration.

As for the Fall, the level of L2 was significantly undershot at faster rates. This effect was gradual, rather than categorical, linearly correlated with local speech rate. When speech rate is fast enough, L2 level becomes the same as H1 level, resulting in a plateau shape between H1 and H2. When speech rate gets faster, the plateau part disappears, resulting in only one rise. This apparent categorical shift from compression to truncation involves gradual phonetic changes (i.e. undershoot). Unlike H1 level, L2 level was vulnerable to local time pressure; in fact, undershoot of L between H peaks is a common strategy of compression of tones in many other languages (English, Greek, Japanese). On the other hand, while H1 alignment varied significantly, L2 alignment was relatively stable. This may mean that for Rise1, maintaining a certain rise shape is important, and for that purpose, the segmental alignment of H1 is violated more. H1 misalignment would be more tolerable than L2, because L2 is too close to the phrase boundary, so there is not enough time left to delay L2. Instead, L2 is adjusted by undershoot of the pitch level, which is a well-motivated, cross-linguistically common strategy. In summary, pitch level is more important in H1 than in L2; alignment is more important in L2 than H1. This kind of asymmetry can be modeled in a quantitatively precise terms, which will be shown in Chapter 3.

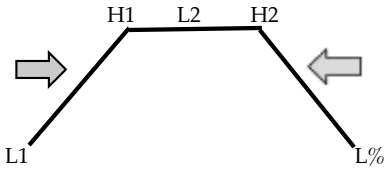
As for Rise2, H2 alignment was fairly stable, but it also shows a systematic deviation depending on local speech rate. The direction of deviation was different from that of H1. Whereas H1 deviation was positively correlated with speech rate, H2 was negatively correlated. That is, H2 occurs earlier with regard to the anchor as speech gets faster. A plausible explanation for this is the existence of a target duration for the fall to the next F0 minimum (L%). A preliminary examination of F0 minima indicates that the L% shows the opposite



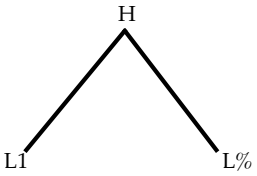
(a) Full scaling



(b) Undershoot



(c) Plateau



(d) Truncation

Figure 2-34: (a) Full scaling, (b) Undershot of L2, (c) Plateau, (d) Truncation

direction of deviation of H2. L% occurs later with regard to the end of AP as speech gets faster. However, if the observed F0 minimum is not a boundary tone, but a phrase-initial tone of the next AP, the F0 minimum must be aligned in the first syllable of the next AP. Depending on the different location of the reference point, the F0 minimum may show a different pattern. A further examination of this F0 minimum with regard to segments in next AP is necessary for a more accurate analysis.

In sum, we may conclude that Rise1 and the fall after H2 have a tendency to maintain a durational target. Figure 2-34 illustrates a possible interpretation of these results. That is, we may hypothesize a model of an Accentual Phrase, where there is a tendency to maintain Rise1 and boundary fall duration. This means that when there is time pressure, the whole LHLH phrase is pushed inward from both sides. As a result, the four tones undergo undershoot (2-34b) and compression (2-34c), and finally truncation (2-34d).

A remaining question is why it is the shape of the initial rise and the final fall that are preserved under time pressure, rather than the Fall from H1 to L2 or the Rise2 from L2

to H2. Cross-linguistic evidence points to a strong tendency that the L between H peaks is undershot. It has been suggested that peaks are perceptually more significant than F0 minima, so H peaks tend to be more consistently realized than the L between H peaks (Caspers, 1994; Prieto, 2005). Skipping L2 is less perceptually costly, so it may be the optimal solution for the production of tones under time pressure. The Korean LHLH AP fits in this cross-linguistic pattern.

Chapter 3

The Weighted-Constraint Model

3.1 The framework

In this chapter, we propose a model that explains phonetic realization patterns of the Accentual Phrase in Seoul Korean, presented in Chapter 2. In particular, the modeling will be focused on the timing of L1 and H1 in Rise1. In Section 2.4.1 in Chapter 2, we have shown that tendencies to both segmental anchoring and target duration are observed in the realization of L1 and H1. The initial H peak was aligned with regard to a segmental landmark (the middle of the second rime) as predicted by the Segmental Anchoring Hypothesis. At the same time, H1 peaks systematically deviated from the anchor depending on speech rate. H1 occurred later relative to the anchor at faster rates, and earlier at slower rates. L1 was more stably aligned than H1, but it also showed systematic deviations from its anchor. The L1 deviations were in the opposite direction from the H1 deviations. That is, L1 occurred earlier relative to the anchor at faster rates, later at slower rates. Based on these findings, we have proposed that there is a *target duration* between L1 and H1. Thus, the precise timing of tones is determined by segmental alignment as well as a duration target. Neither alignment nor duration is invariant, but they are only targets that are realized to different degrees, depending on how important a target is in a given language.

A property is considered important if it remains relatively stable under changes in speech conditions, even at the cost of other features. For H1, pitch level was relatively stable while segmental alignment changed at fast speech rates. This means that for H1, pitch level is more important than segmental alignment. Our analysis of Seoul Korean has also revealed that different parts in the tune have different importance in their realization. E.g. alignment of L1 was more stable than alignment of H1. It would have been also equally possible to vary L1 alignment more while keeping H1 alignment stable, or alternatively, to maintain both L1 and H1 alignment while reducing the magnitude of the initial rise (Rise1) if under time pressure. Then why do speakers delay H1 while maintaining the magnitude of the rise and the alignment of L1, rather than the other way around? We can conjecture that different parts of the tune have different importance; and some properties of a certain tone are more important to realize than other properties of the tone. That is, for the initial rise in Seoul Korean, maintaining a certain duration and magnitude of rise is of more importance to some degree than precisely hitting the H1 alignment target. A similar

example is found with L2. The level of L2 varied gradually as a function of local speech rate, while showing relatively stable alignment. This means that for L2, unlike H1, scaling is less important than alignment, so when there is time pressure scaling is affected more than alignment.

Considering these findings, we need a modeling framework that allows both shape and alignment targets to interact each other, while allowing some properties to be more resistant to changes in speech conditions such as speech rate than other properties. Moreover, for the model of L1 and H1 timing, we need a framework that expresses the relative importance of alignment and target duration. This is to model the experimental observations that H1 deviates to a greater degree than L1 while attempting to maintain a certain duration target. The constraint-based approach in Flemming (2001) is most appropriate for modeling these findings.

In Flemming (2001), a model with weighted constraints has been developed in order to model interactions of the constraints for scalar phonetic representations, such as F2 values of vowels (the acoustic correlate of vowel backness). Back vowels are fronted (phonetically) in the context of coronal consonants in many languages, e.g. English (as in the word 'taught' [$t^hɔt$]). According to Flemming, the partial assimilation between adjacent consonants and vowels can be viewed as a compromise between two constraints: achieving the F2 targets of the consonant and vowel and minimizing the movement between the two targets. As a result, F2 of a back vowel deviates from its target closer to F2 of the preceding consonant, and *vice versa*. Vowel F2 increases in the context of a coronal, and at the same time, F2 of the coronal consonant decreases in the direction of vowel F2. The resulting F2 values for the consonant and vowel are determined by the compromise among the constraints that require the target F2 values for the consonant and the vowel be realized and the constraint that requires that F2 values of adjacent vowels and consonants be the same. In other words, the actual F2 values are the results of a compromise among three factors: target F2 value for vowel (F2(V)), target F2 value for consonant (F2(C)), and minimizing effort. These constraints are illustrated in (1).

(1)

	Constraint	Cost of violation
Ident(C)	$F2(C) = L$	$w_c(F2(C) - L)^2$
Ident(V)	$F2(V) = T$	$w_v(F2(V) - T)^2$
MinimizeEffort	$F2(C) = F2(V)$	$w_e(F2(C) - F2(V))^2$

Ident(C) requires that the F2 value of the consonant (F2(C)) be the target F2 value for the consonant (L). Ident(V) requires that the F2 value of the vowel (F2(V)) be the target F2 value for the vowel (T). MinimizeEffort requires that the actual F2 values of the consonant and the vowel must be the same. These constraints are conflicting: MinimizeEffort penalizes any changes in F2 values in adjacent segments, whereas fully satisfying Ident constraints results in maximum violation of MinimizeEffort.

The conflict is resolved by finding the set of values that minimizes the violations of each constraint. The weights of the constraints determine the relative importance of the constraints. Violations of constraints with a lower weight incur less cost of violations, thus are more easily tolerated, and *vice versa*. The summed cost of violation of these constraints is

shown in (2). The actual F2 values of the consonant and the vowel are selected to minimize this cost function.

(2) $Cost = w_c(F2(C) - L)^2 + w_v(F2(V) - T)^2 + w_e(F2(C) - F2(V))^2$

This framework has been adopted in modeling the timing of H1 in the Seoul Korean Accentual Phrase in Cho (2007). Cho (2007) is the initial work that found the systematic deviation of H1 depending on speech rate. The experimental methods were similar to those in Chapter 2. Six native Seoul Korean speakers read fifteen four-syllable Accentual Phrases at fast, normal, and slow speech rates. Only the timing of H1 was analyzed, and it was assumed that the L was aligned at the beginning of the Accentual Phrase. The timing of H1 and its anchor (the middle of the second rime) had a linear relation. At the same time, H1 was delayed with regard to the anchor at fast speech rate, and anticipated in slow speech. In the linear regression model predicting H1 as a function of the anchor, the intercept was not zero, which means that timing of H1 is determined by the timing of the anchor plus some constant value. The constant is interpreted as the term including the 'rise time' target (the target duration of a rise) . Thus, the proposed model suggests that timing of H1 is determined by alignment and rise time target. The constraints are shown in (3).

(3)

	Constraint	Cost of violation
Alignment	$H = A$	$w_A(H - A)^2$
Rise time	$H = R$	$w_R(H - R)^2$

The alignment constraint $H = A$ requires that the H peak occur at the anchor. The violation of this constraint is calculated based on the differences between the timing of H1 and the anchor. That is, the cost of violation is the squared deviation from the anchor $(H–A)^2$, multiplied by the weight w_A. On the other hand, the rise time constraint $H = R$ requires that the peak occur at the end of the fixed rise time. Deviation from the target rise time is $(H–R)^2$. The cost of violation of this constraint is the squared deviation multiplied by the weight w_R, as also shown in (3). Given this, the cost of violations of a particular H timing is the summed cost of violations of each constraint given that timing, shown in (4).

(4) $Cost = w_A(H–A)^2 + w_R(H–R)^2$

The minimum of the cost function is found where its derivative is zero. By differentiating (4) with regard to H and setting the derivative as zero, (5) is obtained. This means that the timing of H is the weighted average of its anchor (A_H) and the target duration D. Because it was assumed that the rise starts from the phrase onset, the Rise time means the fixed interval from the phrase onset to the H.

(5) $T(H) = w_A \cdot A_H + w_R \cdot R$

This model is a special case of the model that will be proposed in the next Section. The present model in (3) has no constraint regarding timing of L, because it was assumed that timing of L was fixed at the phrase onset. That means, if there was a constraint for the alignment of L (i.e. a constraint requiring L to be at its anchor), that constraint was as-

sumed to be always satisfied. This assumption was approximately correct, given that the alignment of L1 was relatively stable, as shown in Chapter 2, but not precise. In Chapter 2, it has been shown that L1 was not strictly anchored, but systematically deviated from the anchor. That is, L1 occurred relatively earlier than the anchor when speaking fast, and later when speaking slower. In addition, the actual location of the anchor of L1 was not the phrase onset, but the middle of the rime in the first syllable. The model above needs to be elaborated further to accommodate the deviation pattern of L1 as well, which is the aim of the following Section.

3.2 Timing of L1 and H1

Using the weighted constraint model introduced in the previous section, a model of the timing of L1 and H1 in the Seoul Korean Accentual Phrase is developed in this section. The proposed model assumes that all the constraints that concern phonetic realization of F0 movements should be present in a language, and the phonetic realization pattern is explained by the weights of the constraints and targets. We propose the three constraints: alignment of H to A_H, alignment of L to A_L, and a target duration. Each of these factors is translated into a target in the model, where constraints require that these targets be realized. In this section, 'L' and 'H' indicates L1 and H1 for simplicity.

The proposed constraints and the cost of violation of each constraint are shown in (6). $T(L)$ is the timing of L, $T(H)$ is the timing of H. The Alignment constraint $T(L) = A_L$ requires that the L tone occur at the anchor. The cost of violating this constraint is the squared deviation of L timing from the anchor A_L, multiplied by its weight w_L. The same applies to $T(H)$ and A_H. At the same time, the Duration constraint requires that the duration of a rise corresponds to a *target duration*. Because the model consists of Alignment and Duration constraints, it will be referred to as the Alignment-Duration model (or the AD model).

(6) The Alignment-Duration model

	Constraint	Cost of violation
Align(L)	$T(L) = A_L$	$w_L(T(L) - A_L)^2$
Align(H)	$T(H) = A_H$	$w_H(T(H) - A_H)^2$
Duration	$T(H) - T(L) = D$	$w_D(T(H) - T(L) - D)^2$

where,
A_L: anchor for L1
A_H: anchor for H1
D: target duration, a positive constant
w_L, w_H, w_D: positive weights

The relative importance of these constraints is reflected in the weights of the constraints (w_L, w_H, w_D). Constraints with higher weights incur a greater cost of violation, and thus it is more crucial to satisfy them. For example, L was more strictly aligned than H in Korean, so it is expected that Align(L) will have a higher weight than Align(H) ($w_L > w_H$), i.e. alignment of L is more important than alignment of H. The actual timing of L and H is

determined as the values of $T(L)$ and $T(H)$ that minimize the summed cost of violations of these constraints, which is shown in (7).

(7) $$cost = w_L(T(L) - A_L)^2 + w_H(T(H) - A_H)^2 + w_D(T(H) - T(L) - D)^2$$

For illustration, assume that $A_L = 84ms$ (from the mean of the timing of 'rm1' (C1 to rm1) for speaker A1), $A_H = 263ms$ (from the mean of the timing of 'rm2' (C1 to rm2) for speaker A1), and target duration $D = 210ms$. In this example, the distance between A_L and A_H is smaller ($263 - 84 = 199ms$) than the rise duration target ($210ms$). Thus, assuming the weights are all equal, it is expected that the actual timing of L1 and H1 will deviate slightly from the Alignment targets, i.e. L should be a bit earlier than A_L, H should be a bit later than A_H. The evaluation of sample values is shown in (8). In (a), Align constraints are fully satisfied, incurring zero cost for Align(L) and Align(H). However, this incurs a high cost for Duration. In (c), on the other hand, the Duration constraint is fully satisfied ($278.5 - 68.5 = 210ms$), which results in high cost for Align(L) and Align(H). The optimal values for the timing of L1 and H1 are thus determined where the violation of each constraint is least overall so that the total cost is minimized, which is the case in (b).

(8)

	$T(L)$	$T(H)$	Align(L)	Align(H)	Duration	*total cost*
(a)	84	263	0	0	961	961
(b)	74	273	100	100	121	321
(c)	68.5	278.5	240.25	240.25	0	480.5

The minimum of the cost function (7) is found where its derivative is zero. By differentiating the cost function with regard to $T(L)$ and $T(H)$ respectively, we obtain the partial derivative functions for each. The equations in (9) are the expressions for $T(L)$ and $T(H)$ where the relevant partial derivatives are zero.

(9) a. $$T(L) = \frac{w_L}{w_L + w_D} A_L + \frac{w_D}{w_L + w_D}(T(H) - D)$$

b. $$T(H) = \frac{w_H}{w_H + w_D} A_H + \frac{w_D}{w_H + w_D}(T(L) + D)$$

The expressions in (9) correspond to the experimental results presented in Chapter 2, reflecting the effects of segmental anchoring and target duration. The two terms in each equation represent the effects of segmental anchoring and target duration respectively. This means that timing of a tone is determined by the weighted average of the anchor and the position that would yield rise duration of D. The terms with A_L and A_H reflect the effects of segmental anchoring. It was shown that tones have a linear relation with their anchoring points, and the equations in (9) imply linear relationships between $T(L)$ and A_L, and $T(H)$ and A_H respectively. The terms with D reflect the effects of target duration. According to the experimental results, tones deviate depending on speech rate. For example, H peaks occur relatively later than the anchor when speech rate is fast, and earlier when speech rate is slow. (9-b) means that the timing of H is determined by the anchor (A_H) and the fixed duration from the timing of L ($= T(L) + D$), with different relative importance. When

speech rate is fast, the absolute duration between $T(L)$ and A_H decreases. If the duration between $T(L)$ and A_H is shorter than $(T(L) + D)$, the timing of H will more likely exceed the timing of A_H in order to satisfy a target duration D. Similarly, if speech rate is slow, the absolute duration between $T(L)$ and A_H becomes longer, and if the duration between $T(L)$ and A_H is longer than $(T(L)+D)$, H will more likely occur earlier than A_H in order to satisfy a target duration D. Whether the actual timing ($T(L)$, $T(H)$) is closer to the anchor or to the point which satisfies D depends on the constraint weights, i.e. whether alignment is important or duration is important.

Furthermore, using the proposed model, it is possible to predict the L and H timing directly from the timing of A_L and A_H only, without having to know the timing of one of the tones. The direct solutions of optimization for $T(L)$ and $T(H)$ are obtained as in (10), by substituting (9-a) into (9-b) and *vice versa*.

(10) a. $T(L) = u_L(A_H - A_L - D) + A_L$ where $u_L = \dfrac{w_D w_H}{w_H w_L + w_H w_D + w_D w_L}$

b. $T(H) = -u_H(A_H - A_L - D) + A_H$ where $u_H = \dfrac{w_D w_L}{w_H w_L + w_H w_D + w_D w_L}$

The expressions in (10) mean that $T(L)$ and $T(H)$ deviate from their respective anchors (A_L, A_H) by a proportion (u_L, u_H) of the difference between the inter-anchor distance and the target duration D. The proportion (u_L, u_H) depends on the relative weights of the constraints, as (10) shows. Thus, for L, the more heavily weighted the Duration constraint and Align(H) are, the more the L tone deviates from its anchor. For H, the more heavily weighted the Duration constraint and Align(L) are, the more the H tone deviates from its anchor.

3.2.1 Estimating the precise anchor

In the equations in (9), unknown parameters are the three weights (w_L, w_H, w_D) and D. Mixed-effects models will be fitted to the data to obtain the estimates of these parameters. Prior to that, the anchoring points (A_L, A_H) are re-estimated. In Chapter 2, the best anchoring points for L1 and H1 were searched by testing several segmental points for the highest correlation with a tone. From that procedure, the anchor for L1 was the middle of the first rime, and the anchor for H1 was the middle of the second rime. However, the estimation in Chapter 2 was based on the Segmental Anchoring Hypothesis, where the L and H tones are assumed to be independently aligned. Thus, the anchor estimates may not be accurate for the model proposed in this chapter. In this section, the precise locations of the anchors are estimated using the proposed Alignment-Duration model in the following procedure.

The coefficients in (9) sum to 1. That is, for (9-a),

(11) $$\frac{w_L}{w_L + w_D} + \frac{w_D}{w_L + w_D} = \frac{w_L + w_D}{w_L + w_D} = 1$$

The same holds for (9-b). Thus, the expressions in (9) can be simplified by substituting the coefficients with simple variables, as in (12).

(12) a. $T(L) = aA_L + (1-a)T(H) + b$
where $a = \frac{w_L}{w_L + w_D}, b = -(1-a)D$

b. $T(H) = cA_H + (1-c)T(L) + d$
where $c = \frac{w_H}{w_H + w_D}, d = (1-c)D$

The expressions in (12) are further rearranged into (13), by grouping the terms with the same coefficient (a and c respectively). To estimate a precise anchor for L, it is hypothesized that the anchor A_L is somewhere in the first rime, because in the previous estimation, the best anchoring point was the middle of the first rime. Then A_L can be expressed as a point at some proportion into the first rime. A_L in (13-a) is replaced with $v1 + p \cdot rime1$, as in (14), where $v1$ is the beginning of the first vowel and $rime1$ is the duration of the first rime (that is, the distance from the beginning of v1 to the end of the first syllable). Then the anchor can be represented as some proportion (p) into the rime.

(13) a. $T(L) = a(A_L - T(H)) + b + T(H)$

b. $T(H) = c(A_H - T(L)) + d + T(L)$

(14) $T(L) = a(v1 + p \cdot rime1 - T(H)) + b + T(H)$

The value for p can be found by fitting a mixed model to the data. To fit an appropriate mixed model, (14) is rearranged into (15). A mixed model was fitted to the data with $T(L)$ as a dependent variable, $(v1 - T(H))$ and $rime1$ as fixed effects. The coefficient of the $T(H)$ term is known from the model in (13-a), which is 1. For the variables whose coefficient is 1, they are added as an 'offset' in a linear model. Thus the variable $T(H)$ was added to the mixed model as an offset. There were also by-speaker random intercepts and slopes for $(v1 - T(H))$ and $rime1$.

(15) $T(L) = a(v1 - T(H)) + a \cdot p \cdot rime1 + b + T(H)$

In the mixed model, the coefficient estimate of $(v1 - T(H))$ corresponds to a, and the coefficient of $rime1$ corresponds to $a \cdot p$. From a and $a \cdot p$, p can be computed. In the mixed model, $a = 0.833, a \cdot p = 0.247$, thus $p = 0.247/0.833 = 0.30$. So, the precise anchor for L is (16).

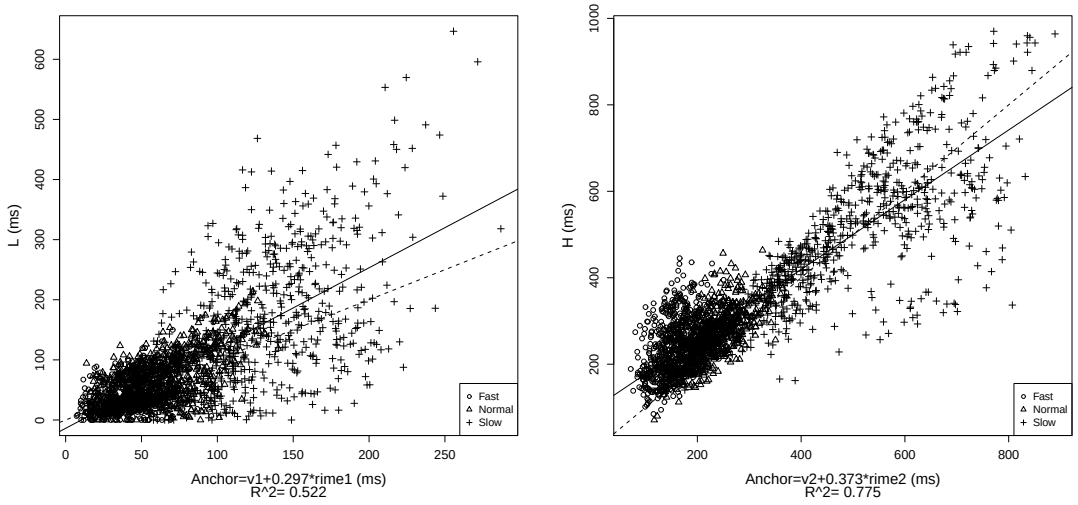
(16) $A_L = v1 + 0.30 \cdot rime1$

The same procedure is applied to $T(H)$. For $T(H)$, A_H is divided into $v2 + p \cdot rime2$ where $v2$ is the beginning of the second vowel and $rime2$ is the duration of the second rime. A mixed model was fitted to find the coefficients for $(v2 - T(H)$ and $rime2$, in (17), with $T(H)$ as a dependent variable, $(v2 - T(L))$ and $rime2$ as fixed effects, offset of $T(L)$, and by-speaker random intercepts and by-speaker random slopes for $(v2 - T(L))$ and $rime2$.

(17) $T(H) = a(v2 - T(H)) + a \cdot p \cdot rime2 + b + T(H)$

The coefficient for $(v2 - T(H)$ was 0.748, the coefficient for $rime2$ was 0.279, thus $p = 0.279/0.748 = 0.373$. Therefore, the precise anchor for H is estimated as in (18).

(18) $A_H = v2 + 0.37 \cdot rime2$



(a) Alignment of L1 (b) Alignment of H1

Figure 3-1: Alignment of L1 and H1 with precise estimates of the anchors (a) $Anchor = v1 + 0.297 * rime1$, (b) $Anchor = v1 + 0.373 * rime2$. The dashed line is $y = x$

In Figure 3-1, the timing of L1/H1 is plotted against their respective anchors, with the precise anchor estimates obtained in (16) and (18). The deviation plots were also redrawn with the precise anchor estimates, shown in Figure 3-2. Compared to the previous deviation plots with the first approximation of the anchor (Figure 2-20 in Section 2.4.1 in Chapter 2), L1 is found closer to the anchor. That is, the regression line did not cross the approximated location of the anchor (middle of the first rime), but it intersects with the precise estimate of the anchor (29.7% of the first rime).

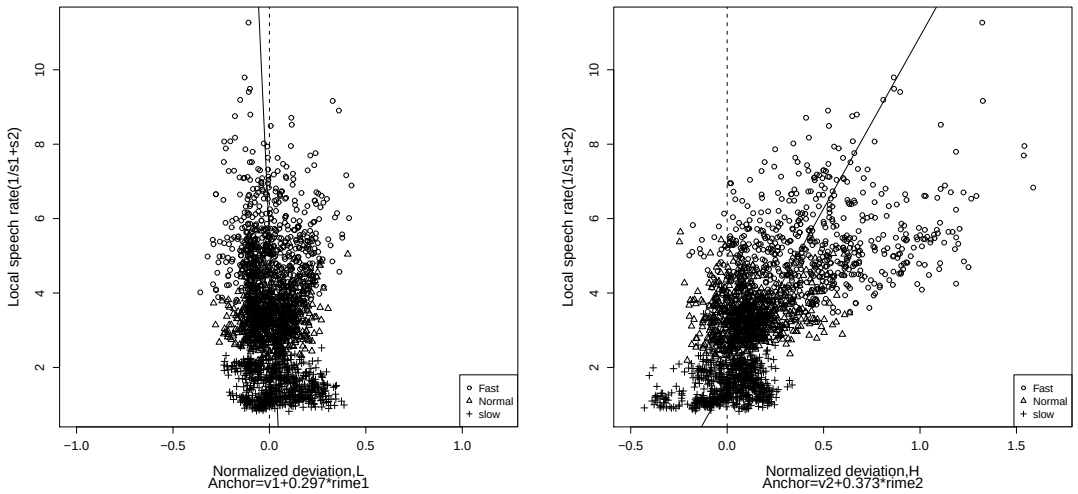
3.2.2 Estimating the constraint weights

Mixed effects models were used to obtain estimates of constraint weights. Constraint weights are calculated using corresponding coefficients in the mixed models fitted to the data. From (9) and (12) (repeated in (19) and (20) below), the relations in (21) hold.

(19) a. $T(L) = \frac{w_L}{w_L + w_D} A_L + \frac{w_D}{w_L + w_D}(T(H) - D)$

b. $T(H) = \frac{w_H}{w_H + w_D} A_H + \frac{w_D}{w_H + w_D}(T(L) + D)$

(20) a. $T(L) = aA_L + (1 - a)T(H) + b$



(a) Deviation of L1 (b) Deviation of H1

Figure 3-2: Deviation of L1 and H1 with precise estimates of the anchors (a) Deviation of L1 ($Anchor = v1 + 0.297 * rime1$), (b) Deviation of H1 ($Anchor = v1 + 0.373 * rime2$). The dashed line is the location of the anchor

b. $T(H) = cA_H + (1-c)T(L) + d$

(21) a. $a = \frac{w_L}{w_L + w_D}, 1 - a = \frac{w_D}{w_L + w_D}, b = -(1-a)D$

b. $c = \frac{w_H}{w_H + w_D}, 1 - c = \frac{w_D}{w_H + w_D}, d = (1-c)D$

(20) is rearranged into (22) (same as (13)), and mixed-effects models were fitted based on this to find a, b, c, and d.

(22) a. $T(L) = a(A_L - T(H)) + b + T(H)$

b. $T(H) = c(A_H - T(L)) + d + T(L)$

The values a and b in (22-a) are obtained by fitting a mixed model with $T(L)$ as a dependent variable, $(A_L - T(H))$ as a fixed effect, and by-speaker random intercepts and by-speaker random slopes for $(A_L - T(H))$, and offset $T(H)$. The precise anchor estimate in (16) was used for A_L. Then the coefficients of the mixed model correspond to the coefficients in the equation (22-a). In the mixed model, the coefficient a of $(A_L - T(H))$ was 0.825, the intercept b was -36.97. For c and d in (22-b), a mixed model was fitted with $T(H)$ as a dependent variable, $(A_H - T(L))$ as a fixed effect, and by-speaker random intercepts and by-speaker random slopes for $(A_H - T(L))$, and offset $T(L)$. The coefficient for $(A_H - T(L))$ was 0.755, the intercept was 92.31. In sum, the obtained values are as in (23).

(23) a. $a = \frac{w_L}{w_L + w_D} = 0.825, 1 - a = \frac{w_D}{w_L + w_D} = 0.175, b = -(1 - a) \cdot D = -36.97$

b. $c = \frac{w_H}{w_H + w_D} = 0.755, 1 - c = \frac{w_D}{w_H + w_D} = 0.245, d = (1 - c) \cdot D = 92.31$

In addition, $w_L + w_H + w_D = 1$ is assumed, because we are interested in the relative ratio of the constraint weights. With this assumption, from (23), we obtain $w_L = 0.54, w_H = 0.35, w_D = 0.11$. The relative weights of the constraints reflect the experimental results in Chapter 2. The weights suggest that the Align(L) constraint is more important than Align(H). In Seoul Korean, L1 was more strictly aligned than H1. The estimates of D, *duration target*, from (23-a) and (23-b) are not equal. From (23-a), $D = 211ms$; from (23-b), $D = 376ms$. However, as will be shown in Section 3.2.4, the two values are not statistically significantly different. The constraint weight for the duration target (w_D) is small (0.11), which makes it difficult to observe its effect. Thus, the D values have a wide confidence interval. In Section 3.2.4, we show that the confidence intervals of the two D values overlap, thus, the D values converge.

3.2.3 Model fitting from the actual solution of optimization

Furthermore, using the proposed model, it is possible to predict the L and H timing directly from the timing of A_L and A_H only, without having to know the timing of one of the tones. This direct solution of optimization of the proposed model is shown in (10), repeated in (24).

(24) a. $T(L) = u_L(A_H - A_L - D) + A_L$ where $u_L = \frac{w_D w_H}{w_H w_L + w_H w_D + w_D w_L}$

b. $T(H) = -u_H(A_H - A_L - D) + A_H$ where $u_H = \frac{w_D w_L}{w_H w_L + w_H w_D + w_D w_L}$

The coefficients u_L and u_H can be estimated by fitting mixed-effects models. For L, a mixed model was fitted to the data with $T(L)$ as a dependent variable, $(A_H - A_L)$ as a fixed effect, offset of A_L, and with by-speaker random intercepts and by-speaker random slopes for $(A_H - A_L)$. For H, a mixed model was fitted with $T(H)$ as a dependent variable, $(A_H - A_L)$ as a fixed effect, offset of A_H, and with by-speaker random intercepts and by-speaker random slopes for $(A_H - A_L)$. Precise anchor estimates were used for A_L and A_H. The coefficient of $(A_H - A_L)$ from the mixed model for $T(L)$ is u_L, the coefficient of $(A_H - A_L)$ from the mixed model for $T(H)$ is u_H. The D values are calculated from the intercepts in each mixed model. We found $u_L = 0.148, u_H = -0.225$. From these, the weights are calculated: $w_L = 0.53, w_H = 0.35, w_D = 0.12$, which is close to the results obtained in Section 3.2.2. The D value from the T(L) model was 142 ms, the D value from the T(H) model was 394 ms.

3.2.4 Estimating the duration target D

As mentioned at the end of Section 3.2.2, the D values from the model of $T(L)$ and the model of $T(H)$ were not the same. We refer to the D value from the model of $T(L)$ as D_L, the D

value from the model of $T(H)$ as D_H. The D values were calculated from the estimates of the intercept and slope in each model. That is, $D_L = b/(a-1)$ and $D_H = c/(1-d)$, where b and c are the estimates of the intercept, and a and d are the estimates of the slope in each model. Since these values are probabilistic estimations, the D values computed from these estimates are also probabilistic. That is, the D values also have a probabilistic distribution depending on the variances of the slope and intercept in the model. Thus, although the computation yielded the different D_L and D_H values, both D values are uncertain. If the computed D values have a high variance, the difference between D_L and D_H might not be statistically significant. However, the variance of D cannot be directly computed from the variances of the intercept and slope, because the calculation of D_L and D_H involves the ratio of two coefficient estimates of the regression model.

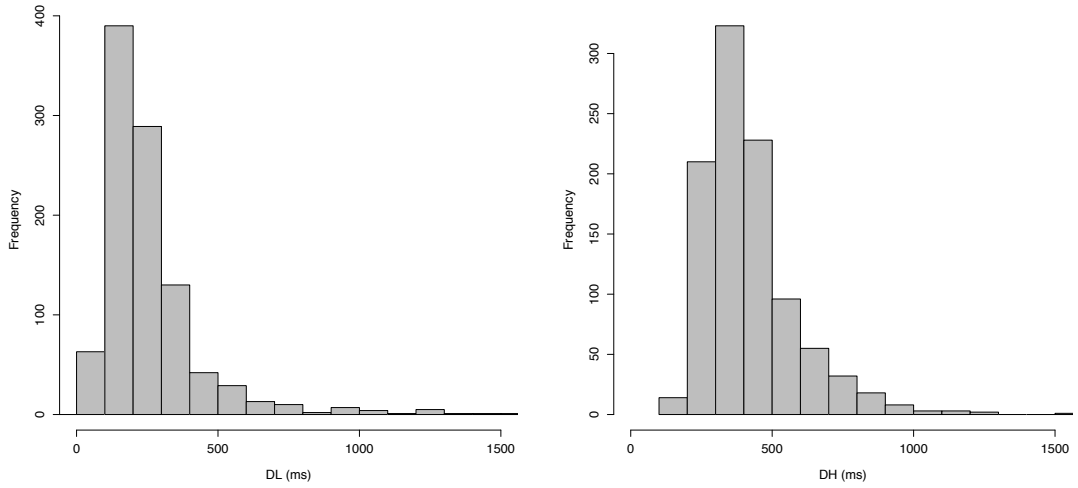
When the object of interest is more complicated than individual coefficients, simulation is the easiest and most reliable way to compute the uncertainty of the estimates (Gelman and Hill, 2007). Simulation is especially valuable in such cases and it is also applicable regardless of whether the distribution is normal or not. This applies to the D values: the D values are computed using two regression coefficients (the slope and intercept of the regression model), the ratio of the D values is likely to have a non-normal distribution. Simulations generate a large number of intercept and slope pairs that are drawn from the standard error distribution of the fitted model. The coefficients in the simulations are centered around the original model coefficients, with variation representing standard errors and covariance of coefficient estimates in the original fitted model. A set of D values are computed from these intercept-slope draws, from which the 95% confidence interval of D can be directly computed.

Prior to the simulation, the fixed effect variables [$(A_L - T(H))$ in the $T(L)$ model, $(A_H - T(L))$ in the $T(H)$ model] were "centered". That is, the mean of the variable was subtracted from the original variable so that its mean is zero. This procedure moves the intercept, but does not affect the slope. Centering yields a proper interpretation of the data when there is a higher-order interaction between terms, because centering reduces the correlation between the terms (Aiken and West, 1991); in our fitted models, the correlation between the intercept and the slope was rather high [0.907 in the $T(L)$ model, -0.664 in the $T(H)$ model], which suggests interactions between the intercept and the slope in the linear model. With centered data, the correlation between the intercept and slope was reduced after centering the variable: -0.519 in the $T(L)$ model, 0.535 in the $T(H)$ model.

Simulations were run to generate 1000 pairs of intercept and slope. 1000 is typically more than enough (Gelman and Hill, 2007). D values were computed for each simulation draw of intercept and slope. For D_L, $b/(a-1)$ values were computed, and for D_H, $c/(1-d)$ values were computed. The 95% of the simulated samples of D_L (i.e. the 95% confidence interval of D_L) was between 78 and 922 ms, and D_H was between 213 and 877 ms. The overlap between D_L and D_H was substantial. That is, the differences between pairs of simulated D_L and D_H were taken, and the 95% of the differences were between -448 and 706. The interval includes zero, which suggests that D_L and D_H are not significantly different.

Other estimation methods that assume a normal distribution (such as Taylor's approxi-

mation, Dieters et al. (1995)) are not appropriate because it turns out that D_L and D_H values are not normally distributed. The 1000 draws of pairs of D_L and D_H values have a skewed distribution, as shown in Figure 3-3. Therefore, the variances of D_L and D_H cannot be calculated from standard errors of the coefficients. Instead, simulations give more direct and accurate estimations of the probability distribution of D values.



(a) D_L (b) D_H

Figure 3-3: Distribution of D_L and D_H. Collected from 1000 draws from simulation.

From the distribution of D_L and D_H values, it is possible to estimate the most probable D value. The method is to find the peak of the combined probability distribution function of the two distributions (Papoulis, 2002). That is, the sampled distributions of D_L and D_H are first smoothed to estimate the probability density function (PDF). The range that will cover most of both distributions was selected. The range 0 to 1500 was selected, considering the range of distribution in Figure 3-3. The two PDF's in this range are multiplied, and the peak in the combined PDF corresponds to the location of the most probable D value. From this procedure, the D value is estimated as 270 ms. D values have a very wide confidence interval, which is the source of the uncertainty. This shows a difficulty in estimating the target value of a constraint which has a very low constraint weight (w_D=0.11).

3.2.5 Model comparison

In this section, the proposed Alignment-Duration model is compared with the model where the tonal timing is determined by its anchor only, not with a relation to the other tone. This is to examine whether the proposed model does better in predicting the timing of tones than the model where it is assumed that the L and H tones in a rising movement are independent of each other. For example, according to the Segmental Anchoring Hypothesis (Ladd et al., 1999), the beginning and end of a rising/falling pitch movement are independently aligned with regard to specific segmental positions. Tonal timing is thus predicted by segmental

alignment only, not by the relation with the other tone. On the other hand, according to the proposed Alignment-Duration model, the duration between the relevant tones is regulated by the grammar, as well as the segmental alignment of the individual tones.

According to the Segmental Anchoring Hypothesis, the timing of L and H are determined by their respective anchors. That is, such a model predicts (25), where $T(L)$ is the timing of L, $T(H)$ is the timing of H, and A_L and A_H are their respective segmental anchors.

(25) The Segmental Anchoring model ("SAH")

a. $T(L) = A_L$
b. $T(H) = A_H$

That is, the timing of L and H is determined strictly by the location of their anchors. However, (25) cannot hold considering the experimental results in Chapter 2. The experiments showed that the timing of L and H has a linear relation with their anchors, but the slope of the anchor in the regression model was not 1, and the intercept was not 0. That is, $T(L)$ and $T(H)$ have the linear relation with their respective anchors as shown in (26), where $a, c < 1$ and $b, d \neq 0$.

(26) The Independent-Alignment model ("IA")

a. $T(L) = a \cdot A_L + b$
b. $T(H) = c \cdot A_H + d$

Strictly speaking, this model is not the model with segmental anchoring only. In particular, the H model in (26-b) is what has been suggested for the timing of the Korean phrase-initial peak (Cho, 2007), which was discussed in (3) in Section 3.1. To recapitulate, the H model in (26-b) is interpreted as the H alignment model with a duration target from the phrase onset. That is, the slope c is the weight of the alignment constraint and the intercept d is the constant that contains the rise duration target (the duration from the phrase onset to H) and its constraint weight. That is,

(27) $T(H) = w_A \cdot A_H + w_D \cdot D$, where $c = w_A$ and $d = w_D \cdot D$

In this model of H timing, it was assumed that the rise starts from the phrase onset: that is, L is always strictly aligned at the phrase onset. The timing of the H tone is the timing that minimizes the deviations from the location of the anchor as well as the point after the fixed interval from the phrase onset. Note that for L, the same interpretation can be applied. That is, the model in (26-a) is equivalent to (28). The timing of L is the weighted average of its anchor (A_L) and a point at a fixed interval (k) from the phrase onset.

(28) $T(L) = w_A \cdot A_L + w_k \cdot k$, where $a = w_L$ and $b = w_k \cdot k$

This means that the model in (26) is not the model of segmental alignment only. The non-zero constant intercept represents the target duration which is the fixed interval from the phrase onset. In this sense, the model in (26) will be referred to as the Independent-Alignment (IA) model.

We compared the IA model in (26) to the proposed AD model, in order to test the con-

tribution of the Duration constraint. The difference between AD and IA is the presence of the target duration which explicitly refers to the difference between $(H - L)$. The AD model is repeated in (29). The L part of the AD model has the H term, and the H part of the AD model has the L term. Thus, L and H are not independent of each other.

(29) The Alignment-Duration model ("AD")

a. $T(L) = aA_L + (1 - a)T(H) + b$

b. $T(H) = cA_H + (1 - c)T(L) + d$

When fitting the Alignment-Duration model in the previous section, the precise anchor was estimated based on the model, as demonstrated in Section 3.2.1. The anchors were estimated for the IA model. The anchor for L is supposed to be in the first rime, and the anchor for H is supposed to be in the second rime. So, A_L is divided into $v1 + p \cdot rime1$, where $v1$ is the beginning of the first vowel, $rime1$ is the duration of the first rime, and p is the proportion into the first rime. A similar approach applies to A_H, which is assumed to be in the second rime. In this way, (26) is expressed as in (30).

(30) a. $T(L) = a(v1 + p \cdot rime1) + b = a \cdot v1 + a \cdot p \cdot rime1 + b$

b. $T(H) = c(v2 + p \cdot rime2) + d = c \cdot v2 + c \cdot p \cdot rime2 + d$

To find the p value for $T(L)$ in (30-a), a mixed model was fitted to the data with $T(L)$ as a dependent variable, $v1$ and $rime1$ as fixed effects, by-speaker random intercepts, and by-speaker random slopes for $v1$ and $rime1$. The random effects were tested with the same fixed effects, and this one was the best model. In this fitted model, the coefficient for v_1 was 1.004, the coefficient for $rime1$ was 0.447, so $a = 1.004, a \cdot p = 0.447$, and from here, $p = 0.45$. Thus, A_L is estimated as (31-a). A_H was estimated following a similar procedure.

(31) a. $A_L = v1 + 0.45 \cdot rime1$

b. $A_H = v2 + 0.23 \cdot rime2$

According to the coefficient and intercept estimates of the fitted models, $T(L)$ and $T(H)$ are expressed as in (32).

(32) a. $T(L) = 1.00A_L - 12.31$

b. $T(H) = 0.91A_H + 86.68$

To compare goodness of fit of IA and AD models, deviance values are used. Both the AD model and the IA model have the same number of parameters, so any reduction in deviance is a significant improvement. Table 3.1 summarizes the deviance values for each model. The deviance values for both $T(L)$ and $T(H)$ were lower in the AD model than the IA model. The differences between the deviances in the IA model and the AD model were significantly large. Therefore, the proposed Alignment-Duration model is better than the IA model in the case of the Seoul Korean phrase-initial rise. This means that H and L are not independent, and the duration constraint is making a significant contribution.

Table 3.1: Summary of deviance of models for Seoul Korean

	T(L)	T(H)
The IA model	21840	23231
The AD model	21679	22892
Difference	161	392

3.3 L2 undershoot

Another possible place where weighted constraints may apply is the relation between H1 deviation and L2 undershoot. Both H1 deviation and L2 undershoot change gradually depending on speech rate. In Section 2.5.2 in Chapter 2, it has been suggested that such changes reflect gradual compression of the tones, which eventually leads to truncation of the later rise. As time pressure increases, H1 deviates close to L2, and at the same time L2 level increases. The degree of undershoot is shown to be gradual as a function of local speech rate. On the other hand, H1 level was not affected by local speech rate. Thus, maintaining the H1 level is more important than alignment of H1. When time pressure increases, H1 deviates from its anchor and gets closer to L2. The alignment of L2 is shown to be relatively stable under changes in speech rate. Therefore, if the level of L2 is to be fully realized, the slope of the Fall from H1 to L2 should get steeper. Thus, we interpret the tendency to undershoot L2 under time pressure as a strategy to avoid a steep slope of the Fall when H1 gets too close to L2. It may be because it is too effortful or because there is a preferred slope.

These two possibilities suggest two different lines of analysis. It has been argued that if an AP is longer than four syllables, the syllables between H1 and L2 are underspecified for tones, so the transition from H1 to L2 is a linear interpolation between them (Lee and Kim, 1997). This suggests that the Fall may not have a slope target; it is entirely determined by the pitch and timing of H1 and L2. If there is no slope target, and the avoidance of a steep slope is explained by an effort constraint only, a slope that exceeds a certain threshold will be penalized, but a shallow slope will never be penalized. On the other hand, if a slope target exists for the Fall, deviations from the slope target in either direction will be penalized. Not just a steep slope, but also a shallow slope will be penalized. We do not have strong evidence suggesting which is the correct analysis; it can only be empirically determined by testing the two different models. In any case, it may be possible to model both analyses using a *slope target*; the difference between an effort constraint and a slope-target constraint is the penalized directions of violations: the effort constraint is one-way, the slope constraint is two-way.

Thus, the ideas above can be modeled in terms of weighted constraints, as shown in (33). Here we assume a slope target analysis rather than an effort constraint, for presentational simplicity.

(33)

	Constraint	Cost of violation
Ident(P_{H1})	$P(H1) = P_{H1}$	$w_{H1}(P(H1) - P_{H1})^2$
Ident(P_{L2})	$P(L2) = P_{L2}$	$w_{L2}(P(L2) - P_{L2})^2$
Slope	$S(F) = S_F$	$w_S(S(F) - S_F)^2$

where,
P_{H1}: target pitch for H1
P_{L2}: target pitch for L2
S_F: target slope for the Fall
$S(F) = \frac{P(H1) - P(L2)}{T(L2) - T(H1)}$ (Slope of the Fall)
w_{H1}, w_{L2}, w_S: positive weights

Ident(P_{H1}) requires that the F0 level of H1 must be the target pitch for H1, and Ident(P_{L2}) requires that the F0 level of L2 must be the target pitch for L2. The importance of these constraints are reflected in their weights, w_{H1} and w_{L2}. Considering that the level of H1 is more stable than the level of L2 under time pressure, we expect w_{H1} to be higher than w_{L2}. The level constraints conflict with the Slope constraint: in order to maintain the level of H1 and L2 when H1 and L2 are close in the temporal dimension, the slope of the Fall should increase. In order to maintain a slope of the fall when H1 and L2 are close to each other, the level of H1 and L2 should change. The importance of satisfying the Slope constraint can be found by the degree to which the slope is susceptible to time pressure. The slope term also includes the timing of H1 and L2. Considering the previous result that the level of L2 was affected by speech rate, we might expect the slope constraint to be more highly weighted than the L2 level target. The actual levels of H1 and L2 are determined by minimizing the violation of these constraints. Any deviations from the required pitch value incur violations of the constraints, and the cost of violation is the weighted sum of squared differences, as shown in (34).

(34) $$Cost = w_{H1}(P(H1) - P_{H1})^2 + w_{L2}(P(L2) - P_{L2})^2 + w_e(S(F) - S_F)^2$$

The level of the H1 and the level of L2 are determined as the values that minimize the cost function. In order to maintain a relatively constant slope, the ratio between $T(L2) - T(H1)$ and $P(H1) - P(L2)$ must be kept stable. This means that changes in $T(L2) - T(H1)$ must accompany changes in $P(H1) - P(L2)$. So, when $T(L2) - T(H1)$ decreases because of fast speech rate, $P(H1) - P(L2)$ must also decrease. This would result in undershoot of L2 at the fast speech rates.

In Chapter 2, it has been suggested that extreme compression of the LHLH intonational pattern leads to truncation of the later rise. We hypothesize that as time pressure increases, the level of L2 increases so that it reaches to the level of H1. At the same time, the timing of H1 gets closer to L2 so that it reaches to the timing of L2. The weighted- constraint model proposed here is expected to provide a tool to quantify this hypothesis. This L2 undershoot model will not be developed further quantitatively in this dissertation, partly because it involves a more complicated procedure, namely, differentiation of fractions and presumably a non-linear relation. It is a subject for future research.

The L2 undershoot model explains the phonological change of tonal sequences (LHLH to LH) as the results of the accumulated effects of gradual phonetic change. The model aims to provide a quantitative analysis of phonetic realization patterns of the LHLH AP from full scaling to undershoot up to the plateau pattern, i.e the point where $P(H1) = P(L2)$. We suggest that at this point, compression ends and truncation starts. A plateau is produced at this point, and tones are truncated after this point.

3.4 Summary of the chapter

In this chapter, a weighted-constraint model with alignment and duration constraints was developed, to account for the experimental results in Section 2.4.1 in Chapter 2. The beginning and ending points of a rising movement were aligned relative to segmental positions, reflecting the effects of segmental anchoring, but there was a systematic deviation of the tones from the anchor. That is, a rise starts earlier and terminates later when speaking faster. This systematic deviation pattern is analyzed as the effect of a constant duration target. The proposed Alignment-Duration model takes the Alignment constraint (Align(L) and Align(H)) and the Duration constraint as interacting terms. The constraint Align(L) requires L to be aligned with regard to its anchor; the constraint Align(H) requires H to be aligned with regard to its anchor; and the Duration constraint requires the distance between L and H to be constant. The conflicts among the constraints are resolved by constraint weighting, so highly-weighted constraints are less violated at a cost of greater violations of lower-weighted constraints. The actual timing of L and H tones is determined to minimize the cost of violation of these constraints. Neither segmental anchoring nor rise duration is invariant, but tonal timing is determined to satisfy all the constraints maximally.

On the other hand, according to the previous Segmental Anchoring Hypothesis, tonal timing is determined by segmental alignment only. The duration between L and H tones has not been considered as a factor that also regulates timing of tones. However, our experimental results show that the strict segmental anchoring model cannot hold. The Alignment-Duration model is compared with a model where the tones are independently aligned to their respective anchors ("the Independent-Alignment (IA) model"). It turned out that the Alignment-Duration model with the target duration between L and H has a lower deviance, so it is better than the model with independently-aligned tones. According to the Segmental Anchoring Hypothesis, it has been claimed that the tones are anchored to segmental landmarks, independently of each other. However, the model with the constraint on the duration between the two tones is significantly better than the model without the relation between tones.

The proposed model suggests that in general, tonal timing can be modeled as a compromise among alignment and shape targets. The constraint-based approach assumes that there are constraints common across languages, and cross-linguistic differences are reflected in the relative weights of constraints. Thus, it must apply across languages, not just Seoul Korean. We expect that the three constraints in the AD model will exist in other languages as well, and different languages will have different relative weights of constraints. To test this hypothesis, in the next chapter, the Alignment-Duration model is applied to

other languages, where the phonological status of tones are different from Seoul Korean.

Chapter 4

Cross-Linguistic Applications

In this chapter, the weighted-constraint model developed in Chapter 3 is applied to other languages with various tonal phonologies. The constraint-based approach implies that languages have common constraints, and cross-linguistic differences are parametric variations. Thus, the approach provides a common framework across languages, which facilitates cross-linguistic comparison. Cross-linguistic differences can be explicable through the relative weights of the constraints. More specifically, we test two hypotheses: that the same constraints (alignment and duration) are applicable cross-linguistically with variation in weights, and that cross-linguistic differences in tonal timing depend on the phonological status of tones in the language, which is reflected in the relative weights of alignment and duration constraints.

In Chapter 2, we have shown that in Seoul Korean, tendencies to both segmental alignment and target duration are observed in the timing of phrase-initial L and H. Both L and H were found to be centered around a segmental position (the anchor) but there were systematic deviations from the anchor depending on speech rate. H peaks tended to occur after their anchor as speech rate gets faster, and L troughs occurred earlier relative to their anchor as speech gets faster. In Chapter 3, the Alignment-Duration model was proposed to model these findings. The model explains the timing of initial L and H tones in the Seoul Korean Accentual Phrase as the interaction of alignment and duration constraints. The Alignment constraints (Align(L), Align(H)) require L and H tones to be aligned with regard to their respective segmental anchors, and the Duration constraint requires the duration between L and H tones to be constant. The actual timing of L and H tones is determined by the values that minimize the weighted sum of the cost of violation of these constraints. In other words, tonal timing is a weighted average of the alignment targets and the duration target, not just a function of the location of the segmental anchor. The constraint weights reflect the characteristics of the tonal alignment pattern in Seoul Korean. That is, Align(L) was more highly weighted than Align(H) ($w_L = 0.54, w_H = 0.35$), reflecting the experimental result that alignment of L was more strict than alignment of H.

In the present chapter, we examine rising F0 movements in other languages to test the generality of the weighted constraint model. We investigate whether the same constraints are applicable cross-linguistically with variation in weights. By examining constraint weights across languages, we test an additional hypothesis that relative weights of

constraints reflect the phonological nature of the relevant tones, that is, lexically-contrastive, prominence-lending, or phrasal boundary. We examine the tones in three languages: Tokyo Japanese, Mandarin Chinese, and English. In Tokyo Japanese, accented words have lexically-contrastive pitch accents, whereas unaccented words are characterized by phrase-initial boundary tones. Mandarin has lexically contrastive tones, and English has prominence-lending intonational pitch accents.

These languages have been chosen because the phonological status of tones in these languages is different from that of Seoul Korean as well as from one another. The tones in these languages are lexically contrastive (Mandarin, Japanese) or marking prominent syllables (English). Thus, the location of the tones is important, so these tones will show more strict alignment than boundary tones in Seoul Korean. In Seoul Korean, tones are not lexically specified but only mark the beginning and the end of a phrase. On the other hand, tones in the other languages are lexically specified, or align with a lexically determined location of prominence. We hypothesize that if tones are specified at the lexical level, the tone-syllable association is expected to be stronger because the location of tones are significant in distinguishing meanings or marking prominence of particular syllables. In Japanese, for example, the location of pitch accent is contrastive (e.g. íma (HL) "now", imá(LH) "living room"), so the tone-mora association is expected to be strong. In Mandarin, every syllable is specified with a lexical tone. In English, pitch accents mark the prominent syllable in the prominent word in the intonation contour. On the other hand, Seoul Korean does not have lexically specified tones, pitch accents, or stress. The Korean Accentual Phrase has one rise at the beginning and the other rise at the end of the phrase, if the phrase is longer than three syllables (Jun, 2000), and this intonational pattern is consistent regardless of the lexical items in the phrase.

The broad hypothesis that is tested in this chapter is that tones that are lexically contrastive or prominence marking will show less variability in alignment of tones than boundary tones as in Seoul Korean. Previous research supports this prediction: in Mandarin, most syllables are specified with contour tones, so misalignment would result in intrusion to the neighboring tones. Thus, tones are realized within or close to the syllables that the tones are phonologically associated with, while the shapes of the contour tones undergo changes through coarticulation with adjacent tones (Xu 1999). In English, lexical stress is realized as an intonational pitch accent when the associated word is prominent. The starred tone in a pitch accent (e.g. L+H*) is associated with the stressed syllable in the prominent word (Pierrehumbert 1980). The alignment of the peaks is shown to be stable regardless of speech rate (Ladd et al 1999). On the other hand, in Seoul Korean, individual words are not specified for pitch, but pitch signals only the beginning and end of the phrase. For such phrasal tones, the location of the H peak was relatively flexible. We have shown in Chapter 2 that the peak of the initial rise systematically deviates from the syllable with which it is phonologically associated.

This chapter presents the experimental results of three other languages with various tonal phonologies, and cross-linguistic comparisons based on the weighted-constraint models. Experiments were carried out to examine rising F0 movements in similar configurations in three languages: Tokyo Japanese (lexical pitch accent), Mandarin (lexical tone), and

English (intonational pitch accent). The experimental methods were comparable to Seoul Korean in Chapter 2. Speech materials were designed to contain a rising pitch movement at the phrase- initial positions. For Japanese, word-initial LH rising pitch movements are analyzed; for Mandarin, word-initial rising tones ("Tone 2") are analyzed; for English, word initial L+H* pitch accents are analyzed.

4.1 Lexical Pitch Accent: Tokyo Japanese

Japanese accented words are characterized by a pitch rise at the beginning of the word (unless the word is initial accented), and an accentual fall at the accented mora. Unaccented words have a word-initial pitch rise, but no accentual falls. According to the theory of Pierrehumbert and Beckman (1988), there is a L% boundary tone at the beginning of the utterance, a H phrasal tone on the second mora, and if the word is accented, the accentual HL on the accented mora. Pitch accents are contrastive in Japanese, e.g. íma (HL) 'now' vs. imá (LH) 'living room', and áme (HL) 'rain' vs. ame (LH) (unaccented) 'candy'. The Accentual Phrase has at most one pitch accent and it is the lowest prosodic unit in Japanese intonation (Pierrehumbert and Beckman, 1988). The prosodic level higher than the Accentual Phrase is the Intermediate Phrase, which consists of one or more Accentual Phrases. The Intermediate Phrase is the domain of the effects of what is known as catethesis. That is, within the Intermediate Phrase, each accentual peak is proportionally lower than the one before (Pierrehumbert and Beckman, 1988: 79).

In our experiment, we focus on the timing pattern of the tones in the pitch rise (LH) that occurs at the phrase-initial position. In the speech materials, the first mora has L and the second mora has H, so that they are comparable to the Seoul Korean experiment. We examine second-mora accented words and unaccented words. The initial-accented words are not the concern of this section (e.g. íma (HL) 'now'). Thus, the first mora has a phrasal boundary tone (L%) in both the accented and unaccented words; the H peak on the second mora is phrasal in the unaccented words, but it is accentual (HL) in the second-mora accented words. By examining the alignment patterns of the L and H tones in these words, we will show how the alignment pattern is different depending on the phonological status of the tones, i.e. accentual or phrasal. We also examine the accentual peaks in different phonological contexts (word-medial or word-final) and show how the alignment pattern is affected by the phonological context. The results are modeled in terms of weighted constraints, so the differences due to phonological status or context can be compared in a quantitatively-precise framework.

F0 movements in Japanese have been extensively studied in both quantitative and qualitative terms (Fujisaki, 1983; Poser, 1984; Pierrehumbert and Beckman, 1988; Kubozono, 1993; Ishihara, 2003). Most of the quantitative models focused on the levels and slopes of pitch movements. On the other hand, alignment of tonal targets with regard to the associated segmental string is relatively less studied. According to Ishihara (2006), the alignment of Japanese pitch accent is stable, confirming the Segmental Anchoring Hypothesis. However, we found that Japanese also shows systematic variation in the alignment of tones depending on speech rate, as was in the case in Seoul Korean. The systematic variations

may be explained under the proposed Alignment-Duration model, because the AD model implies that tonal timing is determined by the interaction of both alignment and duration targets, not by the segmental alignment only, and the variations are the consequences of the compromise between these two factors.

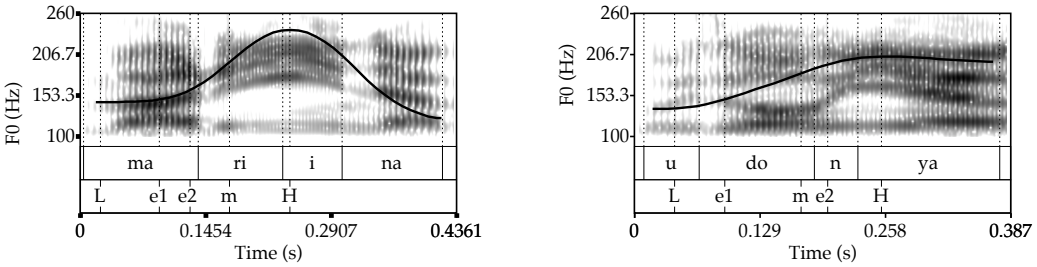
4.1.1 Experiment

Hypotheses

In Japanese, we examine the two hypotheses, segmental anchoring and target duration. In addition, it is expected that the alignment patterns of the tones will vary depending on phonological status (accented, unaccented) and context (word-medial, word-final). Lexically contrastive tones will show more strict alignment than tones that are not contrastive. That is, since accentual peaks in Japanese accented words are lexical, their alignment with regard to segments will be stricter than the alignment of the Seoul Korean initial rise peak, which only marks phrasal boundaries. Also, within Japanese, pitch peaks in accented and unaccented words will show different timing patterns. H peaks in accented words are contrastive, whereas those in unaccented words are boundary tones, so the peak alignment will be stricter in accented words.

Speech materials

The timing of L and H tones in word-initial accentual rises (LH) is analyzed in our experiment. The phrase-initial rise has been analyzed as a sequence of a boundary tone L% and phrasal H (Pierrehumbert and Beckman, 1988). This applies for unaccented words. For second-mora accented words, the H peak on the second mora is viewed as a phrasal H and an accentual HL on top of each other (Warner, 1997).



(a) Accented (b) unaccented

Figure 4-1: F0 movements in accented and unaccented words: (a) Accent on the second mora, [maríina] 'marina', (b) unaccented, [udonya] 'noodle shop' (Speaker: J6). 'L' and 'H' are F0 minima and maxima. 'e1' and 'e2' are the inflection points (lower and upper elbows). 'm' is the location of maximum velocity.

Examples from our data are shown in Figure 4-1. Figure 4-1a shows an example of a second-mora accented word: the H tone is a lexical pitch accent, immediately followed by

an accentual fall. Figure 4-1b shows an example of an unaccented word: the H peak is phrasal, and there is no accentual fall following the rise peak. It has been observed that pitch peaks in accented words are higher than pitch peaks in unaccented words (Pierrehumbert and Beckman, 1988: 46; Kubozono, 1993: 87-90). This was observed in our data as well. The mean of the F0 level of the H peak was 237 Hz for accented words, 224 Hz for unaccented words, the difference was statistically significant [$t(343) = 3.4, p < 0.001$]. In addition, the mean of the magnitude of the rise (the pitch change from L to H) was 61 Hz in accented words, 53 Hz in unaccented words, and the difference was statistically significant [$t(392) = 4.8, p < 0.001$].

Table 4.1: Japanese speech materials

	Accent location	Context of H	Examples	Tones
Medial-accented	2nd mora	Word-medial	[amádo] 'rain shutter'	LHL(L)
Final-accented	2nd mora	Word-final	[inú] 'dog'	LH
Unaccented		N/A	[udonya] 'noodle shop'	LHHH
Fillers	3rd syllable	Word-medial	[monomórai] 'sty'	LHHLL

Speech materials consisted of the four categories as shown in Table 4.1 (for the entire list, see A.2 in Appendix). These words were followed by a nominative particle '-ga' and a carrier phrase (e.g. [amadoga arimasu] 'There is a rain shutter.'). There were 15 words in the medial-accented group. In the medial-accented group, first syllables were light (one mora). Second syllables were light (one mora) in ten words, and heavy in five words (two morae; e.g. [yamámba] 'witch'). The H tone was on the second mora in the word, i.e. the first mora in the second syllable. Because the words in the medial-accented group were at least 3 syllables long and the pitch accent was on the second syllable, there was always at least one syllable between the H tone and the following particle: e.g. [yamámba-ga]. The final-accented group consisted of 7 bimoraic accented words, with the H tone on the second mora. Thus, for the final-accented group, pitch accent is immediately followed by the particle: e.g. [inú-ga] 'dog-Nom'.

The unaccented group consisted of five unaccented words with three to five morae. These were included to compare with pitch accents in accented words. The fillers were five words that had accent on the third syllable. For all groups, only sonorant sounds or voiced consonants were used in the syllables where the rising movements occur. The order of the words was randomized and rearranged to split more than two successive phrases with the same length in order to prevent abnormally rhythmic speech.

Methods and speakers

The speakers were six native speakers of Tokyo Japanese: four females (J1, J2, J3, J4) and two males (J5, J6), in their 20's$\sim$30's. The speakers were asked to read the speech materials with two repetitions, at normal, fast, and slow speech rates. We have learned from the experiment on Seoul Korean that it is difficult to elicit natural slow speech consistently in all speakers, because slow speech can vary much more than fast speech. Thus we tried to elicit slow speech closer to normal speech than to overly slow speech, whereas fast speech was

elicited as fast as possible. To avoid unnaturally slow speech, speakers were instructed to read slowly but still naturally, or "slower than the first (normal) reading". The recordings were made in the sound-attenuated recording booth at the phonetics lab in the MIT Linguistics Department. Other technical details were the same as the Seoul Korean experiment in Chapter 2.

The positions of F0 minima and F0 maxima were manually labeled, and were used as the location of the L and H tones. Between the F0 minima and F0 maxima, the inflection points (i.e. the lower and upper elbows) and maximum velocity points were automatically located using the same techniques used in the Seoul Korean experiment (Section 2.2.3 in Chapter 2). That is, the least-squares three-piece linear regression method was used to locate the two inflection points between the F0 extrema of the accentual rise. Spline fitting was used to locate the maximum velocity points. The shapes of rises were classified using the slopes of the three fitted lines.

4.1.2 Overall shape

In Chapter 2, the shape of rises was classified into sigmoid, scoop, and dome. The schematic illustration of these shapes is shown in Figure 4-2 (repeated from Chapter 2, Figure 2-6). The shape of a rise is determined by the slopes of the three regression lines. That is, the slope of the first line segment is the steepest in domed rises, the slope of the second line segment is the steepest in sigmoid rises, and the slope of the third line segment is the steepest in scooped rises. Linear shape cannot exist under this method, because the slopes of the three regression lines are never exactly the same.

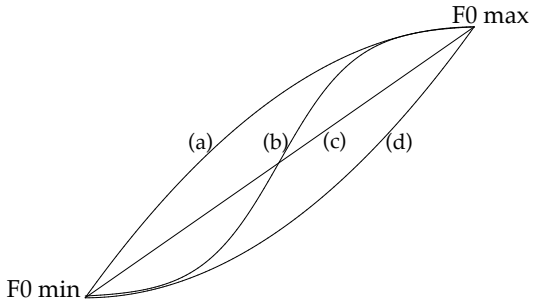


Figure 4-2: Schematic illustrations of the shape of rises: (a) dome (b) sigmoid (c) linear (d) scoop (adapted from Barnes et al. (2010))

The most common shape of the pitch rises in Japanese was sigmoid (85% of the data), as shown in Table 4.2. Domed shapes are found mostly in fast speech. This is expected because in domed shapes, the actual start of the fastest rise is at the beginning of the rising pitch movement, so they are more likely to occur under time pressure. Scooped shapes are mostly found in slow speech. This is also expected because in scooped shapes, the actual start of the rise comes very late, and the beginning of the actual rise can be delayed more easily when speaking slow. 'none' is where the three slopes of regression lines keep increasing, which often results from segmental perturbation during the rise. 'N/A' is the data where L and H points cannot be reliably measured due to perturbations in F0 contours.

Table 4.2: Shape of rises in Japanese

	sigmoid	dome	scoop	none	N/A
All	657 (85%)	38 (5%)	58 (8%)	9 (0%)	12 (2%)
Fast	216 (83%)	19 (7%)	15 (6%)	4 (2%)	6 (2%)
Normal	236 (91%)	7 (3%)	13 (5%)	0(0%)	3 (1%)
Slow	205 (80%)	12 (5%)	30 (12%)	5(2%)	3 (1%)

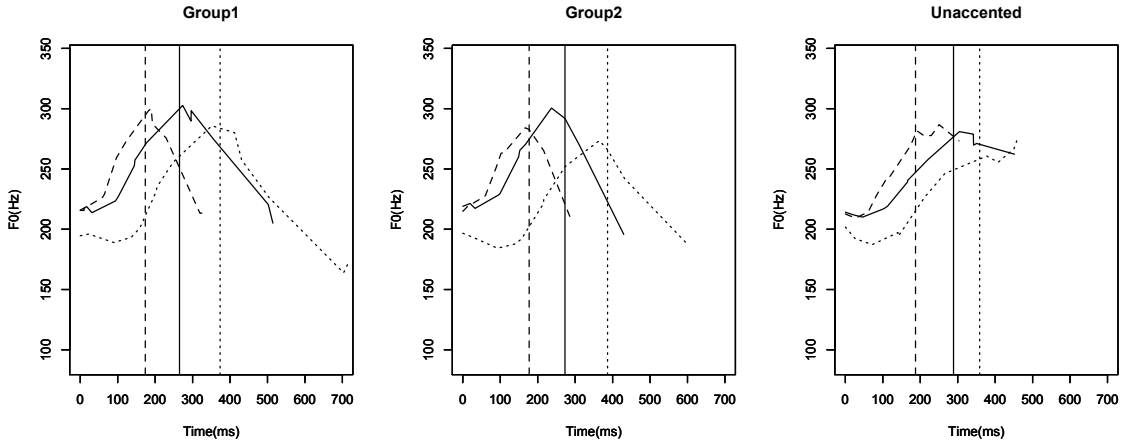


Figure 4-3: Averaged F0 curves: Solid line: normal speech, dashed line: fast speech, dotted line: slow speech, vertical lines: the end of the second mora where H is phonologically associated.

Figure 4-3 illustrates the averaged F0 curves including all speakers for medial-accented, final-accented, and unaccented words respectively. Alignment of the peak was quite stable across all speech rates. In the medial-accented group, the accentual peak in fast speech is not delayed, and the peak in slow speech is not much anticipated, either. In both medial-accented and final-accented groups, the F0 minimum seems to be delayed in slow speech, which may be evidence for a target duration or slope of the rise. The accentual peaks in normal and slow speech in the final-accented group show earlier alignment than those in the medial-accented group. The peak in the final-accented words is immediately adjacent to the word boundary. Thus, we may hypothesize that the presence of the morphological boundary pushed the timing of the accentual peak leftward in the final-accented words.

4.1.3 Effects of segmental anchoring

This section demonstrates the effects of segmental anchoring in the medial-accented words: the H peak on the second mora in the word-medial context. F0 minima and F0 maxima were used as the location of L and H tones. According to the Segmental Anchoring Hypothesis (Ladd et al., 1999), the beginning (L) and the end (H) of a rising pitch movement are stably aligned with regard to segmental landmarks ('anchors'). Thus, it is expected that there will be a positive linear correlation between a tone and its anchor. As a first approximation of the anchor location, we select the segmental landmark that has the highest correlation with

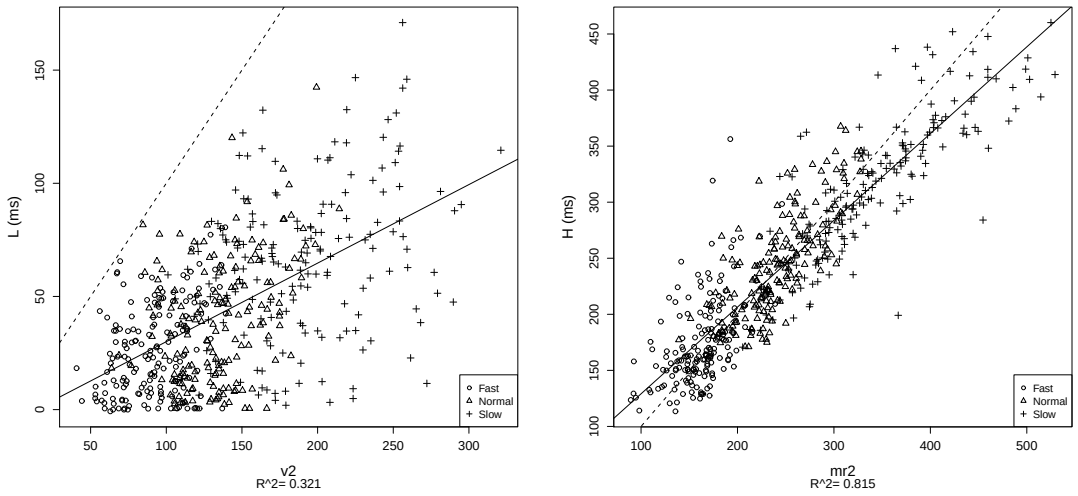
the timing of a tone, as we did in Seoul Korean in Chapter 2. Several candidate segmental landmarks were tested. For L, the segmental positions tested were 'v1' (the beginning of the first vowel), 'vm1' (middle of the first vowel), 'rm1' (middle of the first rime), 'o2' (onset of the second syllable),'c2' (onset of the second consonant), 'v2' (the beginning of the second vowel),'mr2' (the end of the second mora), and 'o3' (the beginning of the third syllable). A linear regression model was fitted for each point with the timing of L as a dependent variable, timing of a candidate anchor, speaker, and their interaction as predictor variables. As can be seen in the model specifications, the search is based on the correlation between the tone and a segmental landmark, without regard to the location of the other tone. The highest correlation between L and a segmental position was found with 'v2', the beginning of the second vowel ($R^2 = 0.49$). The point 'v2' remained as the best in a mixed-effects model with L as a dependent variable, timing of a candidate anchor as a fixed effect, by-speaker random intercepts and by-speaker random slopes for the timing of the candidate anchor ($G^2 = 4758$). Likewise, the best anchoring point for H was estimated. The highest correlation with H was found with 'mr2', the end of the second mora ($R^2 = 0.89$), which remains the best with a similar mixed model ($G^2 = 4930$).

In Figure 4-4, the L and H tones are plotted against their respective anchors, 'v2' and 'mr2', pooling across all speakers. The plots show the positive linear relations between the L or H tone and its anchor, which is expected under segmental anchoring. However, the slope of the regression lines was not 1: for H, slope$= 0.77, t(521) = 47.84, p < 0.001$, for L, slope$= 0.35, t(516) = 15.61, p < 0.001$ (testing difference from 1). This means that as the segmental anchor gets later, the tones do not get later as much as the anchor does. The dashed line in the plots are the $y = x$ line, i.e. the line that is expected if the tone is exactly at the anchor. Figure 4-4b shows the relation that as the anchor gets later, H occurs earlier than the anchor, and as anchor gets earlier, H occurs later than the anchor. In summary, a tendency to segmental anchoring is observed in Japanese, but segmental anchoring was not strict, so there should be other factors that explain the deviations of the tones from the anchor.

4.1.4 Effects of target duration

Along with the effects of segmental anchoring, an effect of a target duration for the rise was observed. Figure 4-5 shows the normalized deviation of L and H from their anchoring points. The data points show the location of the L or H tones relative to the anchor. The x-axis of the plot is the deviation of a tone from its anchor, normalized by local speech rate (the inverse of the duration of the first two syllables). The slope of the regression line was significantly different from zero; for H, slope$= 0.049$, $F(1, 521) = 187.2$, $p < 0.001$; for L, slope$=-0.028$, $F(1, 516) = 77.76$, $p < 0.001$. The plots show the trend that H peaks occur later with regard to the segmental landmark when speech rate is faster, and earlier when speech rate is slower. L troughs show the opposite pattern: they occur earlier when speech rate is faster, later when speech rate is slower.

The results in Figure 4-5 can be regarded as evidence for a duration target. That is, the accentual rises start earlier and terminate later at fast speech rates. This is to maintain a preferred duration of the rising movement under time pressure. These are the familiar



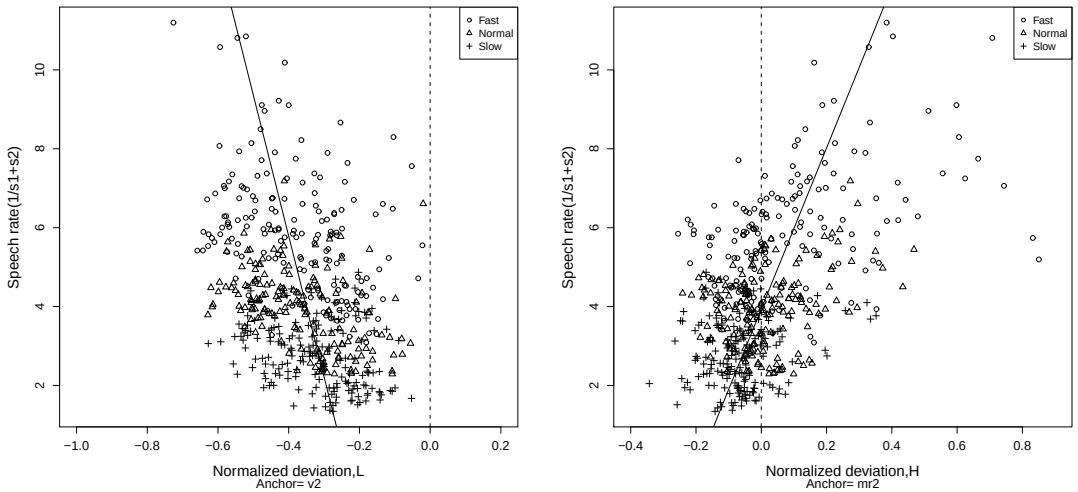
(a) Alignment of L (b) Alignment of H

Figure 4-4: Alignment of L and H in the medial-accented group: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$

deviation patterns which were also observed in the phrase-initial L and H tones in Seoul Korean (Section 2.4.1 in Chapter 2). In summary, in Japanese, both tendencies to segmental anchoring and target duration are found. That is, L and H tones are aligned with regard to a segmental landmark, but there is a systematic deviation of the tone from its anchoring point. This means that the timing of a tone is determined by both tendencies to maintain segmental alignment and a target duration. The similar tendencies in Seoul Korean were modeled in the Alignment-Duration model (Section 3.2 in Chapter 3). Thus, we hypothesize that the same model may be applied to Japanese, which is the subject of the next section.

4.1.5 The medial-accented group

The previous section shows tendencies for both segmental anchoring and target duration in Japanese pitch accent rises. Based on these findings, we apply the weighted-constraint model with the alignment and duration constraints. The hypothesis is that both L and H tones that comprise the accentual rise are aligned with regard to their respective anchoring points, and that there is also a constraint that requires the duration between L and H tones to be constant. Thus, there are three constraints: Align(L), Align(H), and Duration. The actual timing of L and H tones is determined as the values that minimize the weighted sum of the cost of violations of these constraints. Each constraint has to be violated to some degree in order to minimize the sum of weighted violations. This model was introduced in Section 3.2 in Chapter 3, referred to as the *Alignment-Duration* model. According to the model, the timing of L and the timing of H can be expressed as in (1). In (1), $T(L)$ and $T(H)$ are the timing of L and H tones, A_L and A_H are their anchors. a and c substitute relevant weight terms, as shown in (2). In (2), w_L is the weight of the constraint Align(L), w_H is the



(a) L deviation from A_L (b) H deviation from A_H

Figure 4-5: Alignment of L and H in the medial-accented group: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the location of the anchor

weight of the constraint Align(H), and w_D is the weight of the Duration constraint.

(1) a. $T(L) = aA_L + (1-a)(T(H) - D)$
b. $T(H) = cA_H + (1-c)(T(L) + D)$

(2) a. $a = \dfrac{w_L}{w_L + w_D}, 1 - a = \dfrac{w_D}{w_L + w_D}$
b. $c = \dfrac{w_H}{w_H + w_D}, 1 - c = \dfrac{w_D}{w_H + w_D}$

To review, the model in (1) implies that the timing of L is determined by its anchor, A_L, and the point which is at the fixed duration (D) from the timing of H. $T(L)$ is the weighted average of the alignment and duration targets. The timing of H is similar, i.e. $T(H)$ is the weighted average of the timing of A_L and the point which is at the fixed duration (D) from the timing of L. For model fitting, the model in (1) can be rearranged as in (3).

(3) a. $T(L) = a(A_L - T(H)) + b + T(H)$
where $b = -(1-a)D$

b. $T(H) = c(A_H - T(L)) + d + T(L)$
where $d = (1-c)D$

Before fitting the Alignment-Duration model, the anchor is estimated based on the model. To estimate the precise location of the anchor for L, we start with the assumption that A_L is within the first mora, because L is phonologically associated with the first mora.

A_L is expressed as a proportion into the first mora, so it is substituted with $(v1 + p \cdot mora1)$, where $v1$ is the beginning of the first vowel, $mora1$ is the duration of the first mora and p is the proportion into the first mora. Then (3-a) is expressed as (4),

(4) $T(L) = a(v1 + p \cdot mora1 - T(H)) + b + T(H)$

To find the coefficients a and b from mixed modeling, (4) is rearranged as (5).

(5) $T(L) = a(v1 - T(H)) + a \cdot p \cdot mora1 + b + T(H)$

Then, the proportion p can be calculated from the coefficient estimate of $mora1$ from the fitted mixed model. A mixed-effects model of the form shown in (5) was fitted to the data with $T(L)$ as a dependent variable, $(v1 - T(H))$ and $mora1$ as fixed effects, by-speaker random intercepts, and by-speaker random slopes for $mora1$. The by-speaker random slope for $mora1$ was significant in the model [$\chi^2(2) = 45.5, p < 0.001$]. Adding by-speaker random slopes for $(v1 - T(H))$ did not result in significant improvement in fit [$\chi^2(3) = 3.31, p = 0.34$]. The coefficient for $(v1 - T(H))$ was 0.835, the coefficient for $mora1$ was 0.158. The coefficient for $mora1$ was not significantly different from zero [t(518)=1.22, p=0.224]. If we take this best estimate of the coefficient, $a = 0.835, a \cdot p = 0.158$, and $p = 0.158/0.835 = 0.189$. Thus, A_L is expressed as (6-a).

(6) a. $A_L = v1 + 0.189 \cdot mora1$
b. $A_L = v1$

However, because the coefficient of $mora1$ is not significantly different from zero, (6-a) is not significantly better than $A_L = v1$, (6-b). It turned out that in unaccented words, the coefficient of $mora1$ was not significantly different from zero, either. Because we are working on the theory that the L tone is the same boundary L% tone in both accented and unaccented words, we also assume that A_L is the same in both cases. Thus, the anchor estimate of (6-b) will be used as the best estimate of A_L for both accented and unaccented words.

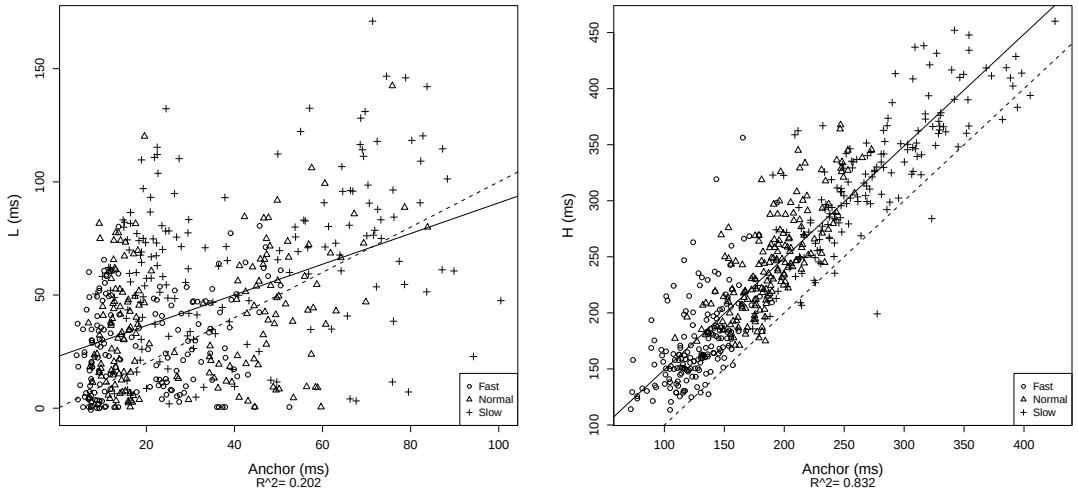
The precise value of A_H was estimated following a similar procedure. Assuming that A_H is within the second mora, A_H in (1-b) is substituted with $(v2 + p \cdot mora2)$, where $v2$ is the beginning of the second vowel, $mora2$ is the duration of the second mora, and p is the proportion into the second mora. From this, (7) holds.

(7) $T(H) = c(v2 - T(L)) + c \cdot p \cdot mora2 + d + T(L)$

A mixed-effects model was fitted to the data with $T(H)$ as a dependent variable, $(v2 - T(L))$ and $mora2$ as fixed effects, by-speaker random intercepts and by-speaker random slopes for $mora2$. Adding by-speaker random slopes for $(v2 - T(L))$ did not significantly improve the model [$\chi^2(3) = 1.02, p = 0.79$]. The coefficient for $(v2 - T(L))$ was 0.946, the coefficient for $mora2$ was 0.486, thus $c = 0.946, c \cdot p = 0.486$, so $p = 0.486/0.946 = 0.513$. Therefore, the precise estimate for A_H is as in (8).

(8) $A_H = v2 + 0.513 \cdot mora2$

The anchors in (6-b) and (8) are estimated using the Alignment-Duration model. Using these estimates, L and H were plotted against these anchor estimates, shown in Figure 4-6. Compared to Figure 4-4, the precise anchors are closer to L and H tones than the previous estimates of the anchor locations. The previous estimates of the anchor were based on a model where tonal timing is determined by the segmental anchor only, without the rise duration target. The segmental position that has the highest correlation with a tone was considered as the anchoring position. On the other hand, the anchors estimated here are based on a model that predicts timing of a tone as a function of segmental anchor as well as the timing of the other tone in the rising movement.



(a) Alignment of L

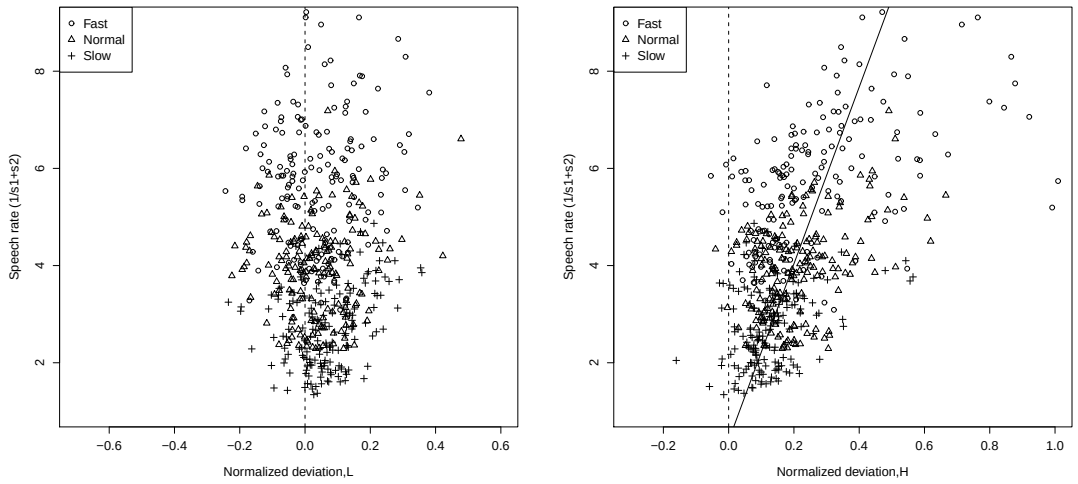
(b) Alignment of H

Figure 4-6: Alignment of L and H in the medial-accented group: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$. The anchor was the precise estimate of the anchor based on the AD model.

With the precise anchor estimates based on the AD model, mixed-models were fitted to the data to find the constraint weights, following the same procedure as in Seoul Korean described in Section 3.2.2 in Chapter 3. The computed constraint weights for Japanese are shown in (9). The relative weights suggest that alignment of both L and H tones are important, but H is more strictly aligned than L, and the Duration constraint is less important than the alignment constraints.

(9) The constraint weights for Japanese medial-accented words (tentative):
$w_L = 0.19, w_H = 0.76, w_D = 0.06$

The D values from the L model and the H model were different: $D_L = 108, D_H = 865$. Simulations were run to estimate the uncertainty of the D_L and D_H values. D_L and D_H values were computed from 1000 intercept and slope pairs drawn from the fitted $T(L)$ and $T(H)$ models. The 95% confidence interval for $(D_H - D_L)$ did not include zero, rang-



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-7: Deviation of L and H in the medial-accented group: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. The anchor was the precise estimate of the anchor based on the AD model.

ing from 398 to 1696, which means that D_L and D_H are significantly different. Because there is only one D value in the proposed model, this suggests a problem in applying the Alignment-Duration model to Japanese pitch accents as it is. We will diagnose the source of the problem by model comparison in the next section, and propose a possible solution in Section 4.1.7.

4.1.6 Model comparison

It turned out that the Alignment-Duration model has a problem in modeling the Japanese accentual pitch rise. The problem was that the D values computed from the fitted $T(L)$ and $T(H)$ models did not converge. To diagnose the problem, the Alignment-Duration model is compared with an alternative model. The simpler model is the model predicting the timing of a tone by the segmental alignment only, without consideration of the location of the other tone. The models in this section have been discussed in Section 3.2.5 in Chapter 3. In the Alignment-Duration model, timing of a tone 'T' is determined by both the alignment and duration targets, i.e. its anchor (A_T) and distance to the other end of the rise. On the other hand, the alternative model that is compared with the AD model is shown in (10). That is, the timing of a tone T ($= T(T)$) is determined by the location of its anchor, A_T.

(10) $\quad T(T) = aA_T + b$

Strictly speaking, (10) is not a model with segmental anchoring only. According to the Segmental Anchoring Hypothesis (Arvaniti et al., 1998; Ladd et al., 1999), the beginning and end of a rise are strictly anchored to segmental landmarks. That means that under the

SAH, a tone has to have a linear relation with the anchor with slope of 1 and intercept close to 0. That is,

(11) $T(T) = A_T$

However, from our experimental results, we have seen that (11) cannot hold. Rather, the L and H tones show a linear relation to their anchors of the form shown in (10), i.e. the slope (a) is not 1, and the intercept (b) is not 0. $T(T)$ is a function of A_T, but not equal to A_T.

In fact, the model in (10) is a special case of the Alignment-Duration model where duration is defined by a fixed interval from the phrase onset to the tone, instead of the duration between L and H tones. This needs more explanation. For the H tone, Cho (2007) found that the timing of the initial H peak in the Seoul Korean Accentual Phrase is in a linear relation with A_H, but with a slope less than 1 and a positive intercept. Thus, $T(H)$ is expressed as (12).

(12) $T(H) = aA_H + b$, where $0 < a < 1$ and $b > 0$

This model is also discussed in Section 3.2.5 in Chapter 3. In Chapter 3, the AD model supersedes this model motivated by the finding of dependency between $T(H)$ and $T(L)$. According to (12), $T(H)$ is a linear function of A_H, with a fixed intercept of b. The intercept corresponds to the 'rise time' target (the time from the phrase onset to the timing of the H peak), multiplied by the weight. In this model, the rise is assumed to start at the phrase onset, so L is always at the phrase onset. (12) is interpreted as (13). The slope of the linear regression is the weight of the alignment constraint ($H = A_H$), and the intercept includes the target duration of rise (R) and its weight (w_R). The expression suggests that the timing of H is a weighted average of the alignment target (A_H) and the rise duration target (R) where the duration is measured from the phrase onset, not from L.

(13) $T(H) = w_A \cdot A_H + w_R \cdot R$

Therefore, the model in (10) contains a target duration term, where duration is defined as the fixed interval from the phrase onset to a tone T, rather than the duration between L and H. Since the model in (10) actually contains a target duration constraint, it is different from the Segmental Anchoring Hypothesis. Thus, when we compare the Alignment-Duration model with the model in (10), we are not comparing the Alignment-Duration model with the Alignment-only model, but the model with alignment target and a fixed interval from the phrase onset.

We now demonstrate the results of the model comparison between the Alignment-Duration model and the model in (10), which is rewritten in (14). We refer to this model as the Independent-Alignment ("IA") model because the tones are aligned with regard to the anchor, independently of each other. In the IA model, the duration target is defined from the phrase onset, not with regard to the other tone.

(14) The Independent-Alignment model ("IA")

a. $T(L) = a \cdot A_L + b$

b. $T(H) = c \cdot A_H + d$

In order to estimate precise anchoring points based on this model, A_L and A_H are divided into ($v1 + p \cdot mora1$) and ($v2 + p \cdot mora2$) respectively, where $v1$ is the beginning of the first vowel, $mora1$ is the duration of the first mora, $v2$ is the beginning of the second vowel, $mora2$ is the duration of the second mora, and p is the proportion in the relevant mora.

(15) a. $T(L) = a(v1 + p \cdot mora1) + b = a \cdot v1 + a \cdot p \cdot mora1 + b$
b. $T(H) = c(v2 + p \cdot mora2) + d = c \cdot v2 + c \cdot p \cdot mora2 + d$

Mixed-effects models were fitted for $T(L)$ and $T(H)$ respectively to find the best estimates of the anchors. For $T(L)$, the dependent variable was $T(L)$, fixed effects were $v1$ and $mora1$, and random effects were by-speaker random intercepts and slopes for $mora1$. By-speaker random slopes for $v1$ were not significant [$\chi^2(3) = 1.44, p = 0.695$]. The same procedure was carried out for $T(H)$. From the fitted models, the p values in (15) were computed, and thus A_L and A_H were obtained as in (16).

(16) a. $A_L = v1 + 1.22 \cdot mora1$
b. $A_H = v2 + 0.51 \cdot mora2$

According to the coefficient and intercept estimates of the fitted models, $T(L)$ and $T(H)$ are expressed as in (17).

(17) a. $T(L) = 0.39A_L + 0.03$
b. $T(H) = 0.96A_H + 57.63$

The lower the deviance, the better the model. A summary of the deviance values is shown in Table 4.3. For L, the deviance was lower in the IA model (4772) than in the AD model (4858). For H, the deviance was lower in the AD model (4813) than in the IA model (4849). This suggests that timing of H is significantly affected by the distance from the preceding L, rather than the distance from the phrase onset, since the AD model was better than IA for H. On the other hand, for L, the IA model was better than the AD model. This suggests that the L tone is timed with respect to the phrase onset, more significantly than to the timing of the H tone. However, the AD model was better than the IA model for the H tone. That is, for L, (18-a) is a better model than (18-b). For H, the AD model (19-b) is still better than the IA model (19-a).

Table 4.3: Summary of deviance

	T(L)	T(H)
The IA model	**4772**	4849
The AD model	**4858**	4813

(18) a. $T(L) = aA_L + b$ (better)
b. $T(L) = aA_L + b(T(H) - D)$

(19) a. $T(H) = cA_H + d$

b. $T(H) = cA_H + d(T(L) + D)$ (better)

In summary, the timing of H is affected by the timing of L, but the timing of L is determined by the distance from the phrase onset, rather than distance to the following H. Predicting the timing of H does require information about the timing of L, but not *vice versa*. This may mean that speech planning is left to right, so the first tone is placed optimally according to its own preferences, but later tones are subsequently constrained by durations to preceding tones. The AD model is still appropriate because the Duration constraint refers to both L and H tones, but an additional constraint is necessary to explain why effects of the Duration constraint on L are not observed.

4.1.7 The constraint 'DelayL'

The model comparison in Section 4.1.6 shows that the timing of the L tone is significantly affected by a duration target, a fixed interval from a phrase onset, rather than by the timing of the following H tone. In the model comparison, it turned out that the model in (20) was a better model for L than the AD model.

(20) $T(L) = aA_L + b$

As explained, (20) implies that $T(L)$ is determined by the timing of A_L and a point at a fixed interval from the phrase onset. The fixed interval is referred to as the variable k. That is, the additional constraint requires L to occur at k, which is conflicting with Align(L) constraint. This constraint is referred to as DelayL, in the sense that the constraint prevents L from occurring at phrase onset, but requires it to be delayed into the first syllable/mora. The additional constraint DelayL explains why the Duration constraint has no detectable effect on L; it is because the Duration constraint is outweighed by DelayL. In summary, the AD model with the DelayL constraint is shown in (21).

(21) The Alignment-Duration model with DelayL

	Constraint	Cost of violation
Align(L)	$T(L) = A_L$	$w_L(T(L) - A_L)^2$
Align(H)	$T(H) = A_H$	$w_H(T(H) - A_H)^2$
Duration	$T(H) - T(L) = D$	$w_D(T(H) - T(L) - D)^2$
DelayL	$T(L) = k$	$w_k(T(L) - k)^2$

where,
A_L: anchor for L
A_H: anchor for H
D: target duration, a positive constant
k: a fixed interval from the phrase onset, a positive constant
w_L, w_H, w_D, w_k: positive weights

The actual timing of L and H is determined as the values that minimize the cost of violation of these constraints. The cost function is shown in (22).

(22) $cost = w_L(T(L)-A_L)^2+w_H(T(H)-A_H)^2+w_D(T(H)-T(L)-D)^2+w_k(T(L)-k)^2$

The minimum of this cost function is found where its derivative is zero. To find the constraint weights, the cost function is differentiated with regard to L and H respectively. Setting the partial derivatives equal to zero, the followings are obtained for $T(L)$ and $T(H)$ respectively.

(23) a. $T(L) = aA_L + bH + c$
where $a = \frac{w_L}{w_L + w_D + w_k}, b = \frac{w_D}{w_L + w_D + w_k}, c = \frac{w_k k - w_D D}{w_L + w_D + w_k}$
b. $T(H) = m(A_H - L) + T(L) + (1 - m)D$
where $m = \frac{w_H}{w_H + w_D}, n = (1 - m)D$

In addition, (24) holds.

(24) $1 - (a + b) = \frac{w_k}{w_L + w_D + w_k}$

Plugging (24) into c in (23-a), we derive $c = (1 - a - b) \cdot k - b \cdot D$

From (23) and (24), the constraint weights can be computed. To obtain the estimates of a, b, c, m, and n, mixed-effects models were fitted to the data for $T(L)$ and $T(H)$ in (23) separately. For $T(L)$, a mixed model was fitted with $T(L)$ as a dependent variable, A_L and $T(H)$ as fixed effects, by-speaker random intercepts and by-speaker random slopes for $T(H)$. The significance of the random effects were tested and this model was the best combination. From here, we obtained $a = 0.176, b = 0.253, c = -21, m = 0.929, n = 61$. With these, the constraint weights were computed as (25).

(25) The constraint weights for the medial-accented group:
$w_L = 0.04, w_k = 0.13, w_H = 0.77, w_D = 0.06$

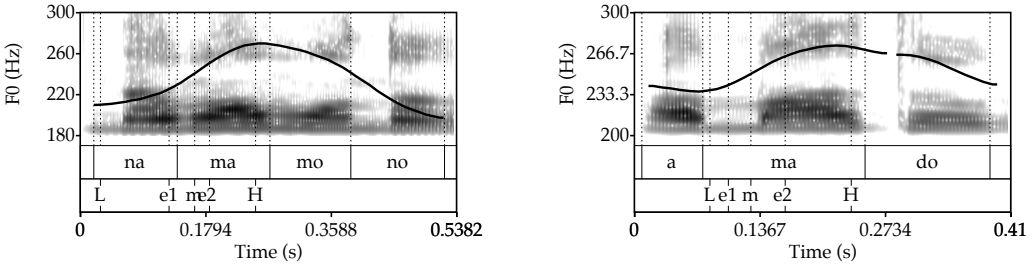
The weights suggest that the H peak is aligned to its segmental anchor more stably than the L troughs ($w_L < w_H$). However, the weight of Align(L) was very low, and lower (0.04) than that of DelayL (0.13). This suggests that the timing of L is more significantly affected by a fixed interval from the phrase onset than by the segmental anchoring point. In addition, w_k was greater than w_D. This confirms that timing of L is better explained in terms of a fixed target interval from the phrase onset rather than its relation to the H peak. Yet, timing of H is related to the timing of L: the weight of w_D was small but not zero. In the Segmental Anchoring Hypothesis literature, it has been claimed that L and H are independently aligned, but given this result, we cannot say that L and H are completely independent, because L was significant in predicting timing of H.

Estimating k and D values from this model is less certain. In the previous Alignment-Duration model for Seoul Korean (without the DelayL constraint), the D value was estimated from the probability distributions of D_L and D_H. To review, we have referred to the D value computed from the L model as D_L, and the D value computed from the H model as D_H. In Seoul Korean, D_L and D_H values had large confidence intervals because

of the large variance in the data. Also, it is difficult to estimate a value that has small effects due to the low constraint weight. The D value was approximated as the peak of the combined PDF (probability density function) of the D_L and D_H distributions. However, in the model developed in this section, D_L cannot be computed because there is another unknown variable k in the $T(L)$ equation of the model. That is, c in (23-a) contains two unknowns: k and D ($= D_L$). Yet, D_H can be computed from (23-b): since $m = 0.93, n = 61$, $D = 61/(1 - 0.93) = 871$. However, this D estimate seems too large to be correct. The confidence interval for D_H was also large: $503 \sim 2016$. If we take the smallest D_H (503), we obtain the smallest estimate of k based on the H model alone, which is 186. However, if we had known the PDF of D_L, and D were estimated from the combined PDFs, the k values would have been different, although the effects of D_L would be small given the low weight of w_D. In any case, we cannot be certain about the D value and we also cannot be certain about the value of k.

The uncertainty of the D value arises from the difficulty in observing the effects of a constraint which has a very low weight ($w_D = 0.06$). We may also try using the mean of the $(H - L)$ value, as a plausible estimate of the value of D. In the model, D is the duration from L to H, so the $(H - L)$ value is the value that is expected if the Duration target is fully satisfied. Given that w_D is low, it is not an accurate estimate, but we may have a sense of what may be the likely value of D. The mean of $(H - L)$ was 201. When this is D, k is 52.

Although the k value is uncertain due to the uncertainty of D and the low constraint weight w_k, we may consider the interpretation of the k value. From observation of the data, the location of L minima can vary, scattered across the whole first mora. Looking at the F0 trajectories, the pitch during the first mora is not always increasing, but it can vary: increasing (Figure 4-8a), decreasing (Figure 4-8b), or level. The F0 minimum is thus found at the phrase onset, at the end of the first mora, or anywhere in between.



(a) F0 minimum at the phrase onset

(b) F0 minimum at the end of the first mora

Figure 4-8: Variations of L location in Japanese: (a) [namamono] 'a living thing', (b) [amado] 'a rain-shutter' (Speaker: J4). Both words are second-mora accented. The location of L minimum varies: L is at the phrase onset in (a), but at the end of the accented mora in (b).

Pitch contours can be perturbed due to segmental effects ('microintonation', 't Hart et al. (1990); Lehiste and Peterson (1961)), which temporarily raises the F0 at the phrase onset if the onset consonant is an obstruent. This creates a dip in the F0 contour in the following vowel, which is measured as the F0 minimum. Thus, some of the variation may be due to

segmental effects. However, the later F0 minima are not always/necessarily due to segmental effects. Figure 4-8b shows a case where the F0 minimum is found at the end of the first mora, but it is not likely to be a segmental effect, because the word is vowel-initial.

Thus, the timing of L is less regulated than the timing of H; L involves more randomness. This explains the reason why the coefficient of $mora1$ (duration of the first mora) was not significant when estimating A_L in terms of a proportion into the rime in the first mora. That is, A_L cannot be meaningfully expressed in terms of the proportion into the first rime because L can appear at random anywhere in the first rime.

The AD model with the addition of the DelayL constraint predicts that L should occur at a particular point between A_L and k. With k, the alignment target for L is determined as the location between A_L and k, that is, the middle of the range of random variation. The random variation around this alignment point shows up as relatively high error in predicting L. Thus, it is expected that k should be the other end of the possible range of random variation. Given that L ranges from the phrase onset to the end of first mora, and that A_L is the beginning of the vowel (i.e. close to the phrase onset), k is expected to be close to the end of the first mora. Our best estimate of k's value was 52 (based on the $(H - L)$) or 186 (based on the $T(H)$ model alone), although it is rather uncertain. Given that the constraint weight is too small to obtain very precise estimates of the values of either D or k, it can be said that the two k estimates are found somewhere targeting the end of the first mora (=94 ms). This may reflect the fact that F0 minimum can be delayed as far as the end of the first mora.

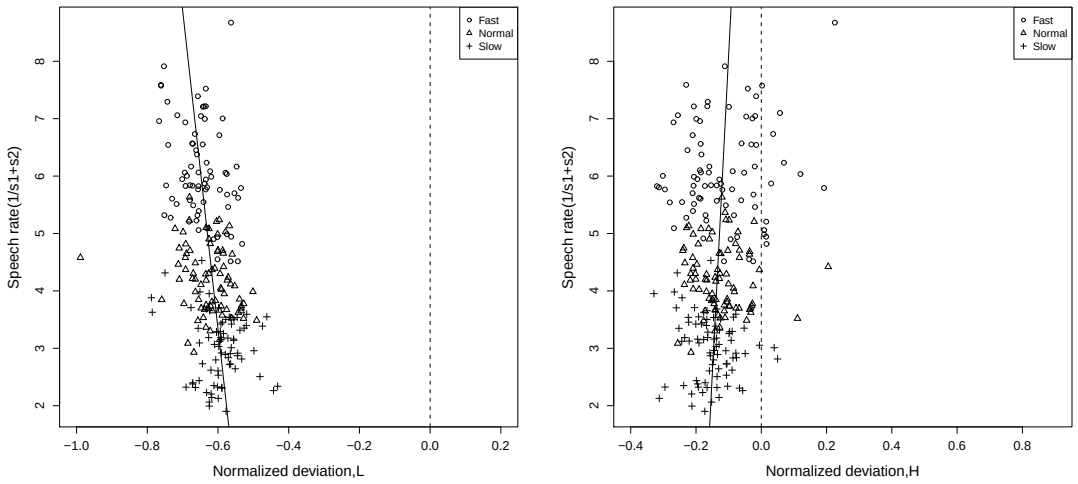
4.1.8 The final-accented group

This section will show differences in alignment patterns that depend on phonological context, based on an analysis of the experimental results of the final-accented words. The final-accented group consists of bimoraic words, accent on the second mora, immediately followed by a nominative particle '-ga'. The difference between medial accented and final accented is the phonological context of the accentual peak: word-medial (26-a) or word-final (26-b).

(26) a. amádo-ga arimasu.
LHL-L
rain shutter-NOM exists.

b. inú-ga arimau.
LH-L
dog-NOM exists.

As a first approximation, several segmental landmarks were tested for the best correlation with the L tone. The best correlation was found with the end of the second mora ($R^2 = 0.45$). In Figure 4-9a, the normalized deviations of L tones from this anchor are plotted against local speech rate. The dashed line is the end of the second mora. There was a small negative trend between L deviation and speech rate, that is, the slower the rate the later the L.



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-9: Deviations of L and H in final-accented words: (a) L deviation from the segmental landmark that has the highest correlation, (b) H deviation from the end of the second mora.

For H, the segmental landmark that has the highest correlation was the middle of the second rime ($R^2 = 0.962$). In the normalized deviation plot, however, the end of the second mora is used as the anchoring point. The correlation of this point was almost the same as the middle of the second rime ($R^2 = 0.959$), and it reveals an interesting generalization that H peaks tend not to cross the end of the accented mora when that mora is at a word-boundary. For this reason, Figure 4-9b shows the deviation of the H peak, with regard to the end of the second mora. The dashed line corresponds to the end of the second mora. The F0 peaks tend not to occur after the end of the accented mora when the mora is word-final. This is different from the word-medial context: in the word-medial context, peaks show a gradual and systematic delay relative to the accented mora (cf. Figure 4-5b). On the other hand, the accentual peaks in the word-final context do not deviate as much as in the word-medial context, in order to avoid crossing the upcoming word boundary.

This phenomenon is not equivalent to the well-known peak retraction in phrase-final position common in many languages (Kinyarwanda, Myers (2003); English, Silverman and Pierrehumbert (1990); Spanish, Prieto et al. (1995)). In many languages, in phrase-final position, H peaks tend to retract because of pressure from the following boundary tone. In the Japanese case, the H tone is adjacent to the upcoming word boundary, but not adjacent to a boundary tone. One might think that the accentual peak in the final-accented words is still closer to the boundary tone, although not immediately adjacent, so the accentual peak is retracted due to the proximity to the boundary tone. However, in the experimental results in this dissertation so far, the affects of an adjacent tone seem to surface as gradual deviations, rather than an abrupt break as seen in Figure 4-9b. For example, the phrase-final peak H2 in Seoul Korean shows gradual deviations from the anchor (Figure 2-31b in

Chapter 2), due to the pressure from the following L tone. The usual phrase-final peak retraction would predict a retracted anchor, but not stricter alignment, thus we may expect to see a similar pattern of deviation to that observed in the medial-accented group, i.e. gradual deviations from a retracted anchor. On the other hand, the alignment pattern of final accented is stricter than that of medial accented, so we suggest that an additional constraint is forcing the peak to stay within the word.

In Japanese, peaks are delayed to a greater degree if the accentual peak is word-medial (cf. Figure 4-5b). That is, peak delay is allowed within words, but not across a word boundary. In previous studies, peak delay is considered to be a natural consequence of physical implementation of tones (Xu, 2001), but our results suggest that peak delay may not be entirely due to automatic physiological effects, because speakers can choose not to delay. That is, the H peaks in the word-final position have a tighter alignment than in the word-medial position. This shows that peak delay is a controllable linguistic factor. If alignment is controllable, it is in principle subject to language-specific manipulation (Keating, 1985: 120). For example, alignment of the accentual peak in word-final positions in Japanese is more strict than the alignment of the phrasal peak in Korean. Also, Mandarin shows a strict alignment pattern, as will be shown shortly in Section 4.2, which is different from Korean phrasal peaks or Japanese word-medial accentual peaks. These findings suggest that tonal alignment is a controllable linguistic factor, which gives rise to language-specific variation.

Model fitting

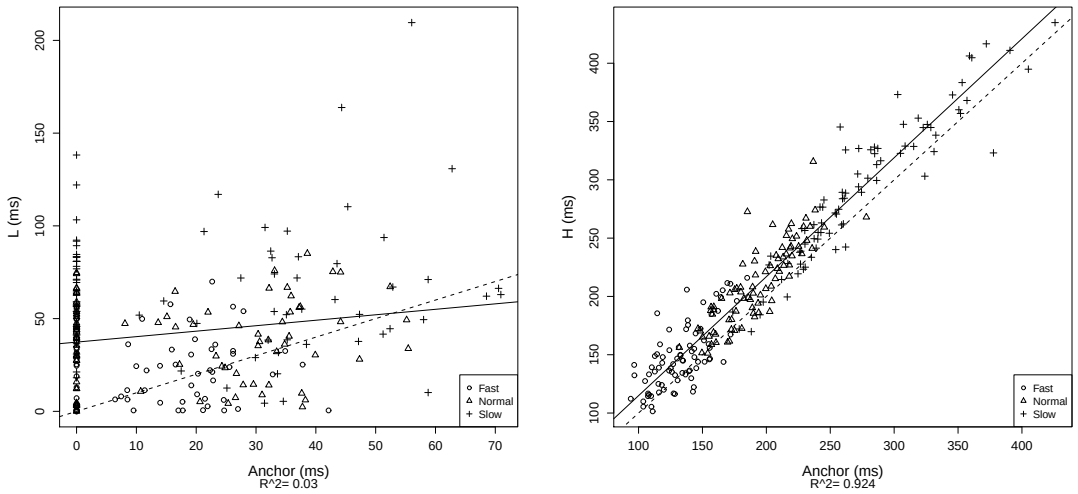
It is assumed that the anchors for the final-accented group are the same as those for medial accented. That is, the anchor estimates for the medial-accented group (in (6-b) and (8)) were used for the final-accented group, repeated in (27). Such a model implies that what is changed according to the phonological context is the constraints or constraint weights, rather than the location of the anchor.

(27) a. $A_L = v1$
b. $A_H = v2 + 0.513 \cdot mora2$

In Figure 4-10, the timing of L and H tones are plotted against these anchor estimates. From the plot for L in Figure 4-10a, we can see that even when the anchor (v1) is at 0 (vowel-initial words), L varies widely from 0 to about 140 ms. This means that the L may be somewhat independent of the location of the anchor. This again supports the idea that there is a factor determining the location of the L that makes the L tend to occur at a fixed interval from the phrase onset, along with other factors. On the other hand, the H tone stays close to the anchor, as shown in Figure 4-10b.

Figure 4-11 shows the normalized deviation plots for L and H tones in the final-accented group, based on the anchor estimates in (27). Both L and H are close to their estimated anchors. In particular, the degree to which H deviates is much smaller than the H tone deviation in the medial-accented group.

The same constraints are assumed for both medial-accented and final-accented groups, which is the Alignment-Duration model with the DelayL constraint. The procedure is the



(a) Alignment of L (b) Alignment of H

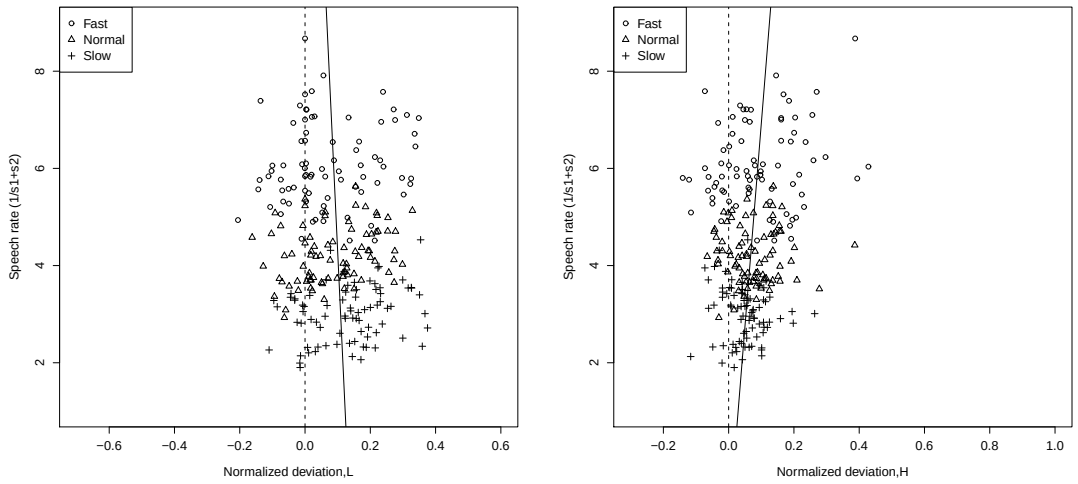
Figure 4-10: Alignment of L and H in the final-accented group: (a)L against A_L, (b) H against A_H, the dashed line is $y = x$. Precise anchor estimates based on the Alignment-Duration model.

same as before. To find the corresponding coefficients, mixed-effects models were fitted to the data for $T(L)$ and $T(H)$ respectively. Using the same model as in the case of the medial-accented group, $a = -0.189, b = 0.285, c = -16, m = 0.99, n = 0.67$. Among these, a was not significantly different from zero [$t(243) = -1.98, p = 0.048$], so we use $a = 0$ instead of the negative value in order to avoid a negative weight. With these values, the constraint weights were computed as in (28).

(28) The constraint weights for the final-accented group:
$w_L = 0, w_k = 0.03, w_H = 0.96, w_D = 0.01$

The constraint weights indicate that the Align(H) constraint is so strong that the effects of other constraints are almost negligible. The weight of Align(H) was almost 1, and the weights of other constraints are almost 0. The weight w_H is calculated from the slope (m) of the H model. That is, $m : (1 - m) = w_H : w_D$. The slope of the H model was 0.99, but it was not significantly different from 1 [$t(243) = -0.05, p = 0.96$]. If we take $m = 1$, the constraint weights are effectively $w_H = 1$ and all others 0. D_H is not calculable, because m is 1 or very close to 1, the intercept has to be divided by zero or a number close to zero (because $D_H = n/(1 - m)$), resulting in infinity (∞).

Under the grammar we have assumed, the anchors are the same for both word-medial and final positions. The computed constraint weights are different: w_H is higher in the final-accented group (0.96) than in the medial-accented group (0.77). However, there is another alternative grammar model that can be suggested. That is, the difference between the two groups is not the constraint weight, but the effect of the word boundary. The resulting difference in the alignment is due to another constraint that reflects the effect of



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-11: Deviation of L and H in the final-accented group: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. Precise anchor estimates.

phonological context. The constraint is referred to as "Boundary". The difference between medial-accented and final-accented groups is whether the peak is word-medial or word-final. From the pattern we observed, we may hypothesize that peak delay within a word is tolerable, but delay across a word boundary incurs a greater cost of violation. This can be formulated in terms of the weighted constraint model in (29).

(29)

	Constraint	Cost of violation
Align(L)	$T(L) = A_L$	$w_L(T(L) - A_L)^2$
Align(H)	$T(H) = A_H$	$w_H(T(H) - A_H)^2$
Duration	$T(H) - T(L) = D$	$w_D(T(H) - T(L) - D)^2$
DelayL	$T(L) = k$	$w_k(T(L) - k)^2$
Boundary	$T(H) < B$	0
	$T(H) > B$	$w_B(B - T(H))^2$

where,
A_L: anchor for L
A_H: anchor for H
D: target duration, a positive constant
k: a fixed interval from the phrase onset, a positive constant
B: timing of a word boundary
w_L, w_H, w_D, w_k, w_B: positive weights

In (29), the Boundary constraint is added to the Alignment-Duration model with DelayL. The cost of violation of the Boundary constraint is 0 when H is before the word boundary (i.e. $T(H) < B$). Otherwise, any deviation from the boundary incurs a cost of violation.

This constraint makes H peaks stay within the word demarcated by the word boundary. As shown in Figure 4-9, H peaks tend not to cross over the end of an accented mora which is at a word boundary. On the other hand, when the end of the accented mora is word-medial (which is the case in the medial-accented group), H peaks may cross over it, as shown in Figure 4-5b. The addition of the Boundary constraint models these observations, because crossing the word boundary incurs a cost of violation whereas not crossing it costs nothing, so it is less likely that the actual H peaks occur after the word boundary,

The weight of the Boundary constraint (w_B) is calculated based on the constraint weights computed for medial-accented and final-accented groups separately. The weights of other constraints were also adjusted relative to w_B. The result is shown in Table 4.4. Because w_B is very high, other weights became very small due to the assumption that the weights sum to 1, e.g. w_H values, 0.77 and 0.96, became 0.18. However, the relative ratios between weights are maintained approximately, e.g. the adjusted weights are about a fourth of those for the medial-accented condition.

Table 4.4: Adjusted constraint weights for Japanese accented words

	w_L	w_k	w_H	w_D	w_B
Medial-accented	0.04	0.06	0.77	0.13	
Final-accented	0	0.03	0.96	0.01	
Adjusted weights	0.01	0.03	0.18	0.01	0.77

w_B is so high that it can hardly be violated in phonetic realization. The next highest is w_H, and other constraint weights are very small after the adjustment. Conceptually, this means that for the Japanese accentual peak, it is most important to stay within the word, while an occurrence outside the accented mora is tolerable as long as the peak is still within the word. The alignment of the H peaks at the anchoring point (i.e. about 51% into the rime in the accented mora, (39-b)) is next most important.

4.1.9 unaccented words

This section discusses the alignment pattern of L and H tones in the phrase-initial rise in Japanese unaccented words. Japanese words are categorized into accented and unaccented words. Accented words have an accentual fall (HL) at the accented mora. Unaccented words do not have accentual falls, but there is an obligatory rise in word-initial position. Thus, the first mora bears a L% tone, and the second mora bears a phrasal H tone. F0 falls steadily after the second mora (Pierrehumbert and Beckman, 1988). In this section, we will show that phrasal peaks are less strictly aligned than accentual peaks.

For unaccented words, it is supposed that the anchor for the initial L is the same as in accented words. Thus, the previous A_L estimate in (6-b) is used, repeated in (30). This means that we are hypothesizing that the initial L in accented and unaccented words has the same phonological nature.

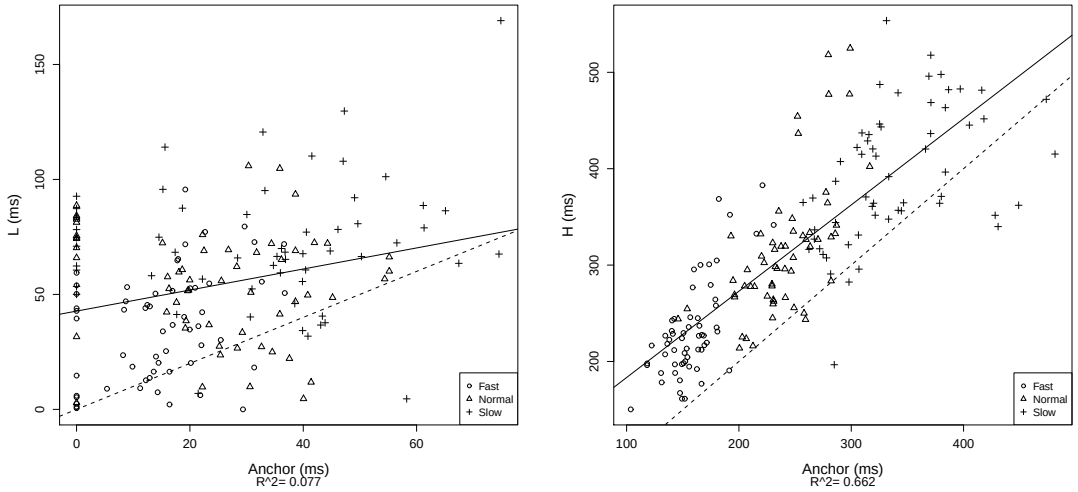
(30) $A_L = v1$

However, the H tones in accented words and unaccented words are not guaranteed to be

the same. The H tone is the peak of the lexical pitch accent in accented words, but it is a peak of the phrase-initial pitch rise in unaccented words. Thus, the anchor for the two H tones may well be different. Following the same procedure as before, A_H for unaccented words was obtained as in (31). The A_H for unaccented words is a bit later than the A_H for accented words (8).

(31) $A_H = v2 + 0.702 \cdot mora2$

Figure 4-12 illustrates the timing of L and H tones against their respective anchors. A noticeable difference between unaccented and accented words can be seen in the alignment of H in Figure 4-12b. The H peaks in unaccented words show a more scattered pattern than the H peaks in accented words. Thus, the portion of the variance in H peaks explained (R^2) is much smaller in unaccented words ($R^2 = 0.662$) than in accented words (For medial- accented, $R^2 = 0.832$; for final-accented, $R^2 = 0.924$). On the other hand, the alignment pattern of L does not seem very different from the L tone of the accented words.



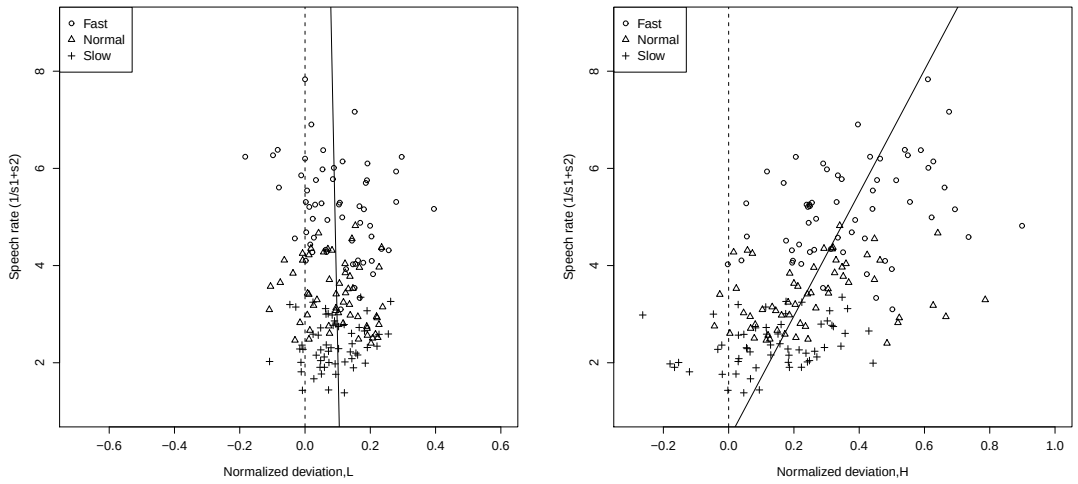
(a) Alignment of L

(b) Alignment of H

Figure 4-12: Alignment of L and H tones in unaccented words: (a)L against A_L, (b) H against A_H, the dashed line is $y = x$. The anchors are the same as in the accented words, estimated based on the AD model.

The deviations of the tone with regard to the anchor are illustrated in Figure 4-13. L stays relatively close to the anchor across all speech rates. On the other hand, H deviates to a greater degree than the H tones in accented words (cf. Figure 4-7b and Figure 4-11b). Given this, we might expect that Align(H) will have a lower weight in unaccented words than in accented words. At the same time, the weight of the Duration constraint will be higher in unaccented words than in accented words.

Hypothesizing that the same constraints will be active in the unaccented words as in the accented words, the Alignment-Duration model with the DelayL constraint, as in (32) (repeated from (23)), is applied to the unaccented words.



(a) L deviation from A_L (b) H deviation from A_H

Figure 4-13: Deviation of unaccented words: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. The anchor is the precise estimate.

(32) a. $T(L) = aA_L + bH + c$
where $a = \dfrac{w_L}{w_L + w_D + w_k}, b = \dfrac{w_D}{w_L + w_D + w_k}, c = \dfrac{w_k k - w_D D}{w_L + w_D + w_k}$

b. $T(H) = m(A_H - L) + T(L) + (1 - m)D$
where $m = \dfrac{w_H}{w_H + w_D}, n = (1 - m)D$

Mixed-models were fitted to the data to find the constraint weights. According to the fitted mixed models, a was -0.02, but because it was not significantly different from zero [$t(173) = -0.216, p = 0.83$], we use $a = 0$, in order to avoid negative weights. Other coefficients were obtained as follows: $b = 0.178, c = -1.54, m = 0.880, n = 91$. From these, the computed constraint weights are in (33). As expected, w_H has a relatively lower weight than in accented words.

(33) The constraint weights for unaccented words:
$w_L = 0, w_k = 0.36, w_H = 0.57, w_D = 0.08$

The D value, computed from the H model (32-b), was 761 ms, which seems too large to be the target duration of a rise. As before, it is difficult to observe the effects of a constraint that has a very small weight. w_D is still very small in unaccented words as well as in accented words, thus it is difficult to estimate D accurately. The mean duration between L and H was 258 ms. The k value was computed from c in (32-a) using the two D values. From D_H (=761 ms), $k = 162$. From the mean of L to H (=258 ms), k was 54. If the real k value is somewhere between these two estimates, the k value may explain how the L varies from the phrase onset to the end of the first mora. The mean of the end of the first mora was 104 ms.

4.1.10 Summary

Table 4.5 summarizes the constraint weights in the three categories of phrase-initial pitch rises in Japanese. The weights for the accented words presented in this table are the values before the adjustment with w_B. After the adjustment, constraint weights other than w_B are so small that it is difficult to compare the ratio between constraint weights, especially when they are very small (such as w_L and w_D). The relative weights of the constraints reflect differences in phonological status of the tones (accented or unaccented) and context (word-medial, word-final). The accentual peaks had a higher w_H than the phrasal peaks. In the same word-medial context, the weight of Align(H) was higher in accented words (0.77) than in unaccented words (0.57). For accented words, word-final peaks are more strictly aligned than word-medial peaks to avoid crossing the word boundary. We speculate that the weights are the same if phonological status is the same (i.e. medial accented and 2), but the weight of 0.96 in the final-accented group in fact includes the effects of the Boundary constraint that requires H peaks to stay within the word.

Table 4.5: Summary of constraint weights in Japanese

	Tone	Context	w_L	w_k	w_H	w_D
medial-accented	Lexical	Word-medial	0.04	0.13	0.77	0.06
final-accented	Lexical	Word-final	0	0.03	0.96	0.01
unaccented	Boundary		0	0.36	0.57	0.08

Table 4.6: w_H for lexical and boundary tones

	Tone	w_L	w_k	w_H(Lexical)	w_H(Bounday)	w_D	w_B
accented	Lexical	0.01	0.03	0.18		0.01	0.77
unaccented	Boundary	0	0.36		0.57	0.08	

Table 4.6 summarizes constraint weights for accented and unaccented words, combining the medial and final contexts of the accented words. The direct comparison between accented and unaccented words is unavailable because unaccented words are all word-medial, so w_B is unknown. The table shows that the grammar must have two separate w_H values: one for lexical tones and one for boundary tones. That is, the phonetic grammar needs to know which w_H should be applied for a given tone, depending on the phonological status of the tone (lexical or boundary). Our proposal is that the distinction between lexical and boundary tones is made in phonological representations, which will be discussed in Chapter 5.

4.2 Lexical Tone: Mandarin Chinese

Mandarin Chinese is a lexical tone language. There are four tones in Mandarin: Tone 1 (High), Tone 2 (Rise), Tone 3 (Low fall-rise), and Tone 4 (High-fall). In this section, Tone 2 is also referred to as the Rising tone. Unstressed syllables are toneless, or said to bear a 'neutral tone' (Chao, 1968). Unstressed syllables in Mandarin occur in restricted positions:

they are usually cliticized to the preceding stressed syllable. Neutral tones do not have a pitch target; rather, their pitch varies depending on the tone in the preceding syllable (Chao, 1968; Li, 2003:41).

In Mandarin, the alignment between the tone and its associated syllable is expected to be stable for the following reasons. First, tones are contrastive: mā 'mother', má 'hemp', mǎ 'horse', mà 'scold'. We have shown that lexically-specified tones (such as lexical pitch accents in Japanese) are more strictly aligned than phrasal boundary tones (such as tones in the phrase-initial rise in Japanese unaccented words or Seoul Korean Accentual Phrases). In Mandarin, tones are specified in the lexicon, so the alignment is expected to be more stable than phrasal tones in Seoul Korean or Japanese unaccented words. Second, because most syllables are specified for tone, misalignment would result in intrusion on the neighboring tones. Previous studies support the stability of alignment in Mandarin. Xu (1998, 1999) shows that regardless of syllable structure or speech rate, the onset of the Rising tone is found near the center of the syllable, and the offset of the Rising tone is found near the offset of the syllable. Xu thus claims that the domain of tone implementation in Mandarin is the associated syllable. The fact the peak occurs a bit later than the syllable offset is attributed to carryover coartiulation effects by Xu (1999).

At the same time, we also expect that Mandarin will have a high weight on shape constraints (i.e. the Duration constraint), because it has been shown that in Mandarin, rising and falling tones have targets for their shape. The peak of Tone 1 (High) stays within the syllable, whereas the peak of Tone 2 is found a bit later than the offset of the syllable (Xu, 1999; Li, 2003). Peak delay in Tone 2 can be explained if Tone 2 has an intrinsically dynamic target so that it requires a minimal time to produce, whereas Tone 1 is static (Xu, 1999). Xu (1998) suggests that rising and falling tones in Mandarin are implemented with dynamic targets rather than static targets, because the slope of the Rising tone is not systematically affected by speech rate or syllable structure. This evidence suggests that the Rising tone in Mandarin has a shape target, so we can expect that the shape of the tone will stay relatively stable, and so the Duration constraint will have a higher weight than in other languages where tones have static targets.

We thus examine the alignment and duration of the Rising tone in Mandarin. In particular, we look at whether tendencies to both segmental anchoring and shape targets are simultaneously observed in Mandarin as was the case in Seoul Korean and Japanese. Despite the widely-accepted stability of tonal alignment in Mandarin, if Mandarin also shows systematic deviation from the anchor depending on speech rate, it means that Mandarin can also be modeled with targets for both alignment and rise duration. Thus, we conducted a comparable experiment to those with Seoul Korean and Tokyo Japanese. As in the previous experiments, the L and H tone in phrase-initial Rising tones will be analyzed.

The phonological context of the Rising tone is varied: the Rising tone is followed by another Rising tone (a sequence of Tone 2- Tone 2), or a neutral tone (a sequence of Tone 2-Tone 0). Li (2003: 58-59) found that there is a significant effect of the tonal context on peak alignment of Tone 2. That is, the peak of the Rising tone is significantly more delayed when it is followed by a neutral tone than when it is followed by a lexical tone (Tone 3). Also, the number of following neutral tones significantly affects peak delay of Tone 2: the more the

later the peak. When Tone 2 is followed by one neutral tone, the peak of Tone 2 is found in the onset consonant of the following syllable. When there are two neutral tones after Tone 2, the peak is delayed into the middle of the following rime or even up to the end of the following syllable. In the experiment in this section, we expect greater deviations of the H peak from its anchor in the context of a neutral tone (Tone 2- Tone 0) than in the context of Tone 2 (Tone 2- Tone 2).

4.2.1 Experiment

Hypotheses

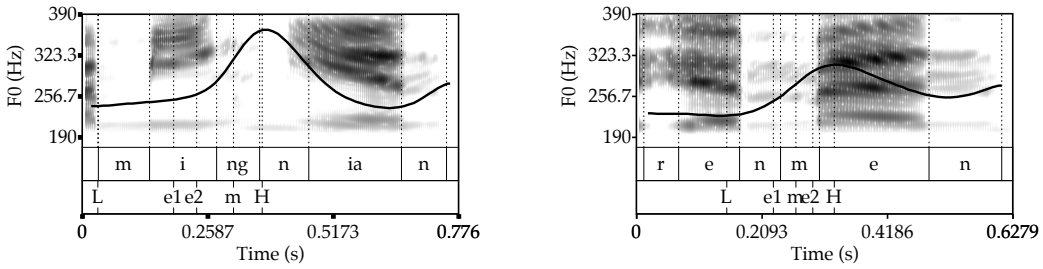
The primary hypothesis tested with Mandarin is the same as in the previous languages. That is, we test whether Mandarin also exhibits tendencies towards both segmental anchoring and target duration of a rise. Secondly, we test whether tonal alignment is affected by the phonological context. In particular, the tonal timing of the Rising tone is compared in the context of another Rising tone and in the context of a neutral tone.

Methods and speech materials

There were four Mandarin speakers: two females (C1, C2) and two males (C3, C4). They were born or lived in Beijing for most of their lives. One female speaker (C2) was in her 20's, one male speaker (C4) was in his 40's, and the two other speakers (one male (C3) and one female (C1)) were in their 60's. The speech materials were presented on a sheet of paper, and speakers read the speech materials at normal, fast, and slow speech rates. The recordings were made in a sound-attenuated recording booth at the phonetics lab in the MIT Linguistics department, all the other technical details were the same as in the previous experiments.

The timing of L and H tones comprising the Rising tone in word-initial positions is examined. The speech materials consisted of 15 disyllabic words, where each syllable carries the Rising tone. Thus, each word has a sequence of two Rising tones (Tone2-Tone2: the Rising tone in the lexical-tone context). An example is shown in Figure 4-14a. In addition, there were 5 disyllabic words, where the first syllable carries the Rising tone, and the second syllable is toneless (Tone 2-Tone 0: the Rising tone in the neutral-tone context), e.g. [rénmen] 'people' in Figure 4-14b. The examples show that as in Li (2003), the peak of the Rising tone occurs later when the following tone is neutral tone. In 4-14a, the peak is at the beginning of the consonant in the following syllable, in 4-14b, the peak is after the beginning of the vowel in the following syllable. There were also 9 fillers with longer words (three or four syllables) with various tonal combinations, to prevent monotony.

F0 mimina, F0 maxima, and segmental boundaries were manually marked. The inflection points (lower and upper elbows) were located using the three-piece linear regression described in Section 2.2.3. The shapes of rises are categorized based on the slopes of the three regression lines. Maximum velocity points were located using the derivative of spline smoothing curve fitting. These procedures are the same as those employed before in Korean and Japanese.



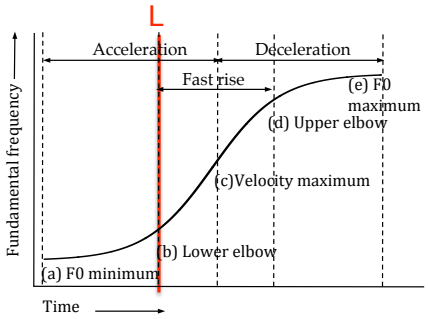
(a) Lexical-tone context (Tone 2- Tone 2)

(b) Neutral-tone context (Tone 2- Tone 0)

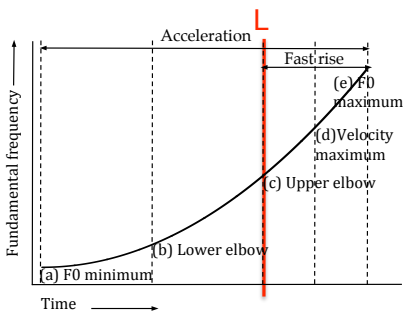
Figure 4-14: Word-initial Rising tone in Mandarin: (a)[mín̊gnín̊an] 'next year', (b) [rénmen] 'people'

In Seoul Korean and Japanese, F0 minima and F0 maxima were used as the locations of L and H tones. However, in Mandarin, F0 minima are not a reliable correlate of the L tone. The shape of the Rising tone consists of two components, a shallow rise followed by a steep rise, as shown in 4-14a. The slope of the portion between the F0 minimum and the start of the steep rise varies from a plateau to a shallow rise. Xu (1998) also notes that F0 minima cannot be used as the location of L in the Rising tone, because F0 minima are far from the actual start of the rise and hard to reliably measure because there is often a long plateau. Instead, Xu takes the point of rapid acceleration point as L, by taking the local maximum in the second derivatives of the F0 curves. That is, the local acceleration maxima can be found by taking the second derivatives of the function fitted to the actual F0 curves. In fitting the F0 curve to find the acceleration maximum, polynomial fitting will not be appropriate: a sigmoid rise can be considered to have two local extrema, for which third-order polynomial fitting may be appropriate. However, because the second derivative of the third-order polynomial is a straight line, a maximum of the second derivative is zero. Instead, we tried spline functions to fit the F0 curves, but because the acceleration maximum found in this way was too often sensitive to segmental perturbation.

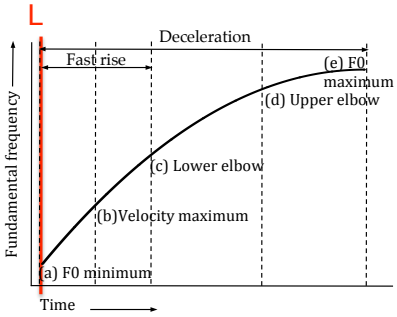
Instead, we use elbow measures to locate L tones. As described in Section 2.2.3, a rise can be categorized as sigmoid if the second line is the steepest, scoop if the third line is the steepest, and dome if the first line is the steepest. In Mandarin, we define the L tone as the start of the steep rise. For sigmoid rises, the fastest line segment is the second, so L is at the beginning of the second line, the lower elbow (Figure 4-15a). For scooped rises, the third line is the fastest, so L is at the beginning of the third line, i.e. the upper elbow (Figure 4-15b). For domed rises, the first line is the fastest, so the F0 minimum would be the L tone in principle (Figure 4-15c) . However, it turns out that in Mandarin, there were no 'dome' shaped rises. In summary, the first inflection point was L for sigmoid rises, the second inflection point was L for scooped rises. There were no dome shaped Rising tones in Mandarin.



(a) Sigmoid



(b) Scoop



(c) Dome

Figure 4-15: Locating L tones using elbows, depending on shape: (a) the lower elbow for a sigmoid rise, (b) the upper elbow for a scooped rise, and (c) the F0 minimum for a domed rise

4.2.2 Overall shape

Table 4.7 summarizes the shapes of the Rising tone in Mandarin. Sigmoid and scooped shapes accounted for most of the Rising tone shapes, with about the same proportions. Yet, scooped shapes are found more in slow speech than in faster speech. This is similar to the previous results with Korean and Japanese; the scooped shape requires a longer time before the fast rise, thus it is less favored under time pressure. It is notable that domed rises were not found in the experiment. Only sigmoid and scooped rises are allowed for the Rising tone, suggesting that the Rising tone has a target for a specific shape. That is, the Rising tone consists of a shallow rise followed by a fast rise, which is possible only in sigmoid and scooped rises. The difference between sigmoid and scooped rises is that the fast rise starts later in scooped rises than in sigmoid rises. Thus, there is a slight tendency for sigmoid rises to be favored over scooped rises under time pressure, as we see in Table 4.7. 'none' is the label applied to shapes that do not belong to any of these categories, and N/A applies to the data where F0 values cannot be measured. 'none' and N/A's were discarded in the analysis.

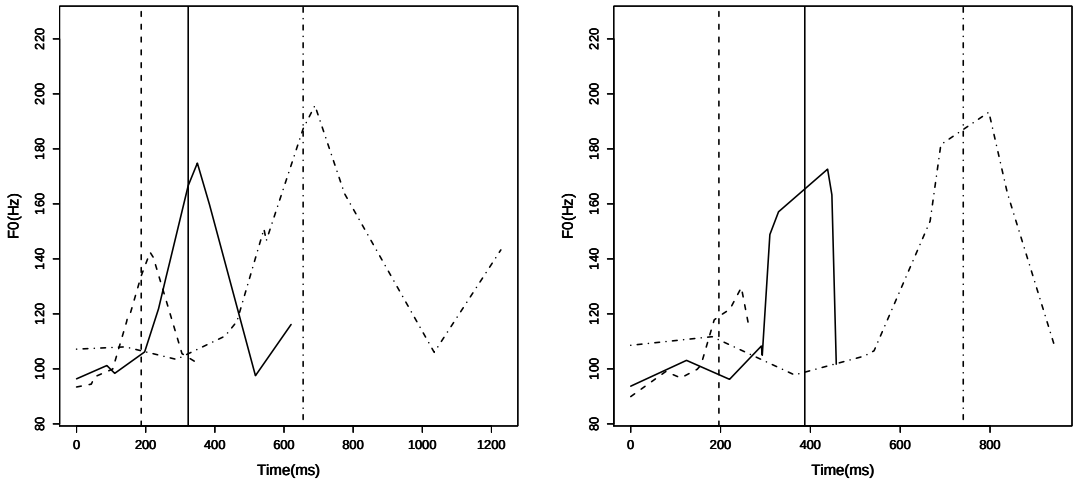
Table 4.7: Shape of rises in each speech rate.

	sigmoid	dome	scoop	none	N/A
All	214 (47%)	0 (0%)	215 (48%)	16 (4%)	7 (2%)
Fast	78 (52%)	0 (0%)	70 (46%)	2 (1%)	1 (1%)
Normal	74(48%)	0 (0%)	67 (44%)	9 (6%)	3(2%)
Slow	62(42%)	0 (0%)	78 (53%)	5(4%)	3 (2%)

Figure 4-16a shows the averaged pitch contours of the initial Rising tone in Tone2-Tone2 words, for all speakers together. The vertical lines correspond to the end of the first syllable in each speech rate. The peak is aligned close to the end of the first syllable for all three speech rates. The magnitude of the rise is reduced in faster speech. This can be interpreted as the maintenance of a constant slope of the rise.

The mean of the maximum velocity, for both lexical and neutral tone conditions together, was 651: by speech rate, 620 (fast speech), 788 (normal speech), and 546 (slow speech). Mixed-effects models were fitted to the data to test whether the slope of the rise is affected by categorical speech rate or not. Maximum velocity was used as a measure of the slope of rises. Table 4.8 shows the compared models. The models (b) and (c) were significantly better than model (a). Because models (b) and (c) are not nested, the deviance values were compared. Model (b) had a lower deviance (5668) than model (c) (5757). Adding categorical speech rate did not improve the fit significantly [(b) vs. (d)]. Thus, the best model is (b). This shows that maximum velocity is not significantly affected by speech rate category, although there were speaker-dependent variations (by-speaker random slopes for speech rate were significant in (b)). This is a similar result as in Xu (1998): the slope of the Rising tone did not vary systematically depending on speech rate.

Figure 4-16b shows the averaged pitch contours of the initial Rising tone in Tone2-Tone0 disyllabic words. The high peak occurs later relative to the end of the first syllable in the neutral-tone context than in the lexical-tone context. At the same time, the peak is not as



(a) Lexical-tone context (Tone 2- Tone 2)

(b) Neutral-tone context (Tone 2- Tone 0)

Figure 4-16: Averaged F0 contours in Mandarin: (a): The initial Rising tone in the lexical-tone context. (b): The initial Rising tone in the neutral-tone context. Solid line: normal speech rate, dashed line: fast rate, dot-dashed line: slow rate. The vertical lines correspond to the end of the first syllable.

Table 4.8: Model comparisons for the effect of speech rate on maximum velocity. The best model is (b).

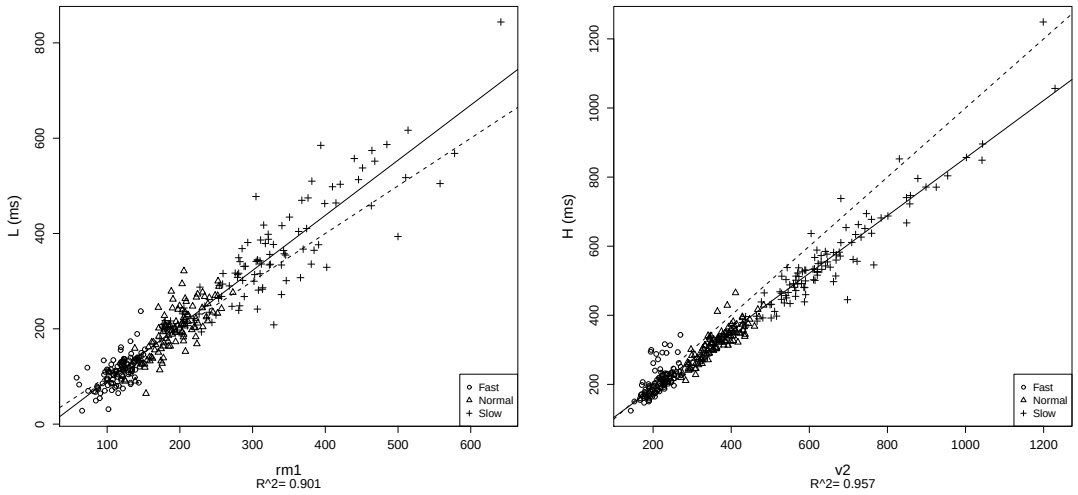
	Dependent variable	Fixed	Random		LRT
			Intercepts	Slopes	
(a)	max.velocity	none	speaker	none	
(b)	max.velocity	none	speaker	speech rate	$\chi^2(5) = 183.93, p < 0.001$ (with (a))
(c)	max.velocity	speech rate	speaker	none	$\chi^2(2) = 95.41, p < 0.001$ (with (a))
(d)	max.velocity	speech rate	speaker	speech rate	$\chi^2(2) = 3.57, p = 0.16$ (with (b))

pointy as the lexical-tone context; there is a shallow rise toward the absolute F0 minimum. This is because the location of rise peaks may vary more in the neutral-tone context, which will be discussed more in Section 4.2.6. The duration of the first syllable is longer in the neutral tone context (the location of the vertical lines is significantly later in the neutral-tone context than in the lexical-tone context, especially in normal and slow speech). So, there is more time to produce the initial Rising tone in the neutral-tone context. Even so, rise peak is delayed more in the neutral-tone context, than in the lexical-tone context.

4.2.3 Effects of segmental anchoring

For Tone 2- Tone 2 words, various segmental points were examined to find the segmental landmark that has the highest correlation with L tones and H tones respectively. For L, 'rm1', the middle of the rime in the first syllable, has the best correlation with L ($R^2 = 0.901$). For H, 'v2', the beginning of the second vowel has the best correlation ($R^2 = 0.957$) with H.

Figure 4-17 shows the timing of L and H tones against their respective anchors. Compared to previous languages, L and H tones in the Rising tone in Mandarin show the highest correlations, suggesting strong effects of segmental anchoring.



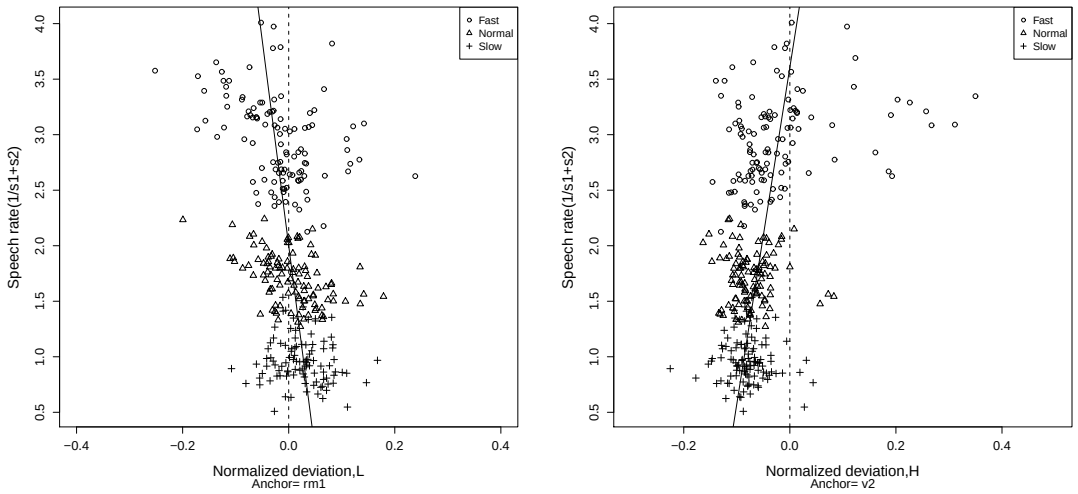
(a) Alignment of L (b) Alignment of H

Figure 4-17: Alignment of L and H in Tone 2- Tone 2 words: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$

4.2.4 Effects of target duration

With the strong segmental anchoring effects, Mandarin also shows a tendency to maintain a target duration. This again suggests that segmental anchoring and target duration are not incompatible with each other, but both effects can co-exist, even when the effects of segmental anchoring are very strong. Figure 4-18 shows the deviation of L and H tones from their respective anchors (as identified in the previous section), normalized by local speech rate (the inverse of the duration of the first two syllables).

Systematic relations between relative deviation from the anchor and local speech rate were observed. That is, L is found earlier than the anchoring point in fast speech, later in slow speech (Figure 4-18a). H is found earlier with regard to the anchor in slow speech, later in fast speech (Figure 4-18b). These results are in line with the results for the other languages, Korean and Japanese. That is, as in other languages, the opposite directions of deviation of L and H tones suggest that there is a target duration. Thus, a rise starts earlier and terminates later under time pressure.



(a) L deviation from A_L (b) H deviation from A_H

Figure 4-18: (a): L deviation from A_L, (b): H deviation from A_H, the dashed line is the position of the anchor.

4.2.5 The Alignment-Duration model for the lexical-tone context tones

Precise anchor estimate

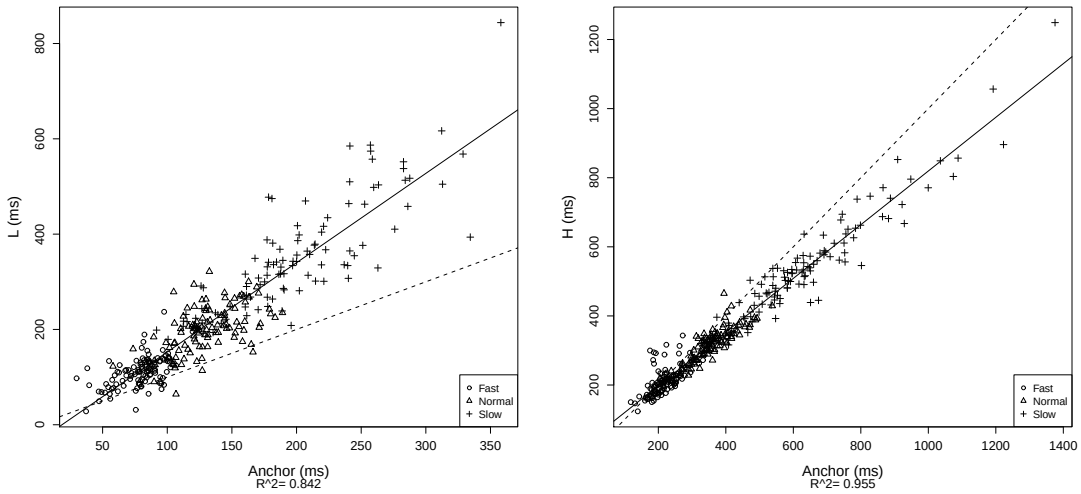
According to the results in the previous two sections, Mandarin also exhibits tendencies to both segmental anchoring and target duration. Thus, we can apply the weighted-constraint model developed in Chapter 3, with the constraints Align(L), Align(H), and Duration.

Before finding the constraint weights, precise locations of A_L and A_H are estimated based on the proposed model. Following the same procedure as in Japanese and Korean, the anchors for L and H are estimated. Assuming that the anchor of L is within the first syllable, A_L is divided into $(v1+p\cdot rime)$, where $v1$ is the beginning of the vowel in the first syllable, $rime$ is the duration of the first rime, and p is the proportion into the first rime. A mixed-effects model was fitted to the data with $T(L)$ as a dependent variable, $(v1-H)$ and $rime$ as fixed effects, by-speaker random intercepts, and by-speaker random slopes for $rime$. The coefficient for $(v1-H)$ was 0.457, the coefficient for $rime$ was 0.084, thus $a = 0.457, a \cdot p = 0.084$, so $p = 0.084/0.457 = 0.183$. Thus, A_L is expressed as in (34).

(34) $A_L = v1 + 0.184 \cdot rime$

The same procedure was applied to find the precise estimate of the anchor for H, based on the weighted-constraint model with alignment and duration constraints. It is assumed the anchor for H is also in the first syllable. A_H is divided into $(v1 + p \cdot rime)$, where $v1$ is the beginning of the vowel in the first syllable, $rime$ is the duration of the first rime, and p is the proportion into the first rime.

A mixed-effects model was fitted to the data with $T(H)$ as a dependent variable, $(v1-L)$ and $rime$ as fixed effects, by-speaker random intercepts, and by-speaker random slopes



(a) Alignment of L (b) Alignment of H

Figure 4-19: Alignment of L and H in Tone 2-Tone 2 words: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$. Precise estimates of the anchors based on the AD model.

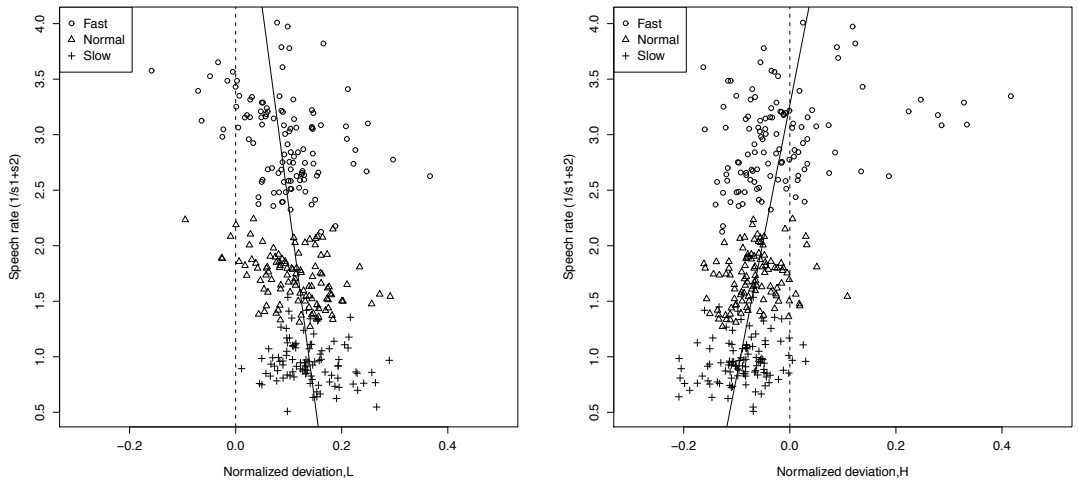
for $rime$. The coefficient for $(v1 - L)$ was 0.493, the coefficient for $rime$ was 0.651, thus $c = 0.493, a \cdot p = 0.651$, so $p = 0.651/0.493 = 1.318$. Thus, A_H is expressed as in (35).

(35) $A_H = v1 + 1.318 \cdot rime$

In summary, A_L is within the first syllable, i.e. 18% into the first rime. On the other hand, A_H is outside the first syllable where the tone is phonologically associated. The mean of the location of H is 356ms, and the mean of the first syllable offset is 328ms, which is earlier than the mean of H. Considering this, the actual H tones occur on average after the end of the first syllable, thus, it is not unreasonable that the computed anchor is found after the end of the first syllable. The results correspond to previous studies finding that the peak of the Rising tone occurs a bit later than the offset of the associated syllable (Xu, 1998; Li, 2003)

Constraint weights

The weights are calculated based on the coefficients of the mixed-effects models fitted to the data. For L, a mixed-effects model was fitted to the data with $T(L)$ as a dependent variable, $(A_L - H)$ as a fixed effect, offset of $T(H)$, by-speaker random intercepts and by-speaker random slopes for $(A_L - H)$. For H, a mixed-effects model was fitted to the data with $T(H)$ as a dependent variable, $(A_H - L)$ as a fixed effect, offset of $T(L)$, by-speaker random intercepts and by-speaker random slopes for $(A_H - L)$. Following the same procedure as in the previous languages, the constraint weights were computed as in (36). The weights of the three constraints are very similar to one another, suggesting that in Mandarin, both alignment and target duration are equally important.



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-20: Deviation of L and H in the Tone2-Tone2 words: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. Precise estimates of the anchors based on the AD model.

(36) The constraint weights in Mandarin (the lexical-tone context):
$w_L = 0.29, w_H = 0.35, w_D = 0.36$

The D values were $D_L = 69, D_H = 91$. The uncertainty of the D values was estimated by simulation through drawing 1000 possible pairs of slope and intercept. The 95% confidence interval was for D_L, 39 to 100, for D_H, 83 to 100. The 95% confidence interval for the difference between D_L and D_H was -10 to 53, which includes zero, so D_L and D_H may not be significantly different from each other. From the distributions of D_L and D_H values, the D value is estimated, by multiplying the probability density functions (PDF) of D_L and D_H values. The peak of the combined PDF is the most probable D value. The D value estimated in this way was 91 ms.

Model comparison

The Independent-Alignment model was compared with the Alignment-Duration model. To review, the IA model is expressed as in (37), as before. The model expresses the idea that the timing of a tone is determined by its anchor and a target at a fixed interval from the phrase onset (indicated by non-zero intercepts).

(37) The Independent-Alignment model

a. $T(L) = a \cdot A_L + b$

b. $T(H) = c \cdot A_H + d$

The anchors were estimated based on the model in (37). In Mandarin, both L and H are associated with the first syllable. Thus, both A_L and A_H are substituted with $(v1{+}p{\cdot}rime1)$,

where $v1$ is the beginning of the first vowel, $rime1$ is the duration of the first rime, and p is the proportion into the first rime. With this substitution, (37) is expressed as (38).

(38) a. $T(L) = a(v1 + p \cdot rime1) + b = a \cdot v1 + a \cdot p \cdot rime1 + b$
b. $T(H) = c(v1 + p \cdot rime1) + d = c \cdot v1 + c \cdot p \cdot rime1 + d$

To find the p value in (38-a), a mixed-effects model was fitted with $T(L)$ as a dependent variable, $v1$ and $rime1$ as fixed effects, by-speaker random intercepts for speakers and by-speaker random slopes for $rime1$. The same procedure was applied to A_H. The anchor estimates based on the IA model is as in (39).

(39) a. $A_L = v1 + 0.63 \cdot rime1$
b. $A_H = v1 + 0.91 \cdot rime1$

According to the coefficient and intercept estimates of the fitted models, $T(L)$ and $T(H)$ are expressed as in (40).

(40) a. $T(L) = 0.62A_L - 19.39$
b. $T(H) = 1.04A_H + 33.99$

Table 4.9: Summary of deviances (the lexical-tone context)

	T(L)	T(H)
The IA model	3168	3308
The AD model	3068	3044

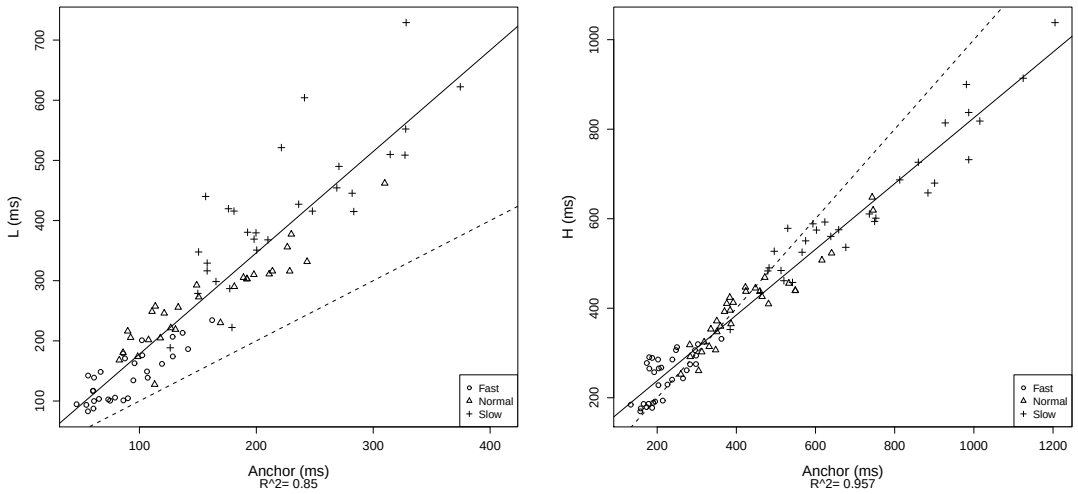
The deviance values are shown in Table 4.9. Deviance values are smaller in the AD model than the IA model. Thus, the Alignment-Duration model is better than the IA model in predicting the timing of the L and H tones in Mandarin Rising tone, so the duration constraint is playing a significant role.

4.2.6 The neutral-tone context

It is expected that the alignment patterns of the H peaks in the word-initial Rising tone will be different in the Tone 2- Tone 0 words (neutral-tone context) and in the Tone 2- Tone 2 words (lexical tone context). In this section, we examine the hypothesis that the alignment of Tone 2 will be less strict in the context of Tone 0 than in the context of another Tone 2.

In the previous experiment with Japanese, it was assumed that the anchor locations are the same in the accented words in different phonological contexts (word-medial or word-final). Likewise, in Mandarin, the same anchors are assumed for different phonological contexts: Tone 2 followed by a lexical tone (Tone 2) or a neutral tone (Tone 0).

Figure 4-21 shows the L and H tones against their respective anchors. Figure 4-22 shows the deviation of the tones from their anchors, normalized by local speech rate (the inverse of the duration of the two syllables). Compared to the deviation plots for the Tone 2-Tone 2 words (Figure 4-20b), the deviation plots for the Tone 2-Tone 0 words show more deviation in the alignment of H tones (Figure 4-22b).



(a) Alignment of L

(b) Alignment of H

Figure 4-21: Alignment of L and H in Tone2-Tone0: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$

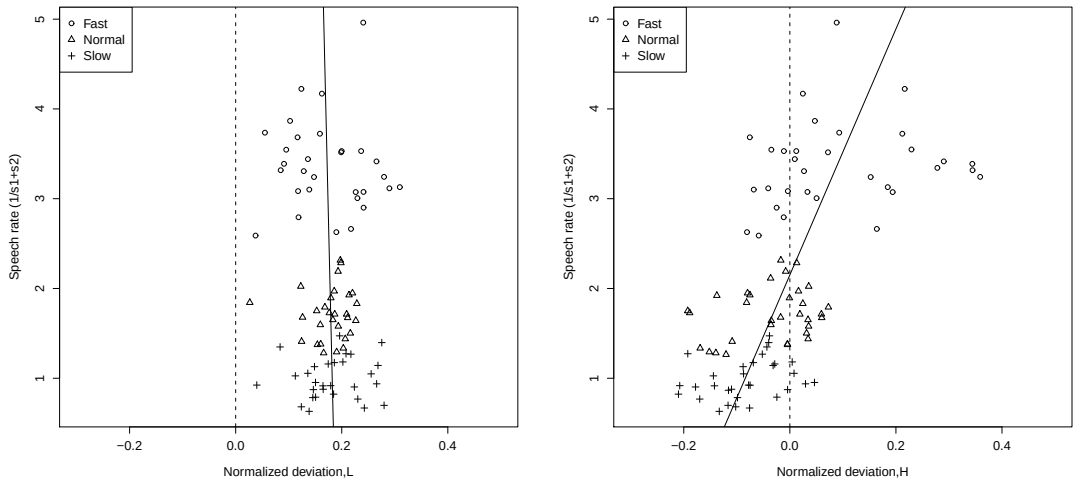
The Alignment-Duration model was applied following the same procedure. According to the model, the constraint weights were: $w_L = 0.32, w_H = 0.25, w_D = 0.43$. The constraint weight for the Alignment of the H tone is relatively lower than in the lexical tone context, reflecting that the H peaks are less strictly anchored when the Rising tone is followed by a neutral tone than by a lexical tone. However, the D values from the L model (58) and the H model (125) did not converge. The 95% confidence interval was 37 to 110.

Nevertheless, the AD model turns out to be better than the IA (Independent-Alignment) model. Table 4.10 compares the deviance values from the IA model and the AD model respectively. For both L and H, the Alignment-Duration model was better than the other model.

Table 4.10: Summary of deviances (the neutral-tone context)

	T(L)	T(H)
The IA model	886	947
The AD model	863	856

It seems that the problem of diverging D values in the data of the neutral-tone context arises from unreliable measurements of L locations due to F0 perturbation by initial consonants (Lehiste and Peterson, 1961; van Santen and Hirschberg, 1994). Compared to the lexical-tone context, most of the neutral-context words start with obstruent consonants: 3 out of 5 words started with /p/ or /z/. On the other hand, only one out of 15 the lexical-tone context words started with an obstruent, the voiceless unaspirated stop /g/. Due to the initial obstruents in the neutral-tone context words, L minima are found somewhere in the vowel, because the initial raising of F0 due to segmental effects creates a dip in the



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-22: Deviation of L and H in Tone2-Tone0: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor

following vowel, as shown in 4-23. In the case of the lexical-tone context, F0 minima were not reliable locations for the L tone, because F0 minima are often found on a plateau or even before the beginning of the target word. On the other hand, in the case of the Tone 0 context, F0 minima are usually within the vowel. The elbow measures based on these F0 minima do not always correspond to the position where the fast rise intuitively begins: the elbow is later than the actual start of the rise. Because the minimum is late due to segmental effects, the three-piece linear regression is fitted to an interval that starts late.

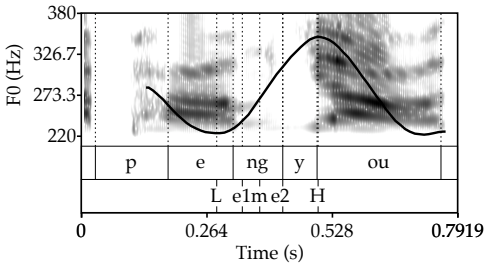


Figure 4-23: Segmental effects delay L minima: [péngyou] 'friend'

4.2.7 Summary

Previous studies have shown that both the alignment and shape of pitch movements are stable in the Rising tone in Mandarin (Xu, 1998, 1999). Our results are broadly consistent with previous results, but crucially, we argue that alignment and shape are not invariant, although they may be fairly stable. Alignment and shape systematically vary depending

on speech rate and segmental duration. Moreover, the results with the toneless syllables suggest that the apparent strictness of H alignment is partly due to competition with adjacent lexical tones.

The constraint weights obtained from our model reflect these patterns. The weights for the Alignment constraint (w_L and w_H) and the Duration constraint (w_D) are similar to one another, as shown in Table 4.11. In Mandarin, the weight of the Duration constraint is higher than in Korean or Japanese. This supports the idea that the Rising tone in Mandarin has a target for the shape of pitch movement, represented by its duration in our model. In addition, we have shown that the slope of the Rising tone (i.e. maximum velocity) is not significantly affected by categorical speech rate.

Table 4.11: Constraint weights in Mandarin

Condition	w_L	w_H	w_D
Tone2-Tone2	0.29	0.35	0.36
Tone2-Tone0	0.32	0.25	0.43

The weight of the Align(H) constraint is smaller in the Tone 2- Tone 0 words than in the Tone 2- Tone 2 words. This means that the H peak of the Rising tone is less strictly aligned when it is followed by a toneless syllable than when followed by a tone-carrying syllable. This result means that the relatively strict tonal alignment pattern that has been found in Mandarin is due to the avoidance of temporal intrusion into the tone in the following syllables.

4.3 Intonational Pitch Accent: English

F0 movements in English are characterized by boundary tones and pitch accents. Pitch accents are realized on the stressed syllables in prominent words (Pierrehumbert, 1980). It has been reported that English has a strong peak alignment pattern. Ladd et al. (1999) studied the beginning and end points of prenuclear accentual rises in British English. They found that the duration of a rise changes depending on speech rate: the faster the rate, the shorter the rise. Speech rate (fast, normal, slow) had only a small and inconsistent effect on the alignment of the beginning and end points of the rises. Based on these results, Ladd et al. (1999) rejected the hypothesis that duration of a rise is constant, and argued that pitch movements should be viewed as a sequence of pitch targets which are aligned with respect to segments. However, in the previous sections, we have already found effects of target duration even in languages that have been reported to have consistent alignment patterns, such as Mandarin (Xu, 1998) and Japanese (Ishihara, 2006). In this section, we examine whether English also exhibits the effects of target duration as well as the effects of segmental anchoring, as in the other languages we previously examined.

4.3.1 Experiment

For the English experiment, there were 19 target phrases and 9 fillers. The target phrases consisted of two or more words, combinations of various parts of speech, e.g. *aménable*

meanings, *malária in Nigeria*. We examine the pitch accent on the stressed second syllable in the first word in the phrase, for example, the rising pitch accent on the second syllable in *aménable* in *aménable meanings*, shown in Figure 4-24. This stressed syllable is followed by at least two unstressed syllables.

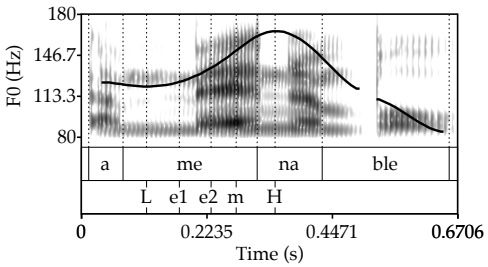


Figure 4-24: Intonational pitch accent in English: 'amenable' [əménəbl] (Speaker: E3)

The phrases were embedded in a carrier phrase: "No, I meant _". Speakers were asked to read the sentences as if they were correcting the first word in the phrase, which is italicized in the list presented to the speakers. This is because pilot tests showed that otherwise, speakers usually put a pitch accent on the second word only. The target word in the experiment thus had a corrective focus on it, and the following word was deaccented. The goal was to consistently elicit a rise that was far from a boundary tone. The sentences were randomized and rearranged so that not more than three target phrases can come in succession. The filler phrases consisted of two monosyllabic words (e.g. *tip toes*), and they were used to prevent monotony.

There were four native speakers of English, two females (E1, E2) and two males (E3, E4), all from the MIT Linguistics Department. One female speaker (E2) was from Canada, and the other speakers spoke American English. The speakers were naive to the purpose of the experiment. The speakers were asked to read the speech materials at normal, fast, and then slow speech rates. The recordings were made in a sound-attenuated recording booth in the phonetics lab at the MIT Linguistics Department. Other technical details were the same as in the previous experiments.

Measurements were carried out using the same methods as in the previous experiments, as described in Section 2.2.3. That is, F0 minima, F0 maxima, and segmental boundaries were manually marked using Praat. The inflection points (first and second elbows) were located using three-piece linear regression, and the maximum velocity points were located by taking the first derivatives of a smoothing spline fitted to the F0 trajectory. F0 minima and F0 maxima were used as the position of L and H tones at first. Later, however, the elbow measures replaced F0 minima as the location of the L tones in Section 4.3.6 because this turns out to provide a better account of the English patterns.

4.3.2 Overall shape

The majority of English rising pitch accents have sigmoid shapes (55%) or scooped shapes (35%), as Table 4.12 shows. To review, sigmoid shapes have the fastest pitch rise in the mid-

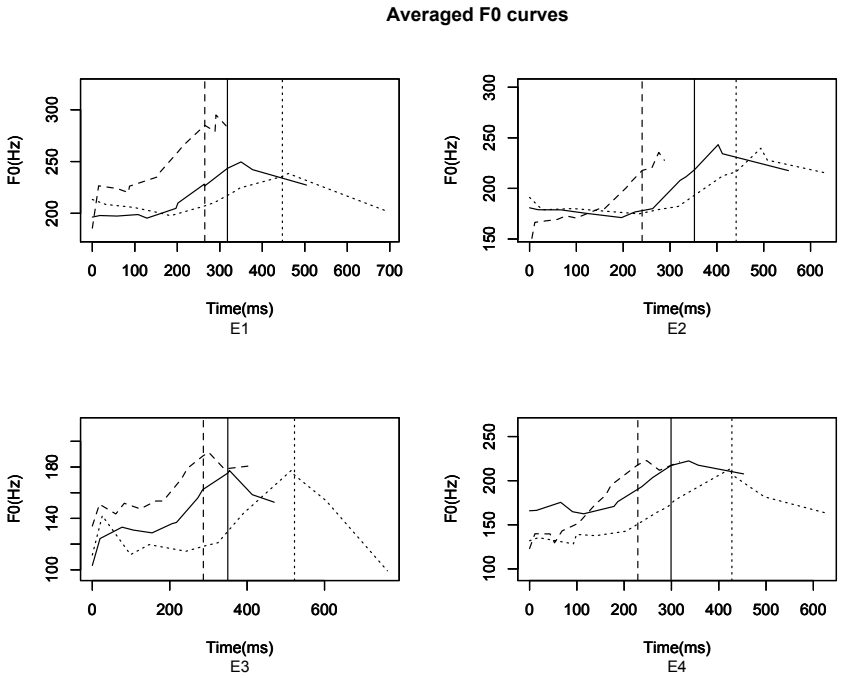


Figure 4-25: Averaged F0 contours over the first two syllables: Solid line: normal speech, dashed line: fast speech, dotted line: slow speech. The vertical lines are the end of the second (stressed) syllable

dle of the rising movement, and scooped shapes have the fastest pitch rise at the end of the rising movement, by definition (Section 2.2.3). Sigmoid and scooped shapes account for the majority in all speech rate categories. In fast and normal speech, sigmoids outnumber scooped rises by about a two-to-one ratio. Domed shapes are very rare in general, accounting for only 1% of the whole data. As before, 'none' refers to shapes that are not classifiable, usually as a result of segmental effects. N/A's are tokens whose pitch contours cannot be measured because of segmental effects or glottalization.

Table 4.12: Shape of rises in each speech rate in English

	sigmoid	dome	scoop	none	N/A
All	249 (55%)	5 (1%)	161 (35%)	19 (4%)	22 (5%)
Fast	90 (59%)	3 (2%)	53 (35%)	2 (1%)	4 (3%)
Normal	91 (60%)	2 (1%)	45 (30%)	7 (5%)	7 (5%)
Slow	68 (45%)	0 (0%)	63 (41%)	10 (7%)	11 (7%)

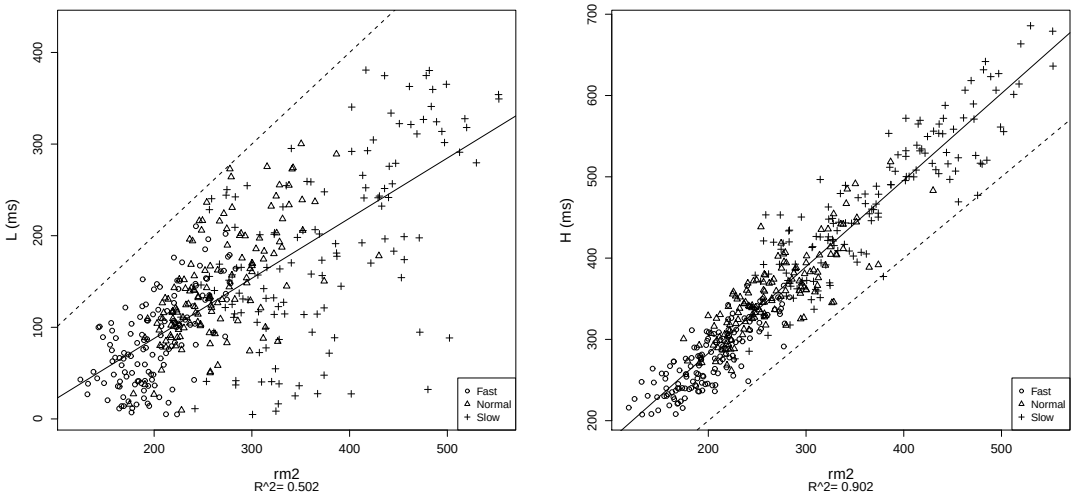
Although we intended to elicit bitonal L+H*, we cannot exclude the possibility that there may have been a few H* pitch accents, probably the ones classified as domed shapes (1%) in Table 4.12. However, the domed shapes are very few, and the average shapes are closer to L+H*. Figure 4-25 shows the averaged F0 curves of the portion from the beginning of the words up to the beginning of the vowel in the third syllable. Vertical lines show the end of the stressed second syllable. The average shape shows that there is a plateau before a fast rise; this means that the elicited pitch accent is L+H* rather than H*. For most speakers,

the average peak of the rise is found at the end of the second syllable in all speech rates. The peaks of the rises of speaker E2 occur later than the end of the second syllable, but the location seems consistent across speech rates. Thus, segmental alignment of rise peaks does not seem to be affected by speech rate to a great degree.

4.3.3 Effects of segmental anchoring

Several segmental landmarks were tested to find the position that has the highest correlation with L and H tones. Initially, F0 minima and F0 maxima were used as L and H tones. Linear regression models were fitted to the data with timing of a tone as the dependent variable, and timing of segmental position, speaker and their interaction as predictor variables. The R^2 values for each segmental position were compared. For L, the middle of the second vowel ('vm2') had the highest correlation with L ($R^2 = 0.661$). However, according to mixed-effects modeling, the best position was 'rm2', the middle of the second rime. Mixed-models were fitted to the data with timing of L as the dependent variable, timing of segmental position as a fixed effect, and by-speaker random intercepts and slopes for timing of a segmental position. The mixed model with the segmental position 'rm2' had a lower deviance ($G^2 = 4354$) than the one with 'vm2' ($G^2 = 4604$). Given that mixed-models are more appropriate when pooling data across speakers, and that the R^2 value of 'rm2' ($R^2 = 0.657$) was not much different from that of 'vm2', we consider 'rm2' to be the best anchor for L.

For H, the middle of the second rime ('rm2') has the highest correlation ($R^2 = 0.925$). This point remains as the best ($G^2 = 4027$) in mixed models with timing of H as a dependent variable, timing of segmental position as a fixed effect, by-speaker random intercepts, and by-speaker random slopes for timing of the segmental position.



(a) Alignment of L

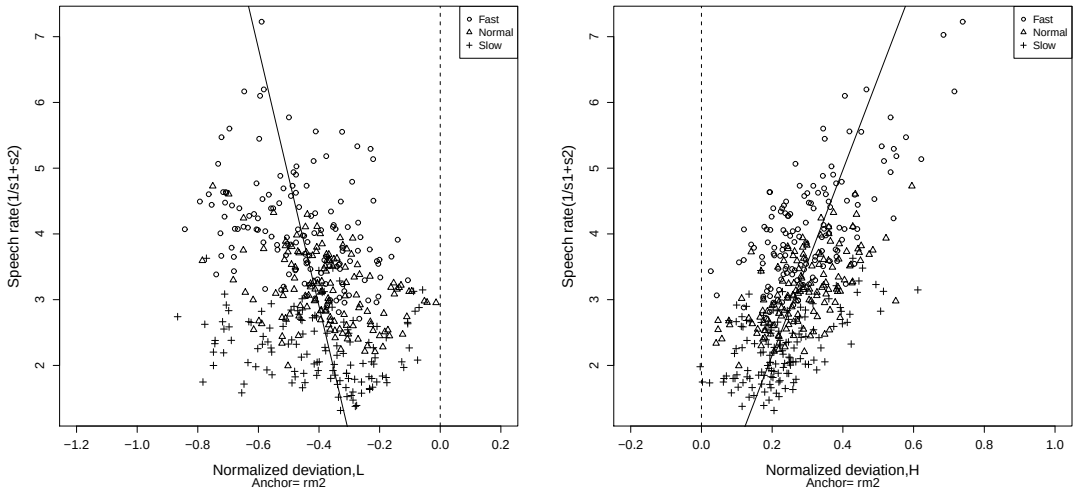
(b) Alignment of H

Figure 4-26: Alignment of L and H: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$. First approximation of the anchor.

Figure 4-26 shows the timing of L and H against their respective anchors. In English, the anchor was the same for both L and H, the middle of the second rime. L occurs before this anchor, and H occurs after the anchor. It is interesting that L and H tones are anchored with regard to the same position. A possible hypothesis is that what is in fact going on is to align some point between L and H to this anchor, and the locations of L and H tones are determined relative to that point. Such a point could be the maximum velocity point, because it is a property of F0 movements that may be significant for speech perception.

4.3.4 Effects of target duration

Along with the effects of segmental anchoring, systematic deviations from the anchoring point depending on local speech rate were observed. L is found earlier than the anchoring point in fast speech, later in slow speech (Figure 4-27a). H is found earlier with regard to the anchor in slow speech, later in fast speech (Figure 4-27b). That is, the rise starts earlier and terminates later in fast speech than in slow speech. As in other languages, these patterns can be interpreted as evidence for a target duration. Thus, we apply the Alignment-Duration model in the next section.



(a) L deviation from A_L (b) H deviation from A_H

Figure 4-27: Deviation of L and H: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. First approximation of the anchor.

4.3.5 Fitting the Alignment-Duration model

Anchor estimation

As in the other languages, tendencies to both segmental anchoring and target duration are observed in English. Thus, we apply the Alignment-Duration model, developed in Section 3.2 in Chapter 3. This model hypothesizes that the timing of L and H in English

involves three constraints: Align(L), Align(H), Duration. That is, there are constraints that require L and H tones to be segmentally aligned, and there is another constraint that requires the duration between L and H to be consistent. The actual timing is determined by minimizing the summed cost of violations of these conflicting constraints. The model for $T(L)$ and $T(H)$ are in (41) (repeated from (13) in Chapter 3).

(41) a. $T(L) = a(A_L - T(H)) + b + T(H)$
b. $T(H) = c(A_H - T(L)) + d + T(L)$

The positions of anchors (A_L and A_H) are estimated, based on (41). The procedure is the same as in the previous languages. In Section 4.3.3, the anchor was estimated from the correlation between a tone and a segmental position only. Given this model, the best anchoring point for L was the middle of the second rime. This anchor estimate is based on a different model from the one in (41), because (41) includes terms with the timing of the tone at the other end of a rise in predicting timing of a given tone. For the precise anchor estimate in this section, it is hypothesized that A_L is somewhere within the second rime, because the second syllable is stressed. A_L is divided into $v2 + p \cdot rime2$, where $v2$ is the beginning of the second vowel, $rime2$ is the duration of the second rime, p is the proportion into the rime. Then, (41-a) is expressed as (42), which is rearranged as in (43).

(42) $T(L) = a(v2 + p \cdot rime2 - T(H)) + b + T(H)$

(43) $T(L) = a(v2 - T(H)) + a \cdot p \cdot rime2 + b + T(H)$

To find the p value in (43), a mixed-effects model was fitted to the data with $T(L)$ as a dependent variable, $(v2 - T(H))$ and $rime2$ as fixed effects, by-speaker random intercepts, and by-speaker random slopes for $(v2 - T(H))$ and $rime2$. The coefficient for $(v2 - T(H))$ was 0.451, the coefficient for $rime2$ was -0.381, so $a = 0.451, a \cdot p = -0.381$. Thus, $p = -0.381/0.451 = -0.848$. This means that the estimate of A_L is $A_L = v2 - 0.848 \cdot rime2$. However, this means that A_L is not in the second rime, but precedes it. Therefore, we estimate A_L again from the first rime. That is, A_L is now substituted with $(v1 + p \cdot rime1)$, where $v1$ is the beginning of the first vowel, $rime1$ is the duration of the first rime, p is the proportion into the first rime. Then (41-a) is rearranged as (44).

(44) $T(L) = a(v1 - T(H)) + a \cdot p \cdot rime1 + b + T(H)$

Following the same procedure, $p = 0.392$ is obtained, so A_L is estimated as in (45).

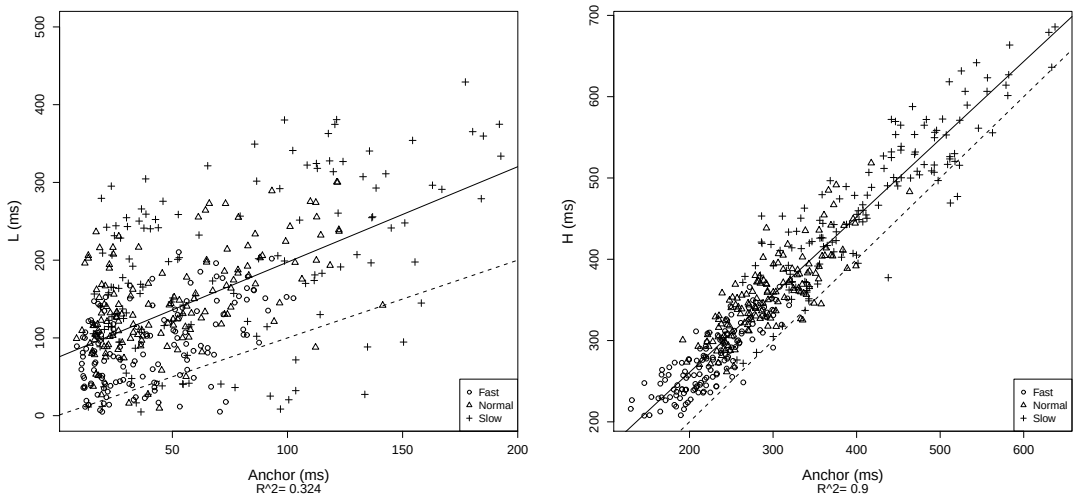
(45) $A_L = v1 + 0.392 \cdot rime1$

The location of A_L obtained in this way is closer to the actual timing of L than the previous anchor estimate ('rm2') in Section 4.3.3. The mean of A_L in (45) was 55 ms, the mean of timing of L was 141 ms, and the mean of 'rm2' was 280 ms.

A_H, the anchor for H, was estimated following the same procedure. It was assumed that A_H is in the second rime. In (41-b), A_H is substituted with $v2 + p \cdot rime2$, where $v2$ is the beginning of the second vowel, $rime2$ is the duration of the second rime, and p is the proportion into the second rime. Following the by now familiar procedure, A_H is estimated

as in (46).

(46) $A_H = v2 + 0.736 \cdot rime2$



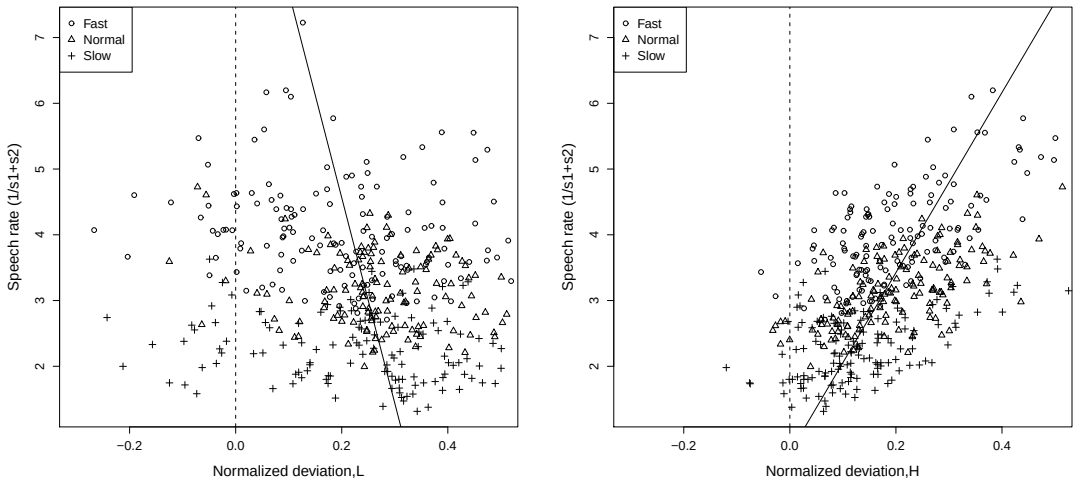
(a) Alignment of L (b) Alignment of H

Figure 4-28: Alignment of L and H with precise anchor estimates: (a) L against A_L, (b) H against A_H, the dashed line is $y = x$. Precise estimates of the anchor based on the Alignment-Duration model.

The precise estimates of the anchors based on the AD model move the anchors closer to the actual occurrence of L and H tones. Figure 4-28 shows the timing of L and H against their respective anchors, replotted based on the precise estimates of the anchor obtained in this section. Figure 4-29 shows the deviation of the tones from these anchors. In Figure 4-28a, the regression line and the $y = x$ line never come very close to each other. This is also true in Figure 4-28b, but the L tones were farther from the $y = x$ line than the H tones. That is, the estimate of A_L is farther from the actual timing of L, than the estimate of A_H is from the actual timing of H. This means that in the deviation plots in Figure 4-29, the regression line for L deviations does not cross the location of A_L (the dashed vertical line), whereas the regression line for H tone deviations almost meets the location of A_H (the dashed vertical line).

Constraint weights

With the anchor estimates in (45) and (46), the weights of Align(L), Align(H), and Duration were computed, following the same procedure as in the other languages. To find the coefficients in the expressions in (41), mixed- effects models were fitted. For (41-a), a mixed-effects model was fitted to the data with $T(L)$ as a dependent variable, $(A_L - H)$ as a fixed effect, by-speaker random intercepts, by-speaker random slopes for $(A_L - H)$, and offset of $T(H)$. A similar mixed model was fitted for $T(H)$. From the fitted mixed models, the con-



(a) L deviation from A_L

(b) H deviation from A_H

Figure 4-29: Deviation of L and H with precise anchor estimates: (a) L deviation from A_L, (b) H deviation from A_H, the dashed line is the position of the anchor. Precise estimates of the anchor based on the Alignment-Duration model.

straint weights were computed: $w_L = 0.16, w_H = 0.72, w_D = 0.12$. The constraint weights reflect the observation that alignment of H peaks is relatively stable in English.

The target duration D values were $D_L = 116, D_H = 562$. The significance of the difference between D_L and D_H was tested by simulation of generating probable intercept and slope pairs based on the fitted models, as in other languages (following the same procedure in Section 3.2.4 in Chapter 3). The 95% confidence interval of the D_L value was 32 to 324. The 95% confidence interval of D_H was 415 to 792. The confidence intervals of the D_L and D_H values did not overlap. The difference ($D_H - D_L$) was significantly different from zero, because the 95% confidence interval of the difference did not include 0 (205 to 682). The difference indicates a problem, so a solution will be discussed in the next section.

4.3.6 Elbows as L tones

The significant difference between D_L and D_H in the previous section indicates a problem somewhere in the model or in the procedure, since the AD model posits a singly value for D. To address this issue, we examine another F0 event that may be more relevant for the English pitch accent rises: the inflection point (elbow). In the previous section, F0 minima and maxima were taken to correspond to L and H. However, the average contours shown in Figure 4-25 suggest that F0 minima may not be reliable locations of L tones, because there is a plateau or a shallow rise before the start of the fast rise, similar to the case of Mandarin. As in Mandarin, the inflection points may be more accurate measures for L tones in English. Thus, the inflection points replaced F0 minima for L. The inflection points are used to locate the start of the fast portion of the rise. By definition (as explained in Section 2.2.3 in Chapter 2), a sigmoid rise is characterized by having the steepest regression line second, thus the

intersection between the first and second regression lines is the start of the fast rise, which corresponds to the first inflection point. For scooped rises, the fastest line is the third, so the intersection between the second and the third regression lines is the start of the fast rise, which is the second inflection point. Thus, a new vector representing L was created as follows: for sigmoid rises, L was the first elbow; for scooped rises, L was the second elbow; for domed rises, L was the F0 minimum. This is the same procedure used in Mandarin (Section 4.2.1). F0 maxima were still representing H tones. For simplicity, the beginning of the fast rise will be referred to as the "elbows", and the L tone represented by the elbows as "elbow L", from now on.

With this modification, the same model was fitted following the same procedure as before. The anchoring points were re-estimated: $A_L = v1{+}0.19{\cdot}rime1$ and $A_H = v2{+}0.93{\cdot}rime2$. The Alignment-Duration model was fitted to the data. The computed constraint weights were $w_L = 0.15, w_H = 0.49, w_D = 0.36$, and the D values were $D_L = 55, D_H = 176$. Simulation yielded 78 to 166 for the 95% confidence interval for the difference between D_L and D_H. This means that D_L and D_H are still significantly different. Thus, taking L to correspond to the lower elbow did not fix the problem. Yet, the deviance values were much lower when elbows were used as L than when F0 minima were used as L, as shown in Table 4.13. Thus, we will use elbows as the correlates of the L tones from now on.

Table 4.13: Deviance

	T(L)	T(H)
L = F0 minima	4564	3885
L = elbows	3714	3409

Table 4.14: Deviance

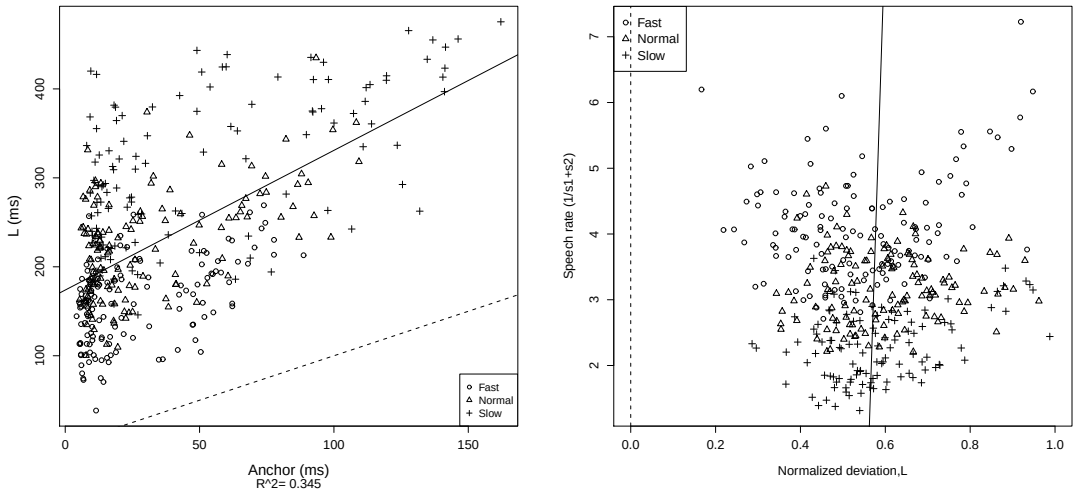
	T(L)	T(H)
The IA model	4295	3729
The AD model	3714	3409

In addition, the Alignment-Duration model turns out to be better than the Independent-Alignment model. The deviance values for the two models are shown in Table 4.14. In both cases, the elbows are used as L tones. Under the AD model, the timing of L and H tones is determined by a compromise between their anchors and the duration from L to H. On the other hand, under the IA model, L and H tones are independent of each other. That is, timing of a tone is determined by a compromise between its segmental anchor and a target duration from the phrase onset, independently of the other tones (as already explained in Section 4.1.6). Thus, the results in Table 4.14 means that the L and H tones are not independent of each other.

4.3.7 The L-offset model

In the previous section, we have shown that the model with elbows as L was better than the model with F0 minima as L in terms of the deviance. However, the D values did not converge in the better model, either. Looking at the data, it seems that the problem lies in the timing of the elbows relative to the anchor. Figure 4-30 shows the alignment and deviation plots of elbow L's. In Figure 4-30a, L is plotted against A_L. In Figure 4-30b, normalized deviations of L are plotted against speech rate. A_L is the anchor estimate based on the AD model, which was in the previous section (i.e. $A_L = v1{+}0.19{\cdot}rime1$). The timing of A_L is far from the actual occurrence of the elbow L's. The mean of A_L was 36 ms, the

mean of the elbow L was 229 ms, and the mean of the difference between A_L and actual L was 194 ms.



(a) Alignment of L

(b) L deviation from A_L

Figure 4-30: Alignment and deviation of the elbow L: (a) L against A_L. The dashed line is the $y = x$ line. (b) L deviation from A_L. The dashed line is the location of the anchor.

This may mean that the anchor of L is actually some offset from the estimated A_L. If so, a more accurate anchor $A_L\prime$ can be expressed as $A_L\prime = A_L + k$, where A_L is the previously estimated anchor, and k is an offset from A_L. Conceptually, this approach models the idea that the F0 contour shape of the English rising pitch accent consists of a plateau followed by a fast rise. The F0 minima are measured somewhere along the plateau, so the location of F0 minima is subject to random fluctuations or segmental perturbations imposed on what is really intended to be a plateau. On the other hand, the elbows are more systematically aligned, so the model with elbow L's had a lower deviance than the model with F0 minima L's. The anchor of the elbow L is estimated as a point much earlier than the actual occurrence of the L. The anchor is found on the plateau. Thus, the L offset model implies that the elbows are aligned with regard to some point during the plateau, but it is delayed by k from that point, producing a plateau before the fast rise. The duration of the plateau varies depending on the speech rate. The faster the speech rate, the shorter the plateau. The difference between A_L and L in Figure 4-30a indicates the duration of the plateau: A_L is a point early in the plateau, and L (elbow) is the beginning of the fast rise.

For this reason, we expect that the adjustment of A_L by an offset k will improve the AD model. In the L model in the usual AD model (47-a), the alignment target A_L is replaced by $A_L + k$, as in (47-b). The fitted models are the same except that now the intercept of the L model has a different interpretation. The intercept of the L model (b) is now $D(a-1)+a\cdot k$, instead of $D(a-1)$.

(47) a. $T(L) = a(A_L - H) + H + D(a-1)$

b. $T(L) = a(A_L + k - H) + H + D(a-1) = a(A_L - H) + H + D(a-1) + a \cdot k$

The slope and intercept of the L and H models of the AD model are the same as in the mixed-effects models fitted in the previous section. According to the model, $a = 0.293, b = -39$. The slope (c) and intercept (d) of the H model are $c = 0.574, d = 75$. The weights are the same as before: $w_L = 0.15, w_H = 0.49, w_D = 0.36$. The D value from the H model is $D_H = d/(1-c) = 176ms$. The D value from the L model cannot be computed directly because there is another unknown, k, in the intercept of the L model. To test whether the proposed model yields a reasonable value for k, D_H is plugged into the intercept of the L model as the D value, $b = D(a-1) + a \cdot k$. From this, $k = 292$. The confidence interval of k is 181 ms to 418 ms.

The resulting k value seems reasonable. k is the offset value of the actual L target from A_L, i.e. it should be close to the difference between A_L and L. The mean of $(A_L - L)$ in the data is $194ms$. By speech rate, the means are 136 ms in fast speech, 195 ms in normal speech, and 264 ms in slow speech. Because w_L is low and the alignment of L is not very strict, the k value cannot be estimated accurately. Considering this difficulty with the estimation, the obtained k value is fairly close to the expected value.

4.4 Summary of the chapter

In this chapter, we examined three languages with tones of varying phonological statuses of tones: Tokyo Japanese (lexical pitch accent), Mandarin Chinese (lexical tone), and English (Intonational pitch accent). Tones in these languages are different from boundary tones in Seoul Korean in that the tones are contrastive or prominence-lending. The common hypothesis we tested across languages was that contrastive or prominence-lending tones will show a stricter alignment pattern than boundary tones. Contrastive or prominence-lending tones include the tones in Japanese pitch accents, Mandarin Rising tone, English intonational pitch accents. Boundary tones include the tones in Seoul Korean Accentual Phrases and unaccented words in Tokyo Japanese were boundary tones. Another hypothesis was that alignment patterns will vary depending on phonological context: word-medial vs. word-final context of Japanese accentual peaks, lexical tone vs. neutral tone context of the Mandarin Rising tone.

In all three languages we examined in this chapter, both tendencies to maintain segmental anchoring and shape target (duration) were observed, as in the case of Seoul Korean in Chapter 2. They all showed a systematic relation between speech rate and deviation of tones. At a fast speech rate, peaks tend to occur later; at a slow rate, peaks tend to occur earlier. This tendency is reversed for L tones: at a fast rate, F0 minima tend to occur earlier; later at a slow rate (in most languages, except for Japanese medial-accented words: Figure 4-7). Based on these findings, we attempted to explain the differences in the alignment patterns in quantitatively precise terms, using the Alignment-Duration model framework proposed in Chapter 3. That is, it was hypothesized that the timing of L/H tones in these languages is determined by the interaction of alignment and duration constraints: Align(L), Align(H), and Duration.

However, some adjustments had to be made when the model was applied to different languages. The adjustments involved adding constraints or selecting different measure-

ment points. In Japanese, an additional constraint was needed to model the asymmetry between L and H tones. The IA model was better than the AD model for the timing of L, while the AD model was better for the timing of H. This means that the timing of L is not significantly affected by the timing of H, but the timing of H is dependent on the timing of L. Thus, the Delay L constraint, which requires L to occur at a fixed distance from the phrase onset, was added. DelayL influences the timing of L independently of the timing of H. DelayL allows better modeling of the variability by putting the effective L target near the middle of the range between the anchor of L (the beginning of the first vowel) and the DelayL target (the end of the first mora).

The interval where the Duration constraint is applied is defined differently in different languages. Initially, we started with the assumption that L and H tones correspond to F0 minima and F0 maxima. This assumption yielded meaningful results for Seoul Korean and Tokyo Japanese: the Duration constraint holds from F0 minimum to F0 maximum. Different measures had to be used in Mandarin and English. In Mandarin and English, elbows were better estimates for the L tones (the beginning of the fast rise), instead of F0 minima. F0 maxima were still the best reflection of the H tones. Thus, in these languages, the AD model holds between the elbow L and the H peak, rather than between the F0 minimum and the F0 maximum.

Although direct statistical tests of the differences between languages are not available because of these language-specific differences, the weights computed for each language can inform us about the cross-linguistic differences. Table 4.15 summarizes the constraint weights computed in each language. The weights for the Japanese accented words are before the adjustment with w_B, to make it easier to compare the proportions between low-weighted constraints.

Table 4.15: Constraint weights by language and phonological conditions

Language	Status	Condition	w_L	w_H	w_D	w_k
Korean	Boundary		0.54	0.35	0.11	0
Japanese	Lexical	medial-accented	0.04	0.77	0.06	0.13
	Lexical	final-accented	0	0.96	0.01	0.03
	Boundary	unaccented	0	0.57	0.08	0.36
Mandarin	Lexical	lexical-tone context	0.29	0.35	0.36	0
	Lexical	neutral-tone context	0.32	0.25	0.43	0
English	Lexical		0.15	0.49	0.36	0

Our hypothesis was that lexical tones will be aligned more strictly than boundary tones, i.e. lexical tones will have a higher w_H than boundary tones. This is observed in Japanese accented vs. unaccented words: w_H is higher in accented words (0.77, 0.96) than in unaccented words (0.57).

However, it is not straightforward to interpret differences in the absolute w_H values across languages, due to a complication caused by w_D values. The complication is that in languages where contour tones are contrastive, w_D values are higher than languages without such contrasts. In Mandarin and English, contour tones are contrastive: the Rising tone and the High tone are contrastive in Mandarin, and L+H* and H* are contrastive in

English ("contour tone languages"). On the other hand, in Japanese and Korean, contour tones are not contrastive ("level tone languages"). We may expect that if contour tones are contrastive, the shape of the pitch movement will be relatively stable. Duration of pitch rising movements is one of the F0 shape properties, so the weight of the Duration constraint w_D is expected to be higher in the contour tone languages than in the level tone languages. This is what is observed, as shown in Table 4.15. When w_D is high, w_H becomes relatively low because the weights are relative and sum to 1. Consequently, due to high w_D values, these languages have w_H values equal or lower (because the constraint weights are relative) than those in the level tone languages, e.g. w_H is 0.35 both Korean and Mandarin (lexical tone context). However, it is unclear whether peak alignment is equally important in Korean and Mandarin. Furthermore, the correlates of L tones are different in Korean and Mandarin: F0 minima in Korean and elbows in Mandarin.

Due to these factors that depend on the phonological status of contour tones in languages, w_H values cannot be directly compared across languages. w_H can be meaningfully compared only when other conditions are equal. That is, within Japanese, w_H is higher for lexical tones than boundary tones, which means that peaks are more strictly aligned if the tone is lexical. Within Mandarin, w_H is higher in the lexical tone context than in the neutral tone context.

Contour tones vs. level tones

As mentioned, the weight of the Duration constraint w_D is relatively higher in Mandarin and English (contour tone languages) compared to Korean and Japanese (level tone languages). This means that in Mandarin and English, the duration of a rise remains relatively stable under the changes in segmental duration, more than in Korean and Japanese. This is a reasonable distinction given that contour tones are contrastive in Mandarin and English. The plots in Figure 4-31 directly show another difference between the two types of languages. In each language, the first panels (Figure 4-31a, 4-31d,4-31g,4-31j) show a plot of the duration between L and H, $(H - L)$ against the duration between the anchors $(A_H - A_L)$. The solid lines are regression lines, and the dashed lines are the $y = x$ lines. In all languages, the duration $(H - L)$ does not change as much as the duration between anchors $(A_H - A_L)$ changes (the slope of the regression line is less than 1). This means that there is a tendency to maintain a certain duration of a rise, rather than tones strictly following the positions of the anchors. However, there is a clear difference between Korean/Japanese and Mandarin/English. $(H - L)$ is much more resistant to change in Mandarin and English than in Korean and Japanese. The weight w_D of the Duration constraint reflects these differences. The higher w_D in Mandarin and English shows that the Duration constraint is important, so rise duration is relatively stable under changes in segmental duration.

The panels in the second column (Figures 4-31b, 4-31e, 4-31h, 4-31k) show $(H - L)$ against $(H - A_L)$. In all languages, $(H - L)$ is more consistent than $(H - A_L)$ (i.e. the slope of the regression line is less than 1). This is due to L deviation from A_L. The third panels (Figure 4-31c, 4-31f,4-31i, 4-31l) show $(H - L)$ against $(A_H - L)$. $(H - L)$ is more consistent than $(A_H - L)$. This is due to H deviation from A_H. In all plots, $(H - L)$ is more resistant to change in Mandarin and English than in Korean and Japanese.

Although $(H - L)$ is resistant to change in Mandarin and English, it is not invariant, i.e. the slope of the regression line is significantly different from zero. Mandarin has been described with dynamic targets (Xu, 1998, 2005), but if the language has dynamic targets only (rise, fall) and without static targets (L, H), $(H - L)$ should be near invariant, but this is not what is observed. The same applies to English. This means that a better model of tonal timing is one with both dynamic and static targets rather than one with static targets or dynamic targets alone. That is, dynamic and static targets both exist in languages and are realized according to their relative importance. Which targets are more important is reflected in the relative weights of the relevant constraints in a given language. Considering the relative constraint weights and the stability of $(H-L)$, static targets are more important than dynamic targets in Japanese and Seoul Korean, whereas dynamic targets are of equal or greater importance than static targets in Mandarin and English.

Further evidence for the distinction between contour tone languages and level tone languages is that the former languages show stepwise movements in extremely slow speech rates, as shown in Figure 4-32. This kind of movement is found in slow speech in Korean and Japanese. That is, each syllable is realized with a long plateau F0 target with a rapid transition made at the syllable boundary (typically during the consonant). This pattern is found in all of the female speakers. In Seoul Korean, two male speakers (B1, B4) show a similar pattern, but instead of plateaux, they have small rises in every syllable. Other male speakers were very monotonous in slow speech, so in many cases it is hard to see any meaningful differences between the pitch levels of the first syllable and the second syllable. A similar stepwise pattern is reported in Buli, a Gur language that contrasts high, mid, and low tones (Akanlig-Pare and Kenstowicz, 2003). In Buli, the syllable is the tone-bearing unit. Each syllable is designated with a pitch target, and this pitch stretches over the entire rime resulting in a plateau. Transitions between pitch targets are made rapidly during the consonant portion at the syllable boundary. Likewise, in Korean and Japanese, syllables or morae are tone-bearing units, so in slow speech, speakers extend each syllable and maintain the same pitch throughout a syllable, producing a plateau for each syllable.

Stepwise F0 movements are never found in English or Mandarin. This may mean that slope targets are more important in languages such as English and Mandarin. That is, the slope of the rise remains relatively stable across speech rate in these languages. On the other hand, in Korean and Japanese, the slope target is less important, so it is affected by changes in segmental duration. Thus, in slow speech, the slope would have been too shallow if alignment is to be maintained (because the duration between the anchors are too great). After a certain threshold, the slope cannot get shallower anymore, and instead a stepwise movement is produced. On the other hand, in Mandarin and English, the pitch transition itself remains relatively stable across speech rates.

In this dissertation, the only aspects of F0 shape that we have analyzed are duration targets, but the stepwise pattern shows that the shape target may be generalized to include the slope of the rise. That is, there is a tendency to maintain not just the duration, but also the slope of a rise. The tendency to maintain a target slope is more important in the languages that do not show a stepwise pattern (Mandarin, English) than in the languages that show a stepwise pattern when segmental duration increases (Korean, Japanese). In

the languages where a target slope is relatively important, it is important to produce an audible pitch transition. Previous experimental studies have claimed that Mandarin tones are associated with ideal pitch targets which can be static [high], [low] or dynamic [rise] [fall] Xu (1998, 2005). Our results also point to the possibility that English rising pitch movements have dynamic targets. On the other hand, audible pitch transitions do not seem to be crucial in Japanese and Korean. In the tones in Japanese and Korean, pitch transition (slope) is a low-weighted target in terms of our model.

Another piece of evidence pointing to the fact that in Korean and Japanese, static targets are more important than dynamic targets, while dynamic targets are more important in Mandarin and English is found in the shape of the rises. In Korean and Japanese, it seems that the shape of rises is not regulated much, thus the shape can vary among sigmoid, scooped, and domed, depending on time pressure. In Korean and Japanese, domed shapes are found (mostly in fast speech), whereas domed shapes are absent in Mandarin, and very rare (1%) in English, even in fast speech. Thus, in Mandarin and English, pitch transition from L to H is specified for shape, whereas in Korean and Japanese, pitch transitions are not specified for shape, but have a sigmoid shape by default (considering that this shape was the most common). In this sense, pitch transitions in these languages are movements from one level target to the next, although the durations between the level targets are regulated. On the other hand, in Mandarin and English, domed shapes are strongly disfavored. It would be interesting to conduct a perceptual experiment to see whether speakers of Mandarin and English would find that domed rises sound strange, but Korean or Japanese speakers would not.

In summary, the timing of L and H tones in the three languages examined in this chapter are determined by both segmental anchoring and duration constraints, although details vary in different languages. The alignment patterns of the tones differ depending on phonological status (lexical, phrasal) and phonological context. Lexical tones are more strictly aligned than phrasal tones; tonal alignment is stricter when it is followed by a word boundary or another lexical tone. In addition, the dynamic targets are more important in languages where contour tones are contrastive (Mandarin, English) than in languages where they are not (Korean, Japanese). These differences are reflected in the relative weights of the weighted-constraint model of F0 movements.

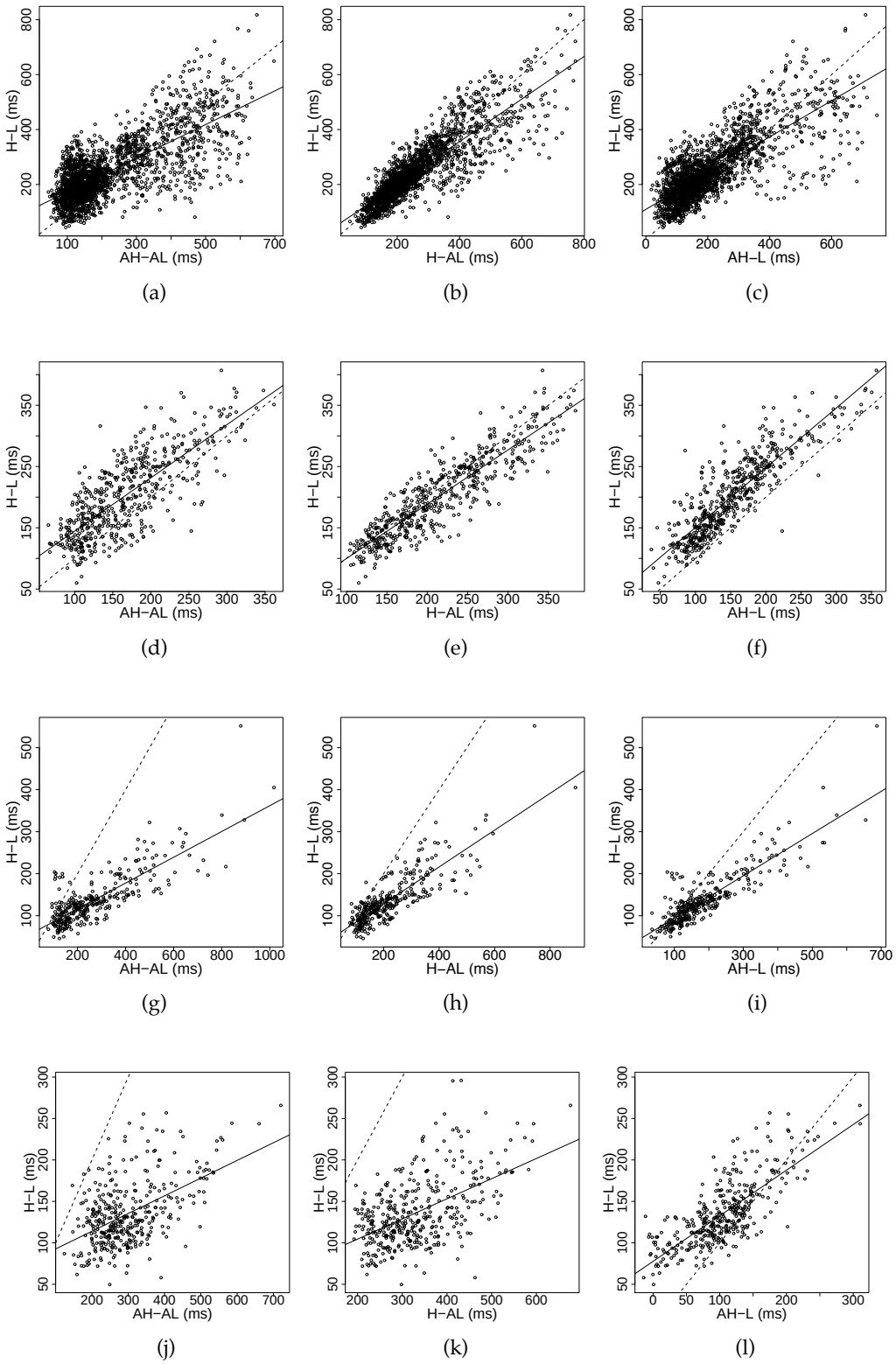


Figure 4-31: H-L plotted against AH-AL, H-AL, AH-L in Korean (a),(b),(c); Japanese (d),(e),(f); Chinese (g),(h),(i); English (j),(k),(l)

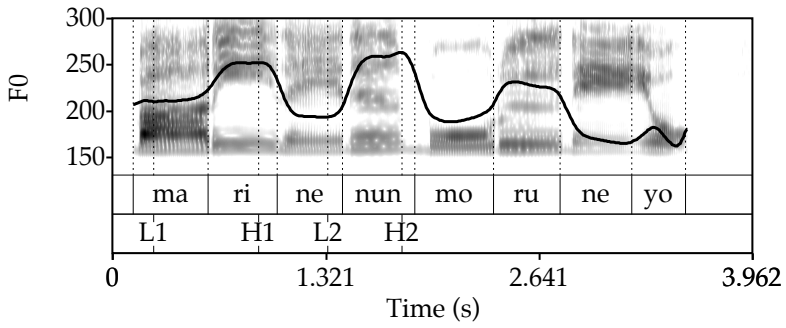


Figure 4-32: Stepwise movements in the Seoul Korean LHLH AP. Speaker A4, slow speech

Chapter 5

Conclusion

The main research question investigated in this dissertation was how underlying tonal representations are phonetically implemented, depending on the phonological nature of tones. The proposed model resolves two conflicting previous views on tonal timing: the Segmental Anchoring Hypothesis and the Constant Duration Hypothesis. Under the segmental anchoring hypothesis, the beginning and end of a rising pitch movement are aligned with respect to segments. Under this approach, tones are independent of each other, and shape properties such as duration and slope are derivable from the locations of level targets. This approach is at odds with constant shape approaches, e.g. the IPO model of Dutch intonation which describes F0 movements in terms of fixed shapes, e.g. 'Type 1 Rise' has a fixed duration of 120 ms.

Recently, a number of studies have supported the Segmental Anchoring Hypothesis (Arvaniti et al., 1998; Ladd et al., 1999) in a number of languages. Experimental studies have shown that the duration of rises changes depending on speech rate. That is, the duration of a rise between two tones (L and H) decreases when segmental duration is shorter, so it cannot be constant (Ladd et al., 1999; Dilley et al., 2005). On the other hand, segmental locations of tones seemed to remain relatively stable. Based on these findings, the SAH literature has claimed that L and H tones are aligned to segmental anchors, independently of each other. Because the duration of rises change, models that assume a fixed duration of rise have been rejected (such as the IPO model of Dutch intonation, 't Hart and Cohen (1973); 't Hart and Collier (1975); 't Hart et al. (1990)). It has been believed that one of them is universally correct, and the other should be rejected (Dilley et al., 2005).

Against this backdrop, one of the main contributions of this dissertation is to have shown that tendencies to both segmental anchoring and target duration are simultaneously observed. Furthermore, these experimental findings are modeled in a constraint-based framework, which provides a compromise solution to reconcile the two conflicting views. The effects of both views can co-exist if the conflicting hypotheses are interpreted as violable constraints, rather than inviolable principles.

The constraint-based approach provides us with a framework to compare cross-linguistic patterns in tonal alignment, because it allows parametric variation with common constraints. Thus, in the previous chapters, we have examined the alignment patterns of tones in languages with varying phonological status. Most of the languages studied in the Seg-

mental Anchoring literature showed strong segmental anchoring effects, but most of the tones are lexically-contrastive or prominence-lending. So, strong segmental anchoring is somewhat expected. The present research started with the question whether segmental anchoring will be different if tones with different phonological status are examined. The phrase-initial rise in the Seoul dialect of Korean was appropriate for testing this hypothesis, because the tones in Seoul Korean are not lexically contrastive, but they only mark the beginning and end of the Accentual Phrase. Thus, in Chapter 2, we thoroughly examined the timing and scaling of the four tones in the Seoul Korean LHLH Accentual Phrase. The experimental results showed that tonal alignment cannot be explained by segmental anchoring alone. There was a systematic deviation of the H peaks from the anchor, depending on speech rate: the faster the rate, the later the peak. At the same time, the L tones also deviate, but in the opposite direction of H deviation: the faster the rate, the earlier the L. That is, when speaking fast, the rise starts relatively earlier and terminates later with regard to the target alignment points. This was taken as evidence for a tendency to maintain a target duration. Based on this, the Alignment-Duration model was developed in Chapter 3. The model consists of constraints for segmental alignment and target duration. The weights reflect the alignment pattern: for example, L in Seoul Korean shows stricter alignment than H, thus the weight of L alignment is higher than the weight of H alignment.

Given these results, in Chapter 4, three other languages with varying phonological status of tones were examined: Tokyo Japanese (lexical pitch accent), Mandarin Chinese (lexical tone), English (intonational pitch accent). The experiments aimed to test the two hypotheses: first, that the constraints for alignment and target duration exist in languages with various tonal phonologies, including languages which have been reported to exhibit segmental anchoring, and second, the relative weights of the constraints reflect differences in the phonetic realization of tones with different phonological status (lexical or boundary) and phonological context (word-medial/final in Japanese or lexical/neutral tone in Mandarin). In Tokyo Japanese, pitch accents are contrastive, so they are specified in the lexicon. In Mandarin, most syllables are specified for contrastive tones, thus misalignment would result in the intrusion into the tones of neighboring syllables. In English, pitch accents are realized on the prominent syllable in a word. In addition, in both Mandarin and English, rises contrast with high tones: Tone 1 vs. Tone 2 in Mandarin, H* vs. L+H* in English, whereas there is no such distinction in Japanese and Korean. Thus, one might expect the former languages place more weight on shape properties such as the slope or duration of the rise which are likely to differentiate high and rising tones.

The experimental results confirmed the hypothesis that languages have constraints for both alignment and target duration. In each language examined, tendencies toward both segmental anchoring and target duration were observed, despite the varying nature of the tones in these languages. At the same time, the degree to which theses constraints are satisfied/violated is different in different languages. For example, in Mandarin, the alignment and duration constraints have similar weights, confirming the previous observations that both alignment and tonal shape are relatively stable in Mandarin. In the Mandarin Rising tone, aligning both L and H tones is important, and maintaining a target duration is also as important as keeping alignment.

The constraint weights reflected differences in tonal alignment that depend on phonological status and context of tones. Tones that are lexically contrastive (lexical pitch accents in Japanese, lexical tones in Mandarin) or prominence-lending (intonational pitch accents in English) were timed close to their anchors which are near or in the tone-bearing units that the tones are phonologically associated with. The alignment was stricter in lexically-specified tones than in tones that are phrasal (phrase-initial L and H in the Seoul Korean Accentual Phrase and in Japanese unaccented words). Phonological context was also a significant factor: tones were more strictly aligned in the context of another lexical tone (Mandarin) or a word boundary (Japanese).

Our experimental results argue against the basic assumption of the Segmental Anchoring Hypothesis that tones are independent of each other. We found that the relation between L and H tones of a rising pitch movement is as significant as the relation between tones and segments. In the proposed model, the relations between tones and segments are regulated by the Alignment constraints, and the relations between tones are regulated by the Duration constraints. The actual timing of a tone is determined by both the segmental and tonal relations of the tone; the relative importance of these relations are reflected in the constraint weights. The actual timing is the weighted average between the alignment target and and the point that would satisfy the rise duration target.

In summary, the major contributions of this dissertation are summarized as follows: first, it showed that tendencies to both segmental anchoring and target duration co-exist; second, it showed that tones that comprise a pitch movement are not independent of each other. The model with dependencies between tones (the Alignment-Duration model) was better than the model with independent tones (the Independent-Alignment model) in almost all cases; third, the conflicting hypotheses were translated into violable constraints; fourth, the proposed framework explicates cross-linguistic differences through relative weights of constraints. Two relevant issues will be discussed before closing this chapter.

Analyzing the rise shape

The Segmental Anchoring approach decomposes a rising F0 movement into L and H tones, arguing that the beginning (L) and ending (H) points of a rise are segmentally aligned. However, defining L and H tones is not straightforward: the acoustic correlates for L and H tones differ across languages. We described an F0 rising movement in terms of a few more sub-components: F0 extrema (minima,maxima), inflection points (lower and upper elbows), and maximum velocity points. Rising movements were decomposed into three regression lines, by fitting three-piece linear regression. From the F0 minimum to the lower elbow is the first line, from the first elbow to the second elbow is the second line, and from the second elbow to F0 maximum is the third line. Depending on the slopes of these lines, the shape of a rise is categorized: dome if the first line is the steepest, sigmoid if the second line is the steepest, and scoop if the third line is the steepest. The two intersections of the regression lines are the elbows. The inflection points were used to locate the beginning of the fast rise. That is, the beginning of the fast rise is the F0 minimum if the shape is domed, the first elbow if the shape is sigmoid, and the second elbow if the shape is scooped.

We have shown that different languages have different patterns of alignment of these

sub-parts of F0 rises. In Seoul Korean, the Duration constraint holds between F0 minima and F0 maxima. In Japanese, our modeling results show that F0 maxima are related to F0 minima. On the other hand, the inflection points at the beginning of the fast rise, are closely related to F0 maxima. However, which point is the beginning of a rising movement is language-specific: F0 minimum or elbow. In Mandarin and English, F0 minima are often found in a long plateau before the actual start of the fast rise; this can vary depending on factors such as language and speech rate. Thus, in Mandarin and English, elbows were used to indicate the location of the L tone. Xu (1998) also used a similar point, the maximum acceleration point, to indicate the location of L tone in the Mandarin Rising tone, and the absolute F0 minima were considered irrelevant for the Rising tone. Thus, whether the L and H tones are independent or dependent depends on which F0 events are considered as the reflexes of L and H tones. The level-target literature argues that a Rising tone can be decomposed into L and H tones, but the acoustic correlates of L and H tones seem to vary across languages, so they have to be defined language-specifically.

Using the three-piece regression, we were also able to classify the shapes of rises in a quantitative rather than impressionistic fashion. The shape of a rise is classified into sigmoid, dome, and scoop categories. In Seoul Korean and Japanese, the shapes are almost free variants. The majority of the rises are sigmoid, domed rises are found more in fast speech, and scooped rises are found more in slow speech. In these languages, free variation among the three shapes is allowed. This is different from Mandarin and English. Domed shapes do not exist among Mandarin Rising tone. The Mandarin Rising tone was either sigmoid or scooped rising tones. English was similar (only less than 1% were domes). Thus, the shape of rises is more restricted in English and Mandarin than in Japanese and Korean. This shows up partly in w_D, but additional shape parameters are required for a full analysis.

Phonological representations

Although we do not make a strong claim about phonological representations, we can discuss what phonological representations are compatible with our model. In Chapter 4, we have shown that phonetic realization of tones depends on the phonological status of the tones: whether they are lexically-contrastive/prominence-lending or phrasal boundary-marking. That is, lexically contrastive (Japanese accented words, Mandarin) or prominence lending tones (English) show stricter alignment patterns than boundary tones (Japanese unaccented words, Seoul Korean). In Japanese, for example, word-medial accented and unaccented words have different w_H values (0.77, 0.57 respectively). Thus, phonological representations that are compatible with the proposed model should be able to distinguish lexically-contrastive or prominence-lending tones from boundary tones, so that the phonetic implementation grammar can see which w_H is applicable for a given tone.

We propose that the difference in the phonological status of the tones is indicated by where the tones are associated. That is, lexically contrastive or prominence lending tones are directly associated with the relevant syllables/morae, but the boundary tones are associated with phrasal boundaries. We illustrate such examples with Japanese accented and unaccented words, as shown in Figure 5-1.

(a) Accented words (b) Unaccented words

Figure 5-1: Suggested phonological representations: (a) a lexically-contrastive tone (b) a boundary tone

In 5-1a, the lexical pitch accent H is associated with the second mora, and the initial rise LH is associated at the phrase- initial position. On the other hand, in the unaccented word in 5-1b, there is only a boundary LH. This approach is similar to what has been proposed for tonal representations for Japanese in Pierrehumbert and Beckman (1988). In Pierrehumbert and Beckman (1988), L and H tones at the beginning of a phrase are associated with the higher phrase nodes, but contrastive HL tones are associated with the accented morae only, not with the phrasal nodes. Thus, in their system, differences between boundary tones and contrastive tones are represented by different phonological association. Differences in phonological nature are differences in the association.

For phonetic implementation of the tones in Figure 5-1, a constraint requires the boundary LH to be realized on the first and second morae respectively (L on the first mora, H on the second mora), for both accented and unaccented words. In the unaccented case in Figure 5-1b, w_H of the unaccented words is applied (0.57). In the accented case in Figure 5-1a, the second mora has both boundary and lexical H's, but realization of the lexical H tone takes precedence over realization of the boundary H tone. So, w_H of the accented words is applied (0.77).

In the phonological representations using association lines, peak delay presents a problem. This is a phenomenon where F0 peak that realizes a H tone appears after the syllable/mora with which it is phonologically associated (English, Silverman and Pierrehumbert (1990); Spanish, Prieto et al. (1995); Mandarin, Xu (2001); Tokyo Japanese, Ishihara (2006)). To address this issue, the Autosegmental Phonology literature has suggested secondary association in order to indicate precise locations of L and H tones. e.g. L is secondarily associated at the left edge of the syllable, and H is secondarily associated with the following vowel (Pierrehumbert and Beckman, 1988; Gussenhoven, 2000; Grice et al., 2000).

However, Ladd (2004) disagrees with this approach on the grounds that using secondary association brings "proliferation of phonological representations for subtly different phonetic details between languages or between language varieties". Instead, he argues that "the fine phonetic detail of segmental anchoring is not a matter of secondary association after all, but of quantitative language-specific phonetic detail in the realization of phonological categories" (Ladd, 2004). In the same vein, although we suggested the phonological representation in Figure 5-1 as one possible modification to the existing representation, they have limitations in representing fine-grained phonetic details. We have shown, throughout this dissertation, that the subtle phonetic details vary systematically depending on the phonological nature and context of the tones. We thus believe that differences in tonal

alignment are best described in terms of a quantitative phonetic realization grammar, such as the weighted-constraint model developed in this dissertation. The phonetic realization grammar provides precise language-specific details, and phonological representations do not have to be more complicated than the ones suggested in Figure 5-1.

Language-specific phonetic grammar

This dissertation has demonstrated that phonetic details of tonal timing vary cross-linguistically. In fact, the question of whether phonetic details are supplied by language-specific or universal rules originates from the very beginning of generative linguistic research. Chomsky and Halle (1968: 259) argue that detailed phonetic properties of speech signals are supplied by universal rules, so it is not necessary to include them in discrete phonological representations. Such properties include various aspects of phonetic realization (e.g. fronting of a back vowel after a coronal onset). Even if it is possible to transcribe all the properties of speech signal, the speaker-hearer's interpretation of the signal is more linguistically meaningful than directly observable physical properties of it (Chomsky and Halle, 1968: 294). Thus, phonological representations are discrete and "coarse-grained" (Flemming, 2001), leaving details of phonetic realization as the task of universal rules. For example, in the Autosegmental literature, representations of tones are L and H for low and high tones, which are assumed to be associated with syllables. The L and H representations do not tell us the precise location or pitch levels of H peaks. However, universal rules cannot explain the experimental findings in this dissertation, because tonal timing patterns are significantly different from language to language.

Studies have suggested that peak delays are due to physiological factors that arise when implementing a rising pitch movement, such as inertia (Xu, 1998, 2005). This approach incorporates an automatic phonetic realization component that reflects physiological implementation of tones into the grammar of segmental anchoring. However, we have seen that peak delay is in fact a controllable factor, which depends on the phonological status and context of the tones. In Japanese, the H peaks gradually deviate into the post-accentual mora if the accented mora is not word-final, but the H peaks stay within the mora if the accented mora is word-final, resulting in a stricter alignment pattern. Speakers are able to control alignment more tightly in that context. In Mandarin, the peak of the Rising tone shows a similar strict alignment pattern. If peak delay is just due to physiological limitations on pitch change, we should observe similar patterns across languages. However, we have shown that alignment patterns can vary depending on phonological contexts (Japanese word-medial/final) and languages (although strictly speaking, this should assume that the accelerations involved are similar across the languages which could differ as well as timing). Controllable factors are in principle subject to language-specific manipulations (Keating, 1985). The variations in alignment patterns across phonological contexts and languages confirm that peak alignment is a controllable linguistic factor.

An analogous example comes from vowel shortening before voiceless consonants. Languages show vowel shortening in this context (Chen, 1970), but the differences between the vowel and consonant durations vary depending on languages (Keating, 1985). For example, the difference is exaggerated in English, but Czech or Polish do not vary vowel dura-

tion according to the following consonant (Keating, 1979). This shows that vowel shortening in English is not physiologically determined, although physical factors influence vowel duration. Therefore, the vowel shortening rule cannot be part of a universal phonetic component, but language-specific phonetic grammars are necessary. Likewise, delayed peaks cannot be explained by imposing physical limitation rules on the Segmental Anchoring Hypothesis. Peak alignment is a controllable property, so it is a part of a phonetic realization grammar. Whether peaks are delayed or anticipated is not due to physiological constraints but due to grammar.

The weighted-constraint model of F0 movements developed in this dissertation provides a framework for the grammar of phonetic realization by allowing constraints common to many languages while at the same time parametrizing language-specific differences by relative weights of the constraints. Language specific phonetic details need to be described quantitatively, not symbolically (Ladd, 2004: 6). Atterer and Ladd (2004) pointed out that alignment patterns are subject to language-specific variation; they found that German speakers align rises later than in other languages previously studied, such as Greek, English, and Dutch. The variation is also found between dialects of the same language: Southern German shows later alignment than Northern German. Moreover, these native patterns of alignment are carried over to the German speakers' pronunciation of English. They claimed that such findings argue against interpreting cross-language alignment differences in terms of distinct patterns of phonological association, and in favor of describing them in terms of quantitative phonetic realization rules. The model with weighted-constraints for scalar representations (Flemming, 2001) is appropriate for this purpose, because it provides a quantitatively precise framework that uses scalar representations. The relative weights of alignment and duration constraints reflect differences in language-specific fine-grained phonetic details.

As a further development of the proposed model, it is worth investigating perceptual effects of the F0 features in order to relate the constraint weights to perceptual importance. The hypothesis is that the weights of pitch features reflect their perceptual effects. That is, highly-weighted features characterize crucial properties of a pitch movement and thus are perceptually more significant. Thus, violation of the highly-weighted targets should be avoided, at the expense of low-weighted targets. Such research may reveal a close connection between production and perception of speech, which can be modeled in a quantitative framework via the weighted constraints proposed in this dissertation.

In conclusion, the main contribution of this dissertation is that the proposed model interprets F0 features in terms of targets, rather than invariant properties. Insistence on identifying invariant properties has been the source of the long-running controversy on which is invariant of F0 movements, alignment or shape. Our answer to this question is that there is no invariant: everything is subject to change, but some features are more resistant to change than others, depending on their relative importance. The interaction of the targets are modeled in a cross-linguistically applicable, constraint-based framework. The proposed model makes quantitatively precise predictions by using weighted constraints for scalar representations, rather than discrete phonological representations. This dissertation also showed that the detailed phonetic implementation of tones is determined by

phonological status/context of the tones. This supports the view that phonetic details in tonal implementation are provided by language-specific phonetic realization grammars, rather than automatically determined by universal phonetic rules. This suggests the possibility of a difference in surface phonological representation that the phonetic implementation grammar can refer to. It is a task for future research to determine how this information is encoded so that the phonetic grammar can make use of it.

Appendix A

Speech materials

A.1 Seoul Korean

There were 36 target APs and 25 fillers of various lengths (13 5-syl APs, 6 6-syl APs, 6 3-syl APs). Fillers are not presented here.

	target phrase	carrier		
1	marinenɨn	morɨneyo	'Mary's family doesn't know'	마리네는 모르네요.
2	narinenɨn	nollaneyo	'Nary's family is surprised'	나리네는 놀라네요.
3	mamurinɨn	miruneyo	'Finishing is postponed'	마무리는 미루네요.
4	manuranɨn	mallineyo	'A wife stops it'	마누라는 말리네요.
5	monaminɨn	morɨneyo	'don't know Monami'	모나미는 모르네요.
6	memorinɨn	morɨneyo	'don't know memory'	메모리는 모르네요.
7	meronanɨn	morɨneyo	'don't know Merona'	메로나는 모르네요.
8	nonaranɨn	mallineyo	'No-dynasty stops it'	노나라는 말리네요.
9	remonanɨn	morɨneyo	'don't know Lemona'	레모나는 모르네요.
10	miri nenɨn	morɨneyo	'don't know milkyway'	미리내는 모르네요.
11	miminenɨn	miruneyo	'Mimi's family postpones'	미미네는 미루네요.
12	minarinɨn	mallineyo	'Dropwort is dried'	미나리는 말리네요.
13	minanenɨn	nollineyo	'Mina's family is surprised'	미나네는 놀리네요.
14	nunanenɨn	nollaneyo	'Sister's family is surprised'	누나네는 놀라네요.
15	manillanɨn	morɨneyo	'don't know Manila'	마닐라는 모르네요.
16	marɨmmonɨn	millineyo	'A lozenge is pushed'	마름모는 밀리네요.
17	norinnenɨn	morɨneyo	'don't know stench'	노린내는 모르네요.
18	amepanɨn	morɨneyo	'don't know ameba'	아메바는 모르네요.
19	inajəŋɨn	miruneyo	'Lee Na-Young postpones'	이나영은 미루네요.
20	əməninɨn	mallineyo	'Mother stops it'	어머니는 말리네요.
21	orencinɨn	mallineyo	'Orange is dried'	오렌지는 말리네요.
22	untoŋcaŋɨn	morɨneyo	'don't know playground'	운동장은 모르네요.
23	torɨrenɨn	millineyo	'Pulley is pushed'	도르래는 밀리네요.
24	cuməninɨn	marɨneyo	'Pocket is dried'	주머니는 마르네요.
25	tenamunɨn	marɨneyo	'Bamboo is dried'	대나무는 마르네요.

26	tallananɨn	morɨneyo	'don't know the moon world'	달나라는 모르네요.
27	cənmuntenɨn	millineyo	'Junior college is pushed'	전문대는 밀리네요.
28	kamnamunɨn	mallineyo	'Persimon trees is dried '	감나무는 말리네요.
29	jəŋminenɨn	millineyo	'Youngmi's family is pushed'	영미네는 밀리네요.
30	nallarinɨn	nollaneyo	'Punk is surprised'	날라리는 놀라네요.
31	min tɨllenɨn	mallineyo	'Dandelion is dried'	민들레는 말리네요.
32	uranjumman	morɨneyo	'don't know only uranium'	우라늄만 모르네요.
33	allatinman	miruneyo	'Only Aladin postpones'	알라딘만 미루네요.
34	tetoŋkaŋman	morɨneyo	'don't know only Taetong River'	대동강만 모르네요.
35	mel laninman	morɨneyo	'don't know only melanin'	멜라닌만 모르네요.
36	ciŋgɨlpelman	morɨneyo	'don't know only Jingle Bells'	징글벨만 모르네요.

A.2 Tokyo Japanese

There were 32 target phrases and no fillers.

		target word	carrier		
Accent(2)	1.	amado	ga arimasu	'There is a rain shutter'	あまどがあります。
	2.	imoya	ga arimasu	'There is a potato store'	芋屋があります。
	3.	muroran	ni arimasu	'is in Muroran'	室蘭にあります。
	4.	amamori	ga simasu	'There is water leak'	雨漏りがします。
	5.	namamono	ga arimasu	'There is a living thing'	生ものがあります。
	6.	oniyuri	ga arimasu	'There is a tiger lily'	鬼百合があります。
	7.	aniyome	ga imasu	'There is a sister-in-law'	兄嫁がいます。
	8.	yonemura	ga imasu	'There is Yonemura'	米村がいます。
	9.	ninomiya	ga imasu	'There is Ninomiya'	二ノ宮がいます。
	10.	erimaki	ga arimasu	'There is a scarf'	えりまきがあります。
	11.	iranzin	ga arimasu	'There is an Iranian'	イラン人がいます。
	12.	mariina	ga arimasu	'There is a marimba'	マリーナがあります。
	13.	yamamba	ga imasu	'There is a witch'	山姥がいます。
	14.	marimba	ga arimasu	'There is a marina'	マリンバがあります。
	15.	orenzi	ga arimasu	'There is an orange'	オレンジがあります。
2syl	16.	inu	ga arimasu	'There is a dog'	いぬがあります。
	17.	mimi	ga arimasu	'has ears'	みみがあります。
	18.	yama	ga arimasu	'There is a mountain'	やまがあります。
	19.	mura	ga arimasu	'There is a village'	むらがあります。
	20.	ami	ga arimasu	'There is a net'	あみがあります。
	21.	mune	ga arimasu	'has a chest'	胸があります。
	22.	iro	ga iidesu	'Color is good'	色がいいです。
Unaccent	23.	nemuri	ga asaidesu	'Sleep is light'	眠りがあさいです。
	24.	yamaimo	ga arimasu	'There is a yam'	山芋があります。
	25.	momoiro	ga iidesu	'Peach color is good'	桃色がいいです。
	26.	nininmae	ga arimasu	'There are two portions'	二人前があります。
	27.	udonya	ga arimasu	'There is a noodle shop'	うどんやがあります。
Accent(3)	28.	munemawari	ga arimasu	'has a bulky chest'	胸回りがあります。
	29.	monomorai	ga arimasu	'has a sty'	モノモライがあります。
	30.	reonarudo	ga imasu	'There is Leonardo'	レオナルドがいます。
	31.	iromoyoo	ga iidesu	'Color and shape are good'	いろもよおがいいです。
	32.	yamanobori	ga iidesu	'Hiking is good'	やまのぼりがいいです。

A.3 Mandarin Chinese

There were 20 target words and 9 fillers (three to four syllable words). Fillers are not presented here.

Carrier phrase: qǐngnínbǎ__ zàishuōyíbiàn.
"Please say __ again."
请您把 __ 再说一遍。

Tone 2 + Tone 2	1	míngnián	`next year'	明年
	2	míngrén	'celebrity'	名人
	3	línglíng	'cool'	泠泠
	4	míngmíng	'clearly'	明明
	5	niánlíng	'age'	年龄
	6	guómín	'people'	国民
	7	nóngmín	'farmer'	农民
	8	rénmín	'people'	人民
	9	ménlíng	'doorbell'	门铃
	10	léiléi	'hang in clusters'	累累
	11	nánnán	'murmuring'	喃喃
	12	lángláng	'clear and ringing'	琅琅
	13	mángrén	'blind person'	盲人
	14	láilín	'arrive'	来临
	15	láinián	'next year'	来年
Tone 2 + Tone 0	16	míngzi	'name'	名字
	17	piányi	'cheap'	便宜
	18	rénmen	'people'	人们
	19	péngyou	'friend'	朋友
	20	zánmen	'we'	咱们

A.4 English

There were 19 target phrases and 9 fillers of various lengths. Fillers are not presented here.

Carrier phrase: "No, I meant __."

1 Amelia Raymond
2 amenable meaning
3 eliminate Mariners
4 illuminated mirrors
5 analogous analysis
6 anonymous writings
7 anomalous meaning
8 immoral ambition
9 Norwegian marinas (for E1, E2)/ remaining minutes (for E3, E4)
10 Orwellian nightmare
11 immunity reaction
12 unmanageable employee
13 unruly employee (for E1, E2)/ Armani employees (for E3, E4)
14 linoleum knives
15 Millennium Resort
16 maligned by the media
17 remunerate a lawyer
18 malaria in Nigeria
19 aluminum mini blinds

Bibliography

Aiken, L. S. and S. G. West (1991). *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications.

Akanlig-Pare, G. and M. Kenstowicz (2003). Tone in Buli. *Studies in African Linguistics 31*, 55--96.

Arvaniti, A., D. R. Ladd, and I. Mennen (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics 26*, 3--25.

Arvaniti, A., D. R. Ladd, and I. Mennen (2006). Phonetic effects of focus and 'tonal crowding' in intonation: evidence from Greek polar questions. *Speech Communication 48*, 667--96.

Atterer, M. and R. Ladd (2004). On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. *Journal of Phonetics 32*, 177--197.

Baayen, R. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R., D. Davidson, and D. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language 59*, 390--412.

Barnes, J., N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel (2008). Alternatives to F0 turning points in American English intonation. *Journal of Acoustical Society of America 124*(4), 2497.

Barnes, J., N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel (2010). The effect of global F0 contour shape on the perception of tonal timing constrasts in American English intonation. *Speech Prosody 100445*.

Beckman, M. E. and P. Welby (2006). elbow-scripts-9feb2006. http://www.icp.inpg.fr/ welby/.

Boersma, P. and D. Weenink (1992-2009). Praat: doing phonetics by computer, version 5.1.12. www.praat.org.

Bruce, G. (1977). *Swedish Word Accents in Sentence perspective*. Gleerup.

Caspers, J. (1994). *Pitch Movements Under Time Pressure*. Dordrecht: ICG Printing.

Caspers, J. and V. van Heuven (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica 50*, 161--171.

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen, M. (1970). Vowel length variation as a function of the voicing of consonant environment. *Phonetica 22*, 129--159.

Cho, H. (2007). The effect of speech rate on the segmental anchoring hypothesis: A model of f0 as a function of alignment and slope constraints. Manuscript, MIT.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York: Harper and Row.

del Giudice, A., R. Shosted, K. Davidson, M. Salihie, A. Arvaniti, A. del Giudice, R. Shosted, K. Davidson, M. Salihie, and A. Arvaniti (2007). Comparing methods for locating pitch "elbows". Proceedings for the 16th International Congress of Phonetic Sciences.

Dieters, M. J., T. L. White, R. C. Littell, and G. R. Hodge (1995). Application of approximate variances of variance components and their ratios in genetic tests. *Theor Appl Genet 91*, 15--24.

Dilley, L. C., D. Ladd, and A. Schepman (2005). Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics 33*, 115--119.

Ewan, W. G. (1975). Explaining the intrinsic pitch of vowels. *Journal of Acoustical Society of America 58*(S1), S40.

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology 18*, 7--44.

Fougeron, C. and S.-A. Jun (1998). Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics 26*, 45--69.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P.F.MacNeilage (Ed.), *The Production of Speech*. Springer-Verlag.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gósy, M. and J. Terken (1994). Question marking in Hungarian: timing and height of pitch peaks. *Journal of Phonetics 22*, 269--81.

Grabe, E., B. Post, F. Nolan, and K. Farrar (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics 28*, 161--185.

Grice, M. (1995a). *The intonation of interrogation in Palermo Italian: implications for intonation theory*. Niemeyer.

Grice, M., D. R. Ladd, and A. Arvaniti (2000). On the place of "phrase accents" in intonational phonology. *Phonology 17*, 143--185.

Gussenhoven, C. (2000). The boundary tones are coming: on the nonperipheral realization of boundary tones. In M.Broe and J.Pierrehumbert (Eds.), *Papers in Laboratory Phonology*, Volume V, pp. 132--151. Cambridge: Cambridge University Press.

Igarashi, Y. (2004). "Segmental Anchoring" of F0 under changes in speech rate: evidence from Russian. *Speech Prosody*.

Ishihara, S. (2003). *Intonation and Interface Conditions*. Ph. D. thesis, Massachusetts Institute of Technology.

Ishihara, T. (2006). *Tonal Alignment in Tokyo Japanese*. Ph. D. thesis, University of Edinburgh.

Jun, S.-A. (1996). *The Phonetics and Phonology of Korean Prosody*. Garland.

Jun, S.-A. (2000, November). K-ToBI (Korean ToBI) Labelling Conventions. *UCLA Working Papers in Phonetics 99*.

Keating, P. A. (1979). *A phonetic study of a voicing contrast in Polish*. Ph. D. thesis, Brown University.

Keating, P. A. (1985). Universal phonetics and the organization of grammars. In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*. Adademic Press, Inc.

Kenstowicz, M. and C. Park (2006). Laryngeal features and tone in Kyungsang Korean: a phonetic study. *Studies in Phonetics, Phonlogy and Morphology*.

Kubozono, H. (1991). Modeling syntactic effects of downstep in Japanese. In G. J. Docherty and D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Segment, Gestures, Tone*, pp. 368--387. Cambridge University Press.

Kubozono, H. (1993). *The Organization of Japanese Prosody*. Kurosio Publishers.

Ladd, D. (2004). Segmental anchoring of pitch movements: autosegmental phonology or speech production? In H. Quené and V. van Heuven (Eds.), *On Speech and Language: Essays for Sieb B. Nooteboom*, pp. 123--131. LOT.

Ladd, D. (2008). *Intonational Phonology*. Cambridge University Press.

Ladd, D., D. Faulkner, H. Faulkner, and A. Schepman (1999). Constant "segmental anchoring" of F0 movement under changes in speech rate. *Journal of Acoustical Society of America 106*(3).

Ladd, D., I. Mennen, and A. Schepman (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of Acoustical Society of America 107*(5).

Ladd, D. and A. Schepman (2003). "Sagging transitions" between high pitch accents in English experimental evidence. *Journal of Phonetics 31*, 81--112.

Lee, H.-J. and H.-S. Kim (1997). Phonetic Realization of Seoul Korean Accentual Phrase. *Harvard studies in Korean Linguistics*, 153--166.

Lehiste, I. and G. E. Peterson (1961). Some basic considerations in the analysis of intonation. *Journal of Acoustical Society of America 33*, 419--25.

Lennes, M. (2003). Collect pitch data at point.praat. http://www.helsinki.fi/ lennes/praat-scripts/.

Li, Z. (2003). *The Phonetics and Phonology of Tone Mapping in a Constraint-Based Approach*. Ph. D. thesis, Massachusetts Institute of Technology.

Martin, S. (1992). *"Yale Romanization" A Referene Grammar of Korean*. Rutland and Tokyo: Charles E. Tuttle Publishing.

Myers, S. (2003). F0 timing in Kinyarwanda. *Phonetica 60*, 71--97.

Papoulis, A. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph. D. thesis, Massachusetts Institute of Technology.

Pierrehumbert, J. B. and M. E. Beckman (1988). *Japanese Tone Structure*, Volume Linguistic Inquiry Monography 15. MIT Press.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Poser, W. J. (1984). *The Phonetics and Phonology of Tone and Intonation in Japanese*. Ph. D. thesis, Massachusetts Institute of Technology.

Prieto, P. (1998). The scaling of the L values in Spanish downstepping contours. *Journal of Phonetics 26*, 261--82.

Prieto, P. (2005). Stability effects in tonal clash contexts in Catalan. *Journal of Phonetics 33*, 215--242.

Prieto, P. and F. Torreira (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics 35*, 473--500.

Prieto, P., J. van Santen, and J. Hirschberg (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics 23*, 429--451.

Ripley, B. (2009). Package 'pspline'. http://cran.r-project.org/web/packages/pspline.

Silverman, K. and J. B. Pierrehumbert (1990). The timing of prenuclear high accents in English. *Papers in Laboratory Phonology 1*.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge University Press.

Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics 7*, 71--79.

Suomi, K. (2009). Durational elasticity for accentual pruposes in Northern Finnish. *Journal of Phonetics 37*, 397--416.

't Hart, J. and A. Cohen (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics 1*, 309--327.

't Hart, J. and R. Collier (1975). Integrating different levels of intonation analysis. *Journal of Phonetics 3*, 235--255.

't Hart, J., R. Collier, and A. Cohen (1990). *A perceptual study of intonation: an experimental-phonetic approach*. Cambridge University Press.

Thorsen, N. (1984). F0 timing in Danish word perception. *Phonetica 41*, 17--30.

van Santen, J. and J. Hirschberg (1994). Segmental effects on timing and height of pitch contours. *Proceedings of 1994 International Conference on Spoken Language Processing. Yokohama, Japan*, 719--722.

Warner, N. (1997). Japanese final-accented and unaccented phrases. *Journal of Phonetics 25*, 43--60.

Welby, P. and H. Lœvenbruck (2005). Segmental "anchorage" and the French late rise. Interspeech 2005.

Welby, P. and H. Lœvenbruck (2006). Anchored down in Anchorage: Syllable structure, rate, and segmental anchoring in French. *Rivista di Linguistica 18*(1), 39--18.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica 55*, 179--203.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics 27*, 55--105.

Xu, Y. (2001). Fundamental Frequency Peak Delay in Mandarin. *Phonetica 58*, 26--52.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication 46*, 220--251.

Xu, Y. and X. Sun (2001). Maximum speed of pitch change and how it may relate to speech. *Journal of Acoustical Society of America 111*(3).