LEXICAL TONE GESTURES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Hao Yi

August 2017

LEXICAL TONE GESTURES

Hao Yi, Ph.D.

Cornell University 2017

This dissertation investigates the lexical f_0 control in Mandarin within the framework of Articulatory Phonology (AP) in two experiments: an imitation study (Experiment 1) and an Electromagnetic Articulography production study (Experiment 2). Empirical results are accounted for by making reference to a gestural model of f_0 control within the AP framework.

The main goal of Experiment 1 is to investigate the speaker control of relative timing between lexical tones and segments in Mandarin. Past research has shown that there exist consistent patterns of relative timing between f_0 turning points and acoustic landmarks. It is unknown how the relative timing patterns are controlled by native speakers, and to what extent the control over relative timing is influenced by perceptual or motoric categories. During the experiment, participants imitated synthesized stimuli with parametrically varied turning point timing and fundamental frequency. By probing into the imitation patterns, it is found that the tone-to-segment alignment is strongly influenced by Mandarin lexical tone categories. It is argued that the relative timing patterns are governed by categorical modes of gestural coordination between lexical tone gestures and oral articulatory gestures. The advantages of the gestural account over the traditional auto-segmental account are further discussed.

Experiment 2 investigates the interaction between lexical tones and sentence-level intonation. Previous research has treated this issue an acoustic one, but few studies have made articulatory arguments about the tone-intonation interaction.

Using both kinematic and acoustic data, it is found that the intra-syllabic relative timing patterns of gestures are altered by the presence of intonational events such as boundary tones and prosodic focus, which favors the idea that intonational tones interact with lexical tones locally, rather than imposing a global f_0 contour. It is proposed that boundary tone (BT) gestures and pitch accent (μ) gestures, respectively triggered by boundary tones and focus-introduced pitch accents, interfere with the intra-syllabic gestural coordination by way of coupling strength, thereby altering the relative timing patterns. The articulatory evidence thus provides an alternative perspective to the extensively debated issue.

BIOGRAPHICAL SKETCH

Hao Yi was born in Guangshun, Rongchang, Chongqing in 1990. He went to Rongchang Middle School and Chongqing Nankai Secondary School. He attended Renmin University of China where he majored in Statistics. After graduating from RUC in 2012, he travelled to United States of America to study Linguistics at Cornell University. His research mainly focuses on experimental phonetics, and his research interests include speech planning, articulation, prosody, tone languages, etc. Currently, he works as a Speech and Data Scientist at Nuance Communications, Inc.

This dissertation is dedicated to my parents, Yunhua and Baiqi, who have always stood by me and supported my dreams.

ACKNOWLEDGEMENTS

It is an understatement to say that I am thankful to my advisor, Sam Tilsen, an insightful mentor throughout my five years at Cornell. For every project we worked on, he always challenged me to think over the problem from alternative perspectives. He liked to emphasize on the importance of having scientific hypotheses and predictions before conducting any experiments. He also encouraged me to play with experiment data, explore different ways of analyzing data, and come up with various interpretations. Most importantly, Sam's rigorous work ethic and deep passion for linguistics research taught me invaluable lessons of working hard and having fun at the same time. I also feel immensely grateful to the other two professors on my special committee—Abby Cohn and Draga Zec, for their time, patience, and guidance. Abby always provided the most detailed comments to every draft I showed her, asking a wide range of thought-provoking questions. Thanks to her, this dissertation can have a theoretical backbone in addition to a large volume of empirical data. Draga, who served as the chair of my second qualifying paper, was also constructive to this dissertation and many other projects of mine. For the confidence she placed in me in our one-on-one meetings, the suggestions she offered to me at my various presentations, and the feedback she gave to me on my numerous drafts, I am deeply indebted to her.

I also want to extend my gratitude to my fellow classmates and other graduate students whom I have the chance to meet, talk to, and occasionally have intense discussions with. We went through ups and downs, shared laughs and tears, and most importantly, we grew together as linguists. Especially, I would like to acknowledge Robin Karlin for the countless scintillating discussions that helped me inch closer to the finish line. I owe thanks to Yanyu Long for participating in the pilot experiment and giving valuable feedback. My gratitude also goes to my friends

and colleagues in the department: Ferdinan Okki Kurniawan, Chris Sundita, Zac Smith, Naomi Enzinna, Amui Chong, and Jixing Li. The last five years would have been much less memorable without them. Even though our paths crossed only briefly, I will not forget the sage advice and kind help from Zhong Chen, Nan Li, Linda Heimisdóttir, Anca Chereches, and Becky Butler.

Thanks to our Research Systems Consultant Bruce McKee, who helped troubleshooting lab equipment. A heartfelt thank-you to Holly Boulia and Michael Duane Williamson, our Administrative Assistant and Administrative Manager, who were tremendously effective in providing administrative help. I also want to thank Xiaorong "Alicia" Zhou, who assisted in administering many experiment sessions. This work would not have been possible without their help.

I owe special thanks to my mentors Aijun Li and Jun Gao in Institute of Linguistics at Chinese Academy of Social Sciences. They opened a window on linguistics to me, for which I will be forever grateful. I am also indebted to my teachers in School of Statistics at Renmin University of China for the statistical way of solving problems they taught me.

I cannot accomplish what I have accomplished so far without the support from my friends. I want to specially acknowledge Jiahui Huang, Long Feng, Xuchen Wang, Mark McGuire, Zeng Zeng, Hanyi Wang, Qian Gong, Yuezhu He, Siyang Wang, and Zhe Wang. Each and every one of them is a beam of light in my life. I would not trade anything for their friendship.

Last but not least, I would like to express my deepest gratitude to my parents—Yunhua Yi and Baiqi Ji. They are my staunchest supporters, and have always stood by me whether in joys or trials. Every bit of my accomplishment is as much their doing as it is mine..

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

This dissertation investigates the lexical f_0 control in Mandarin within the framework of Articulatory Phonology (AP) (Browman and Goldstein, 1986, 1988, 1989, 1990a,b,c, 1992; Saltzman and Munhall, 1989). We ask two main questions: 1) How do speakers control the relative timing of lexical tones in relation to segments in Mandarin? 2) How do lexical tones interact with sentence-level intonation? To investigate these questions, an imitation study (Experiment 1) and a production study using Electromagnetic Articulography (Experiment 2) have been carried out. Empirical results are subsequently accounted for by making reference to a gestural model of f_0 control within the AP framework.

Experiment 1 was conducted using an imitation task. Participants imitated synthesized stimuli with parametrically varied turning point timing and fundamental frequency. By probing into the imitation patterns, it is found that the tone-to-segment alignment is strongly influenced by Mandarin lexical tone categories, although some degree of imitation is accomplished by some participants. It is argued that the relative timing patterns are governed by categorical modes of gestural coordination between lexical tone gestures and oral articulatory gestures.

Experiment 2 was a production study conducted using Electromagnetic Articulography (EMA). The target syllables, bearing lexical tones, occurred at different prosodic contexts, thus introducing interaction between lexical tones and intonation. By analyzing the production results, it is found that the intra-syllabic relative timing patterns of gestures are altered by the presence of intonational events such as boundary tones and prosodic focus, which favors the idea that intonational tones interact with lexical tones locally.

Segmental Anchoring The association of f_0-related events (such as lexical tones, pitch accents, and boundary tones) with segments has long been one of the most hotly researched area in the field of laboratory phonology. Recent studies (Arvaniti et al., 1998; Atterer and Ladd, 2004; Xu, 1997; D'Imperio et al., 2004; Mücke et al., 2012) have shown that f_0-related events do not occur randomly in relation to the segments. Instead, the relative timing pattern is governed via some mechanism controlled by speakers.

One account of the governing mechanism stems from the Autosegmental Metrical (AM) theory. The AM account argues that f_0-related events operate independently on a different tier from the segments ("autosegmental"), and are aligned in time to the segments by way of association line. These segments that f_0-related events are aligned to are termed "anchors" or "anchoring sites".

It can be argued that AM provides a phonetic account for the temporal relationship between f_0-related events and segments in the aforementioned studies. However, one of potential issues with this account is the proliferation of the seemingly arbitrary acoustic anchors. Moreover, the role of segments in production has been cast into doubt by various studies (Browman and Goldstein, 1990b; Pierrehumbert, 1990; Tilsen, 2016). Therefore, the notion that speakers associate f_0-related events with segments might be misguided.

Articulatory Phonology Articulatory Phonology (AP) offers an alternative approach to tone-to-segment alignment. Within this framework, the fundamental units of speech are articulatory gestures, which are associated with vocal tract articulators that contribute to the formation and release of constrictions. Two or more gestures can be coordinated to form a syllable, a word, etc. (Tilsen, 2016).

Relative timing patterns arise as a result of phasing control in the network of coupled gestures. The most successful example is the c-center effect: in a CCV syllable, the activations of the two consonant (C) gestures are displaced equally in opposite directions in time from the activation of the vowel (V) gesture. The c-center effect occurs because of the competition between the in-phase and anti-phase coupling relations: the onset C gesture is in-phase coupled to the nuclear V gesture, and that two C gestures in an onset cluster are anti-phase coupled to one another (Browman and Goldstein, 1989; Nam and Saltzman, 2003).

Recently, the control of f_0 has been modeled as tone gestures within the AP framework (D'Imperio et al., 2004; Gao, 2008; Mücke et al., 2012; Katsika et al., 2014). The relative timing patterns of f_0-related events and segments thus arise out of different modes of gestural coordination between f_0-related gestures and oral articulatory gestures. Gao (2008) proposed that Mandarin tones consist of two invariant lexical tone gestures— high (H) and low (L). The lexical tone gestures are coordinated with the oral articulatory gestures, i.e., consonant (C) and vowel (V) gestures, in various coupling modes. For instance, for Tone1-bearing syllables, the H gesture, like an onset C gesture, is anti-phase coupled to the C gesture and in-phase coupled to the V gesture. The competition of in-phase and anti-phase coupling relations leads to the sequential gestural activations of the C, V, and H gestures, which resembles the c-center effect.

Experiment 1 Against the backdrop of the gestural model of f_0 control, this dissertation consists of two experiments on Mandarin tones. Experiment 1 investigated the speaker control over the relative timing of lexical tones. The experiment used a series of synthesized bi-tonal stimuli that varied parametrically in the relative timing and fundamental frequency of f_0 turning points. Experiment 1 is further

broken down into Experiment 1A and Experiment 1B. Specifically, in Experiment 1A, the parametric variation occurred within the first tone-bearing syllable, while in Experiment 1B, it occurred across syllable boundaries.

In each of these two experiments, there were three parts: an AX discrimination task, an imitation task, and a second AX discrimination task. In both AX discrimination tasks, the participants were asked to judge whether two stimuli are the same. In the imitation task, they were instructed to imitate the bi-tonal stimuli as accurately as possible. Perceptual tone categories were identified through discrimination results. More importantly, the imitation results were analyzed to investigate how the variation in stimulus is reflected in imitation. The research question outlined above can be addressed by both perception and production results.

The relationship between the relative timing and fundamental frequency of f_0 turning points was also investigated in Experiment 1. Specifically, the gestural model of f_0 control argues that in a bi-tonal sequence, the f_0 turning point is indicative of the activation of the second tone gesture. Therefore, the relative timing and fundamental frequency of f_0 turning points vary with one another due to gestural overlap or underlap. However, evidence from lexical tone languages is still lacking possibly due to the difficulty of manipulating f_0 in lexical tones. The current experiment can provide more evidence to back up the notion that f_0 control can be modeled as gestures in a gestural framework.

Experiment 2 Experiment 2 examined the interaction between lexical tones and intonation within the AP framework. Two categories of models, i.e., the overlay model and the unification model, have been proposed to account for the tone-

intonation interaction. Under the overlay model, the overall f_0 is the result of local perturbations of lexical tones being imposed onto the global f_0 contour associated with intonational processes. In other words, lexical tones and intonation are implemented in parallel. Under the unification model, lexical tones and intonational tones are the same type of phonological entity, and they interact with each other locally. The two models were tested in Experiment 2 by investigating the temporal coordinative patterns of f_0 gestures and oral articulatory gestures in different prosodic constructions. The relative timing of gestural activations served as a diagnostic for the influence of intonation on the intra-syllabic alignment of lexical tones to segments.

An Electromagnetic Articulography (EMA) experiment was conducted. Both articulatory and acoustic data were collected for the target syllable [ma] bearing either Tone2 or Tone4, which is embedded in a sentence. The target syllable occurred either at the phrase-medial position or phrase-final position, and can be accented or un-accented. Each sentence ended with either a question mark or period, eliciting a question or a statement, respectively. The temporal lags between the onsets of the C and V gestures, and between the V and f_0 gestures, were measured. The onset lags and the derived metric CV%, which characterizing the relative phase of the V gesture in relation to the C and f_0 gestures, were subsequently compared among different prosodic constructions. The effects of boundary tones, prosodic focus, lexical tones, and their interaction on gestural coordination were investigated to assess the tone-intonation interaction.

Overview of Dissertation The rest of the dissertation is organized as follows: Chapter 2 outlines the general background for both experiments Experiment 1 and 2. It begins with a detailed review of the AM theory using Bruce's 1977 seminal

work on Swedish Accent I and II words as the example. Bruce (1977) has inspired research on the segmental anchoring of f_0 events in various languages, such as Greek (Arvaniti et al., 1998), German (Atterer and Ladd, 2004), and Mandarin (Xu, 1997). An alternative view of segmental anchoring is discussed within the framework of AP (Browman and Goldstein, 1989), and is further backed up by more studies on the f_0 alignment with articulatory data (D'Imperio et al., 2004; Mücke et al., 2012; Gao, 2008).

Experiment 1 and 2 are presented in Chapter 3 and Chapter 4, respectively. Each chapter begins with a brief introduction, which is followed by additional literature reviews that provide specific background for each experiment. Then the hypotheses and predictions are laid out. After that, the experiment methodology, including participants, stimuli construction, procedure, data processing, and data analysis, is elaborated. The results are presented in the second to last section. Finally, each of these two chapters concluds with a discussion section in which attempts to account for the results are made under the gestural model of f_0 control.

Chapter 5 concludes the dissertation by discussing possible directions of future research.

CHAPTER 2

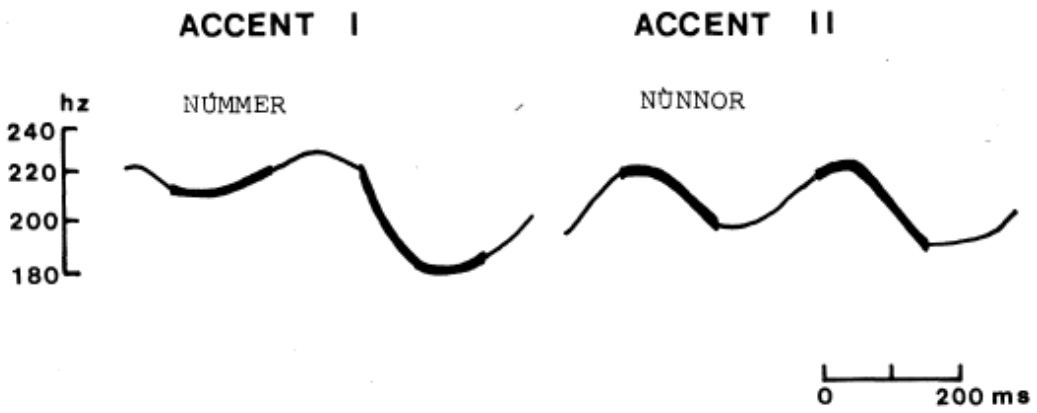
GENERAL BACKGROUND

2.1 Autosegmental Metrical Theory

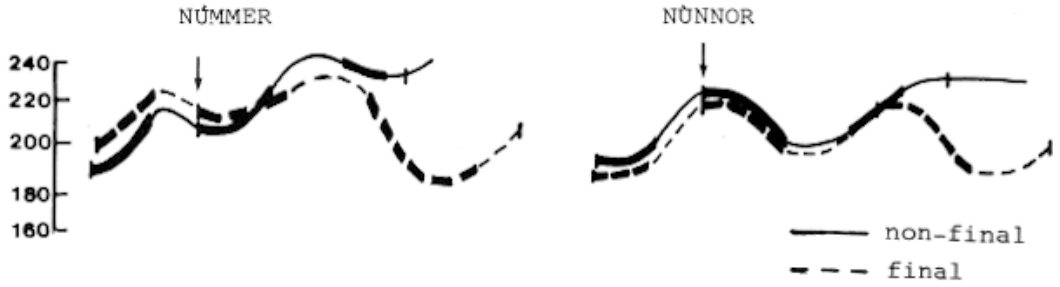
Native speakers of both intonation and tone languages have been reported to exert a high degree of f_0 control (Bruce, 1977; Prieto et al., 1995; Xu, 1997, 1998; Xu and Wang, 2001; Arvaniti et al., 1998; Ladd et al., 1999, 2000; Atterer and Ladd, 2004; Arvaniti et al., 2006; Prieto et al., 2005). The association between segments and tones, including pitch accents, phrase accents, boundary tones, and lexical tones, is relatively constant. In other words, tonal events, such as f_0 extrema, are consistently aligned in time with some segmental landmarks, such as syllable boundary. This phenomenon was termed as *segmental anchoring* by Ladd et al. (1999).

Segmental Anchoring has its roots in Autosegmental Metrical (AM) theory of intonational phonology, which in turn was greatly inspired by Bruce's (1977) seminar work on Swedish word accents. In Stockholm Swedish, the f_0 contours of Accent I and Accent II words can be superficially summarized as Accent I words have one single f_0 peak and Accent II words have two f_0 peaks. For instance, in Figure 2.1(a), *nú*mmer ("number", Accent I) has one f_0 peak to the right of the vocalic offset of the accented syllable, whereas *nú*nner ("nuns", Accent II) has two f_0 peaks: the first peak is aligned with the vocalic onset of the accented syllable and the second peak the following unaccented syllable.

While the phonetic difference between the two accents is striking in citation forms, it is eliminated in other contexts, as seen in Figure 2.1(b). When the two



(a) f_0 contours of Accent I and Accent II words in citation forms. Vowels are shown in thick lines.



(b) f_0 contours of Accent I and Accent II words following an unstressed syllable. Vertical arrows indicate the vocalic onsets of the accented syllables.

Figure 2.1: f_0 distinction between Swedish Accent I and Accent II words in citation forms (a) and following an following an unstressed syllable in final and non-final positions. From Bruce (1977)

aforementioned words are preceded by an unstressed syllable in a non-final position, the superficial f_0 distinction between the two words becomes absent in that both f_0 contours have one peak and the f_0 fall in Accent II disappears. However, there is still a difference regarding the alignment of the f_0 peak: it precedes the vocalic onset of *múmmer* (Accent I) but coincides with that of *núnnor* (Accent II). Moreover, the second f_0 peak occurs in both accents in final position whereas the difference in the alignment of the first peak still holds.

The absence of the second f_0 peak in Accent II words in a non-final position

suggests that this accent is not a perceptual representation at the word level. Instead, Bruce suggested the second peak in Accent II is the phonetic implementation of a sentence-level accent, present only phrase-finally. Note that citation forms are also short but complete sentences, therefore contain this sentence-level intonation feature. Hence, the sole f_0 peak in Accent I citation form words is also the manifestation of the sentence-level accent.

Regarding the difference between the two accents in the alignment of the f_0 peak following an unstressed syllable both phrase-medially and phrase-finally, Bruce also offered an elegant explanation: a word-level accent is present in both Accent I and II words, and the two word-level accents are in fact the same except for their temporal alignments. In Accent I words, because this accent precedes the accented syllable, it is only phonetically present when the accented syllable follows an unstressed syllable, but absent in citation forms. However, in Accent II words, the synchrony of the f_0 peak and the accented vowel renders this accent always present.

To sum up, the difference manifested in the two types of words is attributed to the difference in the temporal alignment of the word-level accent. The word-level accent is followed by a sentence-level accent phrase-finally (such as in citation forms). The word-level accent is aligned earlier in Accent I words than in Accent II words. When the word-level accent is associated with the word-initial syllable, it is phonetically present in Accent II words because of its late alignment. On the other hand, the sentence-level accent is always present in both types of words, giving rise to the second—sometimes the sole—peak in the f_0 contour.

Two Types of Tone Events Two types of tone events were discussed in Bruce (1977): word-level and sentence-level accents.

The notion of "word-level" does not imply that the accent itself is a lexical feature. Instead, the word-level accents are intonational features, and are associated with certain syllables to serve as perceptual cues to prominence or stress (Ladd, 2008). This type of "word-level" accent is usually referred to as pitch accent. The word-level accent in Swedish is one example. Another example is English pitch accents. According to Pierrehumbert (1980), there are six or seven possible configurations of pitch accents, which can be divided into monotonic and bi-tonal pitch accents. A monotonic pitch accent (e.g., H*, L*) is associated with a stressed syllable. In bi-tonal pitch accents (e.g., L*+H, L+H*), only the starred tone is associated with a stressed syllable. Note that the asterisk (*) indicates a tone is associated with the stressed syllable, according to the ToBI (Tone and Break Indices) convention (Beckman and Ayers Elam, 1997).

The other type of tone events, also described as the "sentence-level" accent by Bruce (1977), is also an intonational feature. This type of accent is usually referred to as edge accents or edge tones within AM framework. For instance, English has phrase accent such as L-, noted with a hyphen (-), and boundary tones such as L%, noted with a percent sign (%). Phrase accentd and boundary tones are associated with the syllables at the end of an intermediate phrase and an intonation phrase, respectively.

Localness of Tone Events Bruce's 1977 thesis established the fundamental idea underlying the framework of AM that the "tone structure" (Ladd, 2008) is a string of local events, such as High (H) and Low (L). These tone events can be

word-level accents or sentence-level accents.

The tone events are local, which means that the f_0 contour between two adjacent accents is mere a transition from one accent to another. The difference between the tone events and transitions highlights the important underlying assumption that these Highs and Lows are categorically distinct, and can adequately provide a phonological characterization of a continuous f_0 contour. Further phonetic interpolation provides a mapping from the discrete tone events to the continuous acoustics. This distances AM from theories such as Xu and Wang (2001) that argue f_0 rises and falls are part of the phonological inventory of tone events. Moreover, global trends in f_0 are therefore derived from one-by-one specifications of the local tone events (Ladd, 2008).

Segmental Anchoring of Tone Events Another important assumption of the AM approach is that the local tone events constitute a tone sequence, including both word-level and sentence-level accents. While operating on a different tier than the segments, i.e., "autosegmental", the sequence of tone events are aligned in time to the segments by way of association lines. That is, the tone events are anchored to certain segments. Note that AM does not distinguish between the two types of accents, word-level accents and sentence-level accents, in terms of their temporal alignment to the segmental string.

As shown in Bruce (1977), both the word-level and sentence-level accents are anchored to certain points in the segmental string, and the interplay between their temporal alignments results in the f_0 contours associated with Accent I and II words. In both types of words, the sentence-level accent can be preceded by the word-level accent. The sentence-level accent is always fixed at the end of the phrase,

therefore always present phrase-finally (including in citation forms). However, the alignment of the word-level accent is earlier in Accent I words than in Accent II words, which could induce the difference in the f_0 contours.

2.2 Segmental Anchoring in Various Languages

Following Bruce (1977), a large body of research has investigated segmental anchoring in various languages, showing that both intonational tones (pitch accents, phrase accents, and boundary tones) and lexical tones are aligned in time to the segmental string. Three studies on segmental anchoring in Greek, German, and Mandarin Chinese are summarized in Section 2.2.1, Section 2.2.2, and Section 2.2.3, respectively.

2.2.1 Greek Prenuclear Accents (Arvaniti et al., 1998)

A sequence of two pitch accents can be independently aligned in time to different segmental strings.

Arvaniti et al. (1998) investigated the alignment of the H tone in Greek prenuclear accents that were previously analyzed as bi-tonal accents L*+H. In this account, H was treated as the trailing tone, while L* was the starred tone. It was established that the L* tone coincided with the accented syllable, and the H tone usually occurrred within the post-accentual syllable. This study particularly looked for the factors affecting the alignment of the H tone.

It was found that the temporal lag between the L tone and the H tone was not a

fixed excursion, and was influenced by the post-accentual segmental composition: the nasals resulted in earlier alignment of the H tone, and fricatives resulted in later alignment. This ran against the hypothesis that the "trailing"" tone is aligned at a fixed distance from the starred tone (Grice, 1995).

The authors then showed that the H tone was aligned relative to the post-accentual syllable rather than the accented syllable. It was found that the f_0 peak occurred on average 10.6 ms after the onset of the post-accentual vowel, irrespective of the duration or the segmental composition of the accented and post-accentual syllable. This is shown in Figure 2.2, where the distance in time between the onset of the post-accentual vowel and the f_0 peak (V1toH) was in an orthogonal relationship with the duration of the post-accentual vowel (V1toC2).

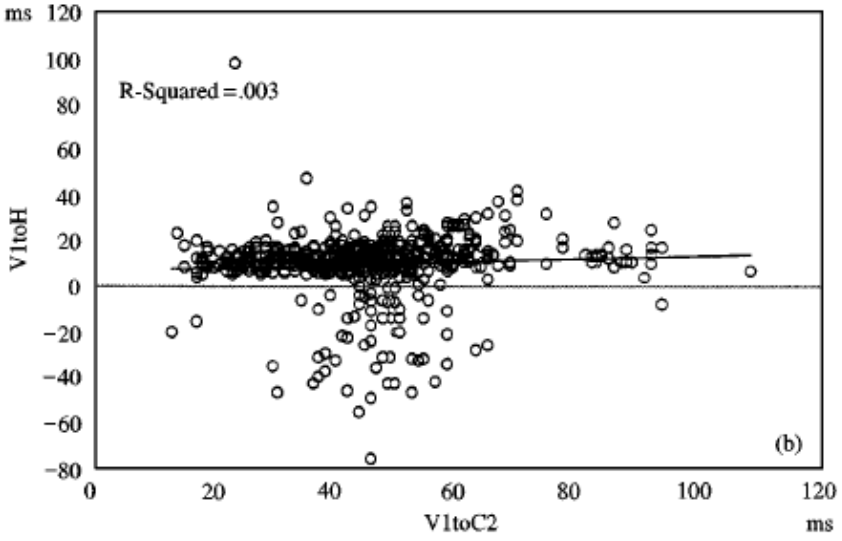


Figure 2.2: The distance between the onset of the post-accentual syllable and the f_0 peak as a function of the duration of the post-accentual vowel. From Arvaniti et al. (1998)

Moreover, as shown in Figure 2.3, the distance in time between the onset of the accented syllable and the f_0 peak (C0toH) was highly correlated with the distance

between the onset of the accented syllable and the post-accentual vowel (C0toV).[1] More than 80% of the variation in C0toH could be attributed to C0toV1.

Hence, the study showed that while the L tone coincided with the beginning of the accented syllable, the H tone was consistently aligned just after the onset of the first post-accentual vowel. The data did not support the bi-tonal accent being analyzed as L*+H, where the "trailing" tone (H) was aligned at a fixed distance from the starred tone (L*). Instead, it was argued that the bi-tonal accent consists of an L target and an H target, and the L and H tones were independently aligned to the segmental string.

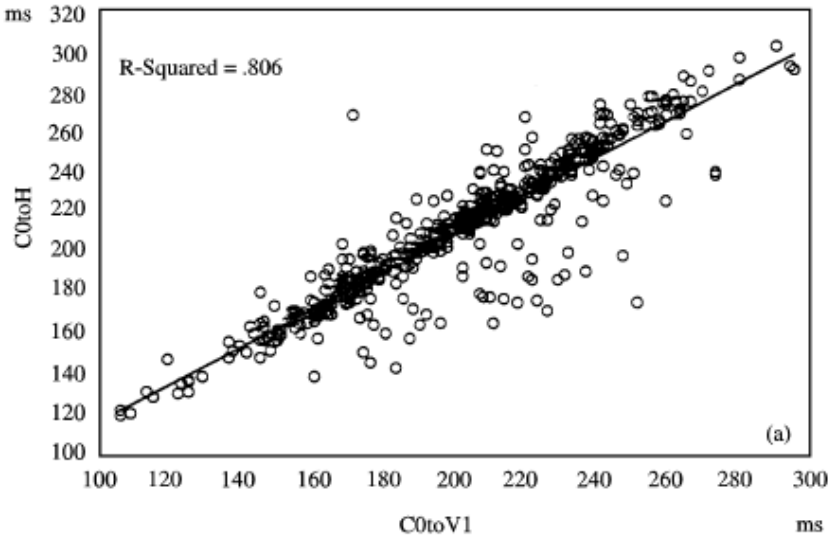


Figure 2.3: The distance between the onset of the accented syllable and the f_0 peak as a function of the distance between the onset of the accented syllable and the post-accentual vowel. From Arvaniti et al. (1998)

[1]The latter distance was slightly shorter than the former, as a result of the f_0 peak occurring just after the beginning of the post-accentual vowel.

2.2.2 German Prenuclear Rising Accents (Atterer and Ladd, 2004)

The same pitch accent may exhibit similar but different segmental anchoring patterns in different languages or dialects.

Atterer and Ladd (2004) investigated the segmental anchoring patterns of German rising pitch accents L+H in both Northern and Southern dialects. The authors found that the alignments of German prenuclear rising accents to segmental anchors were consistent, supporting previous research findings. They also found that the alignment of the L+H pitch accents was later in Southern German than in Northern German. On average, the distance in ms between the f_0 minimum and the onset of the stressed vowel (V0-L) was -40 ms in Northern German and -3 ms in Southern German. The distance in ms between the f_0 maximum and the onset of the post-accentual vowel (V1-H) was 21 ms in Northern German and 34 ms in Southern German. Moreover, the alignment of the rising pitch accents in both German dialects was later than in English (Ladd et al., 1999) and Greek (Arvaniti et al., 1998), as shown in Figure 2.4.

It was argued that the differences in alignment between the two German dialects should not be regarded as reflecting the differences in phonological associations but rather different realizations on one phonetic continuum of alignment. Because the phonological account would inevitably result in the proliferation of anchoring sites for tones such as the left periphery of the accented syllable (for Greek and English), the onset consonant of the accented syllable (Northern German), and the left periphery of the accented vowel (Southern German). Obviously, these anchoring sites were too arbitrary and theoretically hard to account for. Thus, despite

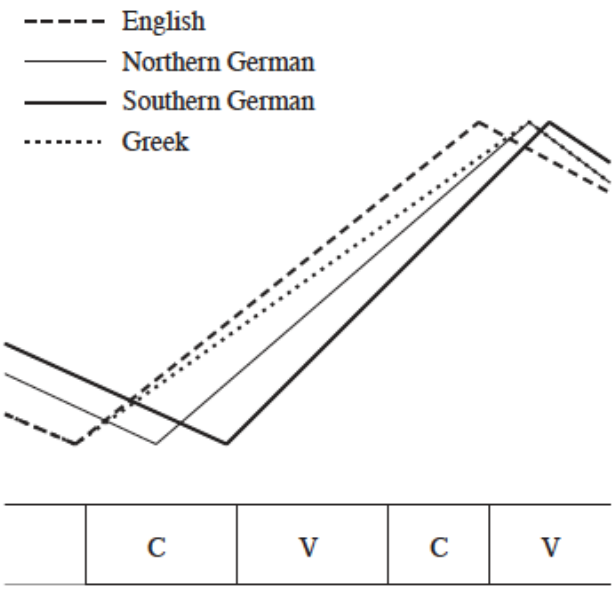


Figure 2.4: Comparisons in L+H alignment among Northern German, Southern German, English and Greek. From Atterer and Ladd (2004)

the consistent differences in alignment pattern, the prenuclear accents in Greek, English and both German dialects should fall into one category of rising accent.

Taking into consideration the alignment differences among German, English, and Greek, the authors went on to draw an analogy between tone alignment and voice onset time in stop consonants. Most languages have the voiced stop [p], the cross-linguistic phonetic realizations of [p] render differences in voice onset time from language to language. [p] in some languages are highly aspirated whereas [p] in other languages have little to no aspiration. Similarly, the rising pitch accents compared here should fall into a single category, and the cross-linguistic phonetic realizations of this rising accent can render differences in alignment from language to language.

2.2.3 Mandarin Tones (Xu, 1997)

The temporal alignment of lexical tones to segments in Mandarin, a lexical tone language, is also quite stable.

In Mandarin, each syllable bears one of the four lexical tones: Tone1 (high), Tone2 (rising), Tone3 (low), and Tone4 (falling). Xu (1997) found that Tone2 f_0 contour started low, and fell slightly before rising (Figure 2.5). The rising did not start until approximately 20% into the vowel. Similarly, Tone4 f_0 contour started high, and rose to an even higher position before falling sharply. The initial rise also lasted until about 20% into the vowel.

With respect to tonal coarticulation, it was found that during the production of a Tone2+Tone2 (rising-rising) sequence, the first Tone2 reached f_0 maximum after the acoustic offset of the first tone-bearing syllable, as illustrated in Figure 2.5. This was referred to as the carryover effect in the sense that the previous tone progressively influenced the following tone. As a result, the second Tone2 did not start to rise until well into the latter half of its tone-bearing syllable. A similar carryover effect, albeit of smaller magnitude, was found in falling-falling sequences.

It was noted later in Xu and Wang (2001) that for bi-syllabic (bi-tonal) sequences, the boundary between two syllables "serves as an anchor for both the offset of the preceding tone and the onset of the following tone" in the sense that "the preceding tone continuously approaches its target value up to the boundary, while the following tone starts to depart from that value right at the syllable boundary." Xu and Wang argued that phonetic implementation of lexical tones in Mandarin was closely associated with the tone-bearing unit, i.e., the syllable, which resulted in syllable boundaries serving as anchoring sites for lexical tones.

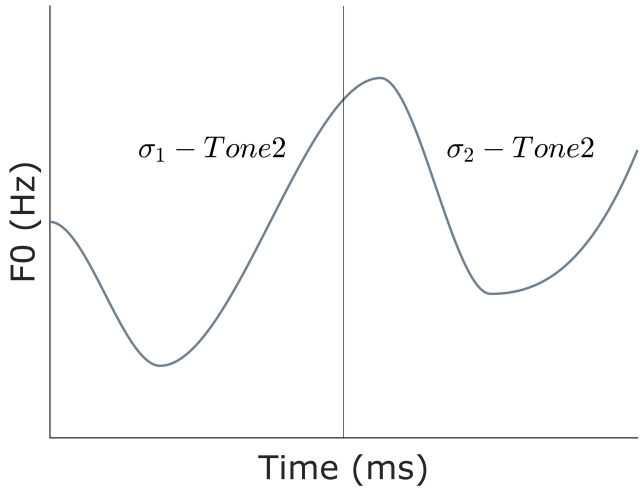


Figure 2.5: f_0 contour of a Tone2+Tone2 (rising-rising) sequence. The vertical line indicates the syllable boundary.

2.3 Inadequacy of Segmental Anchoring

There are several issues that underlie the segmental anchoring hypothesis. Firstly, as noted in Atterer and Ladd (2004) and some other studies such as Mücke et al. (2012), the differences in the acoustic alignment of similar tones (L+H) between different languages or dialects could result in the proliferation of the seemingly arbitrary acoustic sites, such as 20 ms into the accented syllable or the right periphery of the post-accented syllable. It is unlikely that speakers can exert such timing control that f_0 events are precisely aligned in time to segments. In fact, Atterer and Ladd (2004) argued the differences in acoustic alignment between different languages and dialects were acoustic in nature, and were indicative of different phonetic implementations of the same tone-to-segment alignment in different languages and dialects. However, the segmental anchoring hypothesis itself failed to offer explanations for these differences in tone-to-segment alignment.

Secondly, the notion of segment in production has come under challenge. Pier-

rehumbert (1990) argued that there was no evidence that segments can be viewed as a discrete representation in mind. This view was echoed by Browman and Goldstein (1990b) by suggesting that segments do not correspond to important informational units of the (cognitive/physical) phonological system. It was reasoned that the widespread use of the segment as "a practical tool" by linguists has contributed to the confusion that the segment serves as some kind of phonological representation in production. From this standpoint, the notion that f_0-related events are aligned in time to "segments" is in itself misguided.

Thirdly, acoustic signals are the result of simultaneous articulatory movements, therefore the acoustic alignment patterns can vary with different types of segments. For instance, the acoustic onsets of stops are often identified at the start of closure of the lips, which in itself is more difficult to measure for voiceless stops than for voiced stops, while the acoustic onsets of nasals are identified at the point in time when the intensity drops sharply in the spectrogram. The drop in intensity is achieved by lowering the velum while the oral cavity is blocked by closing the lips. However, the articulatory-to-acoustic mapping is yet unclear in the time domain, therefore posing difficulties in identifying comparable acoustic landmarks for different types of segments. This problem can be easily overcome by using kinematic data, which will be detailed in Section 2.4.

2.4 Articulatory Phonology

From a gestural perspective, segmental anchoring can be construed as an indirection reflection of the timing control accomplished via coordination of articulatory gestures based on kinematic movement of articulators. This alternative view is

developed from the Articulatory Phonology (AP) framework (Browman and Goldstein, 1986, 1988, 1989, 1990a,b,c, 1992; Saltzman and Munhall, 1989).

Under the AP framework, speech is conceptualized as constellations of overlapping articulatory gestures, the basic speech units. Gestures are associated with target values of vocal tract variables in the task dynamic model. A total of eight vocal tract variables were identified in Browman and Goldstein (1989): six for the oral gestures, one for the velic gesture, and one for the glottal gesture. The tract variables corresponding to the oral gestures are arranged in two dimensions: the lip aperture (LA) and lip protrusion (LP) for the lip (LIP) tract variable, the constriction location (CL) and constriction degree (CD) for both the tongue tip (TT) and tongue body (TB) tract variable. The tract variables velic aperture (VEL) and glottal aperture (GLO) correspond to the velum and glottis gestures. Vocal tract variables are in turn associated with the articulators (such as lips, tongue, and velum) that contribute to the formation and release of oral constrictions. For example, a bilabial closure gesture is associated with the task of achieving a negative target value of LA by coordinating the movements of the articulators—the upper and lower lips, and the jaw.

Each gesture is modeled as a one-dimensional oscillating system with the point-attractor being the target value of the vocal tract variable, and the system is critically damped. Under critical damping assumption, the trajectory generated by the oscillating system approaches the equilibrium target position increasingly slowly, rather than oscillating around it, as under zero damping assumption. Figure 2.6 shows trajectories generated by undamped (dashed) and critically damped (solid) oscillating systems. The solid trajectory demonstrates that it takes an infinite amount of time for the oscillating system to achieve the "real" equilibrium

target.

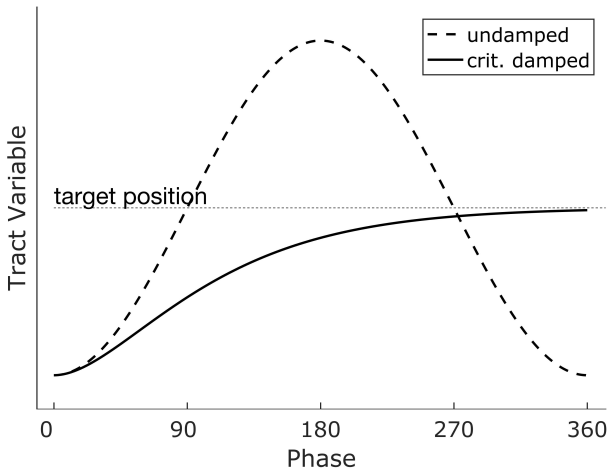


Figure 2.6: Comparison of undamped (dashed) and critically damped (solid) oscillating systems. Adapted from Browman and Goldstein (1990b)

For empirical trajectories generated by critically damped systems, the effective achievement of the target (Browman and Goldstein, 1990b) is defined by a velocity-related criterion (Figure 2.7) as the point in time that is 70% into the velocity range (between the minimum and maximum velocity), starting from the maximum velocity. Similarly, the onset is defined as the point in time that is 30% into the velocity range, starting from the minimum velocity. Note that the 30% threshold is arbitrarily chosen, yet is consistently used throughout the experiment.

The duration of a gesture is determined by the stiffness parameter (Browman and Goldstein, 1990a), which characterizes the responsiveness of the dynamical system in response to displacement. The more responsive the dynamical system is, the faster it drives the articulators to achieve the target. In other words, for a given target position, the stiffer the tract variable, the less time it takes to reach the effective target. As illustrated in Figure 2.8, the amount of time (activation time) taken to reach the same target for a gesture with high stiffness (dashed line) is shorter than for a gesture with low stiffness (solid line).

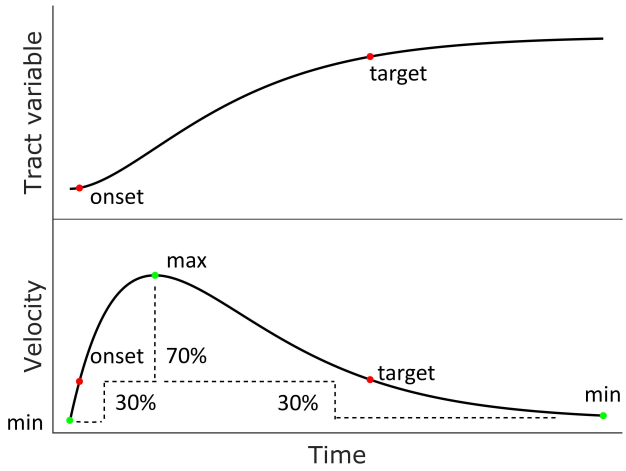


Figure 2.7: Example of a tract variable trajectory (top) and its corresponding velocity profile (bottom). Onset and target of a gesture are determined by the 30% velocity-related criterion. The onset and target are marked by red dots, and the velocity minimum and maximum green dots.

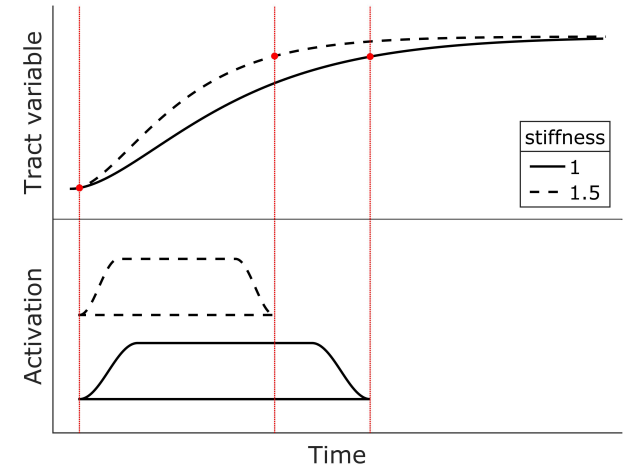


Figure 2.8: Comparison of two tract variables (top) and their corresponding activation times (bottom). Higher stiffness (dashed line) gestures take shorter time to reach the same target than lower stiffness (solid line) systems. Red dotted lines indicate the onsets and targets.

Two Coupling Modes A typical utterance consists of more than one gesture. Two or more gestures can be coordinated in specific ways to form units that may correspond to segments moras, syllables, and words, etc. (Tilsen, 2016). Recall

that a gesture can be modeled as oscillating systems. Moreover, a virtual cycle can be defined for each gesture based on the stiffness parameter (c.f. Tilsen, 2016 and references therein), and each gesture is activated at a defined phase (such as 0°) of the virtual cycle.

Gestural coordination could be characterized in terms of the relative phasing control in the network of the coupled oscillators. Specifically, the speech planning oscillators associated with articulatory gestures collectively derive a stabilized relative phase in the transient stabilization process prior to the initiation of the articulatory movements (Tilsen, 2017).

There are two preferred modes in which two oscillations can be coordinated: in-phase coupling and anti-phase coupling, with the former being the more stable coordination. In the top panel of Figure 2.9, the two planning oscillators (C and V) are in the in-phase coupling mode—they are coupled with a relative phase of 0°. In the bottom panel of Figure 2.9, the two planning oscillators (C1 and C2) are in the anti-phase coupling mode—they are coupled with a relative phase of 180°. Note that besides these two coupling modes, a pair of planning oscillators can also be coordinated with other relative phases.

The relative phase further induces the relative timing pattern of articulatory gestures, as will be seen next.

C-V and C-C Coupling Figure 2.9 illustrates two important relative phase coupling relations in the coupled oscillators model: the C-V in-phase coupling and the C-C anti-phase coupling. An onset consonant (henceforth C) gesture is in-phase coupled to the vowel (henceforth V) gesture. That is, the C and V gestures are activated with a relative phase of 0°, which is the most appealing mode of

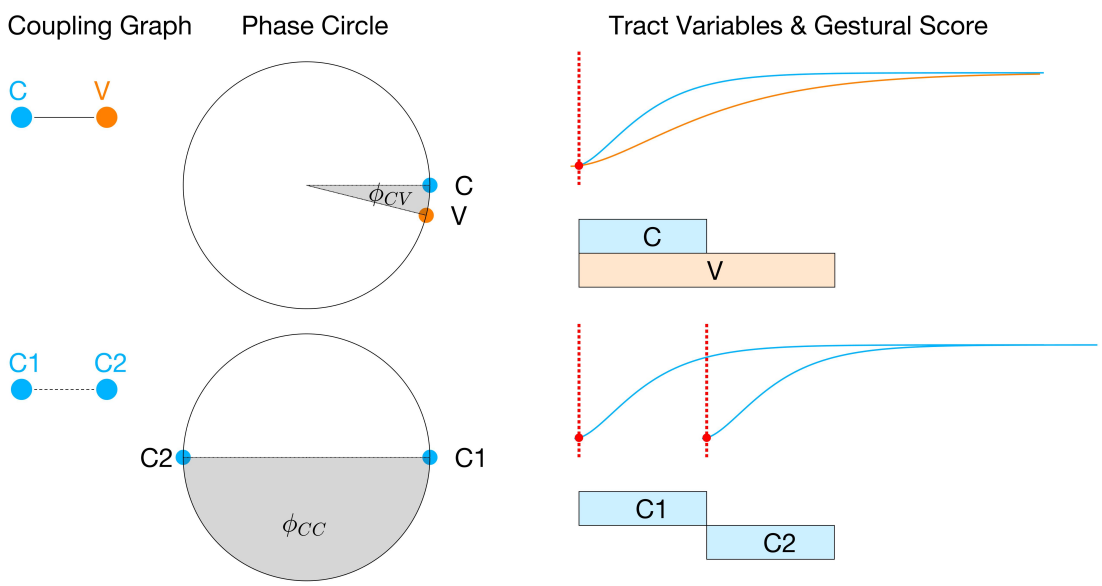


Figure 2.9: Illustrations of in-phase coupling (top) and anti-phase coupling (bottom). The first column show the coupling graph: solid line indicates in-phase coupling, and dashed line indicates anti-phase coupling. The second column shows the relative phasing of planned oscillators on a phase circle. The last column shows hypothesized tract variables and their corresponding gestural scores. Adapted from Tilsen (2016).

coordination due to its stability. Two C gestures in an onset consonant cluster are anti-phase coupled to each other. That is, the C1 and C2 gestures are activated with a relative phase of 180°.

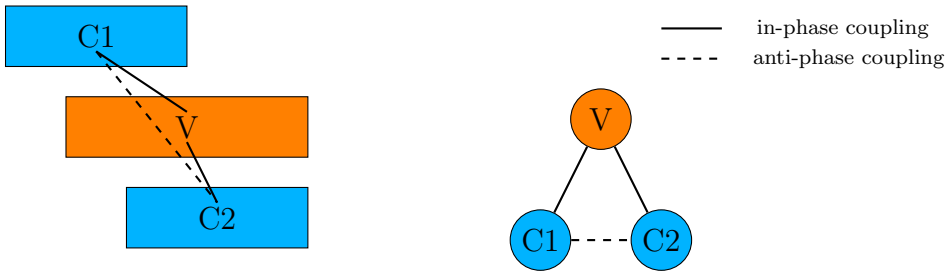
In CV syllables like *me*, the bilabial gesture (C) associated with the onset consonant [m] is in-phase coupled to the tongue body gesture (V) associated with the nuclear vowel [i]. The articulatory synchronization of the C and V gestures is in stark contrast with the apparent sequentiality observed in the acoustic signals.

In CCV syllables like *plea*, the bilabial gesture (C1) associated with the first consonant [p^h] and the tongue tip gesture (C2) associated with the second consonant [l] are anti-phase coupled to each other. Meanwhile, the C1 and C2 gestures are both in-phase coupled to the V gesture. The collective force of these coupling

relations results in the c-center effect (Browman and Goldstein, 1988; Nam and Saltzman, 2003): the onset of the C1 gesture is shifted leftward, and the onset of C2 gesture is shifted rightward, compared to their timing in CV syllables.

The shifts in the onset timing were confirmed in words with complex onsets in English and Georgian (Goldstein et al., 2009). For English nonce words *speets*, the onset of the C1 gesture (for /s/) was shifted earlier by 47 ms than in *seats*, while the onset of the C2 gesture (for /p/) was shifted later by 25 ms than in *peets*. The left-right asymmetry could be attributed to differences in C-V coupling strength. For Georgian, there was a rightward shift of 19 ms in the front-to-back consonant clusters: the latency between /k'/ and vowel was 133 ms in words with /k'/ onset and 114 ms in words with /t'k'/ onset. The direction of shift was consistent across speakers.

When the C1-V coupling strength is equal to the C2-V coupling strength, the activations of the C1 and C2 gestures are displaced equally in opposite directions in time from the activation of the V gesture. The midpoint between the activations of the bilabial gesture (C1) and the tongue tip gesture (C2)—the c-center, corresponds to the activation of the tongue body gesture (V), as illustrated in Figure 2.10.



(a) Gestural scores

(b) Coupling graph

Figure 2.10: Illustrations of the c-center effect in a CCV syllable in English: Gesture scores (left) and coupling graph (right). The onsets of the C1 and C2 gestures are displaced equally in opposite directions in time from the onset of the V gesture. Solid lines indicate in-phase coupling and dashed lines indicate anti-phase coupling.

Tones as Gestures The control of f_0 can also be conceptualized as a gesture, an oscillating system driving its articulators to achieve certain f_0 target. For example, a low tone (L) gesture involves an f_0 movement to achieve a low target (Figure 2.11). Note that the onset of the L gesture is the point in time from which f_0 starts to move downwards to approach the target—the f_0 minimum. Therefore, the L onset is located near the f_0 maximum that immediately precedes the f_0 target. Both lexical tones and intonational tones (pitch accents, phrase accents, and boundary tones) can be considered as tone gestures (Gao, 2008; Mücke et al., 2012).

The idea of treating f_0 as gestures was first brought up by McGowan and Saltzman (1995). They argued that in order to make the task dynamic model a more complete model of speech production, aerodynamic and laryngeal components had to be incorporated in addition to supralaryngeal activities. The notion of tract variable was extended to aerodynamic and laryngeal quantities like subglotal pressure, transglottal pressure, and f_0. For f_0, the proposed additional articulators to

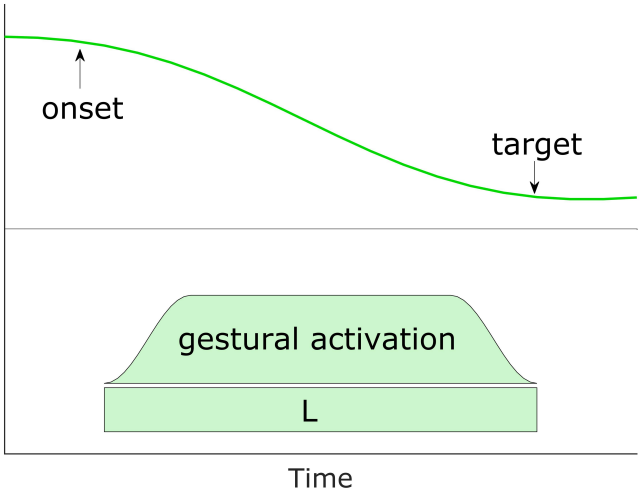


Figure 2.11: Illustration of a low tone (L) gesture within the AP framework: tract variable f_0 (top) and gestural score (bottom).

control these tract variables included vocal fold tension, force on the lungs, etc. The f_0 dynamics was subsequently modeled by a set of linear second-order differential equation.

Note that the "articulators" controlling f_0 proposed by McGowan and Saltzman (1995) are not the real articulators in the strictest sense. If the notion of articulatory gestures should be extended faithfully from segments to suprasegments, the f_0-related articulators such as the cricothyroid and thyroarytenoid muscles and the thyroid and cricoid cartilages should be tracked, because muscle contractions cause the cartilages to rotate, altering the tension of the vocal folds (Honda, 2004). Moreover, the vertical movement of larynx also plays a supportive role in facilitating f_0 change (Moisik et al., 2014). Consequently, instead of a tract variable that describes vocal tract geometry, f_0 —the acoustic measurement—is chosen as the tract variable.

This is because the physiology involved in f_0 production is so complex that it cannot be reduced to a simple movement as can segment production, which in-

volves an oral constriction action. Currently, there is not enough knowledge both theoretically and methodologically to model f_0 as oral articulatory gestures. However, it does not mean that the extension of gestures and tract variables to f_0 lacks theoretical consideration.

As McGowan and Saltzman (1995) argued, in the task dynamic model, the choice of tract variables is dependent on what the modeler considers to the linguistic goal of the speaker, and articulators are the effector system variables functionally related to the chosen tract variables. The linguistically relevant goal is to produce certain f_0 contour that evolves in time. Moreover, it has been shown that f_0-related "articulators' such as vocal fold tension and force on the lungs can be coordinated to model f_0 dynamics. Therefore, f_0 is treated as an abstract tract variable in AP, an idea that has been entertained in previous studies (Gao, 2008; Mücke et al., 2012; Katsika et al., 2014). Naturally, both lexical tones and intonational tones can be modeled as tone gestures (Gao, 2008; Mücke et al., 2012).

Conceptualizing tones as gestures allows for segmental anchoring to be analyzed with an intergestural timing model. In such a model, tone gestures are hypothesized to be coordinated with constriction gestures, i.e., consonant and vowel gestures, offering an alternative explanation of segmental anchoring patterns.

2.4.1 Articulatory Anchoring in Various Languages

In this section, we turn to the research that investigates the articulatory alignment of intonational and lexical tone gestures to oral articulatory gestures in Italian and French (Section 2.4.2), Catalan and Viennese German (Section 2.4.3), and Mandarin Chinese (Section 2.4.4).

2.4.2 Italian and French Pitch Accents (D'Imperio et al., 2004)

Articulatory anchoring tends to be more synchronous than acoustic anchoring. That is, the latency between f_0 events and articulatory landmarks tends to be smaller.

D'Imperio et al. (2004) investigated the anchoring of the H targets of nuclear rises in Neapolitan Italian, comparing the timing between the f_0 events and acoustic landmarks and the timing between the f_0 events and articulatory landmarks. The target words with stress on the first syllable, were embedded in either a statement or a question. The stimuli were read by two speakers at two self-paced rates, normal and fast. Both acoustic and articulatory measures were recorded.

It was found that the acoustic onsets of the accented vowel and the post-accentual vowel could contend for the anchor site of the H target. The latency between the H peak and the accented vowel onset was approximately 80 ms in statements and 150 ms in questions. As for the onset of the post-accentual vowel, the latency was shorter in both statements and questions. In statements, however, rate of speech was significant, suggesting the post-accentual vowel is not a stable anchor. In questions, despite that the effect of rate was not significant, the latency averaged around 50 ms. To sum up, the anchoring between the acoustic landmarks and the H peak is, in general, quite stable. However, stability with respect to tonal anchoring to these acoustic events does not equal synchronicity due to the rather large latencies.

Turning to articulatory anchoring, the data showed a more synchronous, though

not more stable, alignment between the f_0 events and the articulatory landmarks. It was found that in Neapolitan Italian statements, the latency between the H target and the peak velocity of the consonant trajectories—the vertical movement of the lower lip for the labial onsets [m], and the tongue apex for the apical onsets [n]— converged around 0 ms, while in questions the H target is timed to occur at the zero velocity of the consonant trajectories. Even though the effect of rate of speech was still significant in both statements and questions, the differences between fast and slow speech were quite small. Subsequent ANOVAs also confirmed that the articulatory alignment is more synchronous than the acoustic alignments in Neapolitan Italian.

This trend was less clear for French: the L tone of the early rise was closely aligned to the peak velocity of the consonant trajectories, although the acoustic alignment to the word onset also exhibited small latencies.

2.4.3 Catalan and Vienna German Pitch Accents (Mücke et al., 2012)

Cross-linguistics differences in alignment patterns can be explained by differences in gestural coordination.

Mücke et al. (2012) investigated the temporal coordination of intonational tone gestures—LH pitch accents—with oral constriction gestures in Catalan and Viennese German. In Catalan, the accented syllable was placed in the word-medial position in the target word, followed by a low boundary tone. In Viennese, the accented syllable was flanked by a minimum of one unstressed syllable. The definite article preceding the target was assumed to form a single prosodic word with the

target word. The target words were read in answer to questions that were designed to render focus. Both acoustic and articulatory measurements were recorded.

Acoustic measurements showed that the f_0 minimum occurred around 30 ms before the onset of the consonant in Catalan, whereas it was much delayed in Viennese German, occurring on average 70 ms after the onset of the consonant. On one hand, it was argued that the differences in f_0 alignment between Catalan and Viennese German should not be regarded as reflecting differences in phonological associations. Because to attribute the f_0 alignment differences to some phonological distinction, one would have to entertain somewhat arbitrary assumptions that "the left edge of the onset of the consonant" and "the middle of the onset of the consonant" are two different acoustic anchors. On the other hand, the authors also argued against treating the differences in acoustic alignment as being gradient in nature. Because that would beg the question of how speakers directly control the number of milliseconds between the acoustic onset of the consonant and the f_0 minimum.

Turning to the articulatory alignment between the H gesture and the oral articulatory gestures, there were also differences in tone alignment between Catalan and Viennese German. In Catalan, the C, V and H gestures were close to being initiated synchronously, occurring in the order of C-V-H, whereas in Viennese German, the H gesture was initiated considerably later than the gestural onsets of the C and V gestures, which were still close to being synchronous.

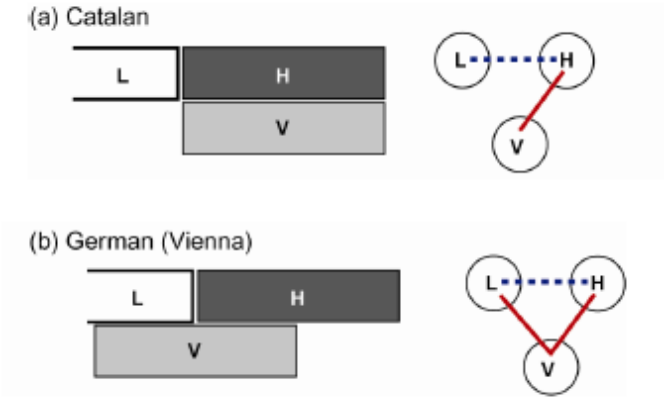


Figure 2.12: Two different coupling modes of LH pitch accents in Catalan (a) and Viennese German (b). From Mücke et al. (2012)

It was argued that native speakers controlled the alignment through different coupling modes. Therefore, the differences in tone alignment can be regarded as an reflection of the differences in coordinative patterns between the two languages, which is illustrated in Figure 2.12. In Catalan, the H gesture was in-phase coupled to the V gesture, and the L gesture was not directly coupled to the V gesture and starts within in the pre-accentual syllable. Therefore, the V and H gestures were synchronously initiated. In Viennese German, both L and H gestures were in-phase coupled to the V gesture, and anti-phase coupled to one another. The competitive coupling resulted in the L-V-H order, with the H gestural onset occurring on average 100 ms into the V gesture. Note that in both non-tonal languages, the C and V gestures were always initiated in synchrony, regardless of the alignment of the H gesture. It was argued that the presence of intonational tone gestures did not alter the intra-syllabic coordinative patterns of the C and V gestures, which differed from lexical tone gestures.

2.4.4 A Gestural Account of Mandarin Tones (Gao, 2008)

The relative timing of lexical tones is also controlled by gestural coordination within the framework of AP.

Gao (2008) investigated the articulatory alignment of Mandarin tones for the first time under the AP framework by examining the temporal alignment of Mandarin lexical tones to oral articulatory gestures such as bilabial closure, tongue tip raising, and tongue body lowering.

All four Mandarin tones were modeled as consisting of one or two invariant lexical tone gestures—H and L. Specifically, Tone1, a high tone, consists of a H gesture; Tone2, a rising tone, consists of a L gesture followed by a H gesture; Tone3, a low tone, consists of a L gesture; Tone4, a falling tone, consists of a H gesture followed by a L gesture.

Based on empirical data, Gao (2008) argued that Mandarin lexical tone gestures behaved like additional onset C gestures in that the lexical tone gestures were in-phase coupled to the V gesture, and anti-phase coupled to the C gesture. Patterns similar to the c-center effect in CCV syllables with complex onsets were observed in Mandarin CV syllables, where the lexical tone gesture functioned essentially as a second onset C gesture.

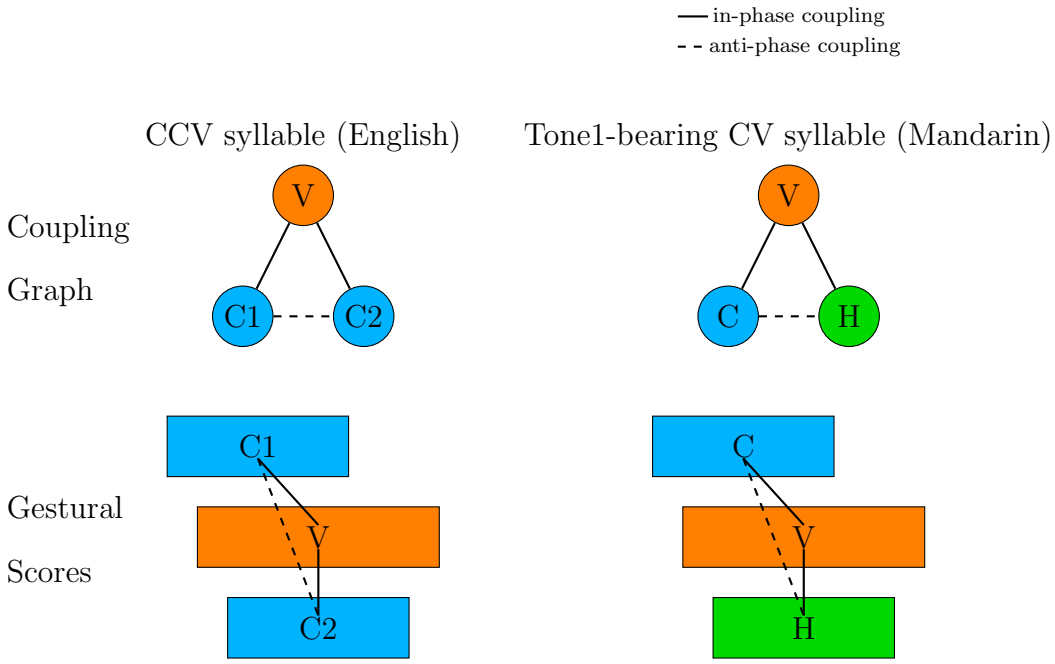


Table 2.1: Comparison in coupling graphs and gestural scores between Mandarin Tone1-bearing CV syllable and English CCV syllable.

Table 2.1 compares English CCV syllables (left column) to Mandarin Tone1-bearing CV syllables (right column) in coupling relations and gestural scores. In English CCV syllables, both C1 and C2 gestures are in-phase coupled to the V gesture, and the two C gestures are anti-phase coupled to each other. The collective force of the in-phase and anti-phase couplings results in the c-center effect, i.e., the gestural onsets of the C1 and C2 gestures are displaced equally in opposite directions in time from the gestural onset of the V gesture. Replacing the C2 gesture in English CCV syllables with the H gesture in Mandarin Tone1-bearing CV syllable, a similar c-center effect emerges in Mandarin. The C and H gesture is in-phase coupled to the V gesture, and the H gesture is anti-phase coupled to the C gesture. As a result of the balance between in-phase and anti-phase coupling forces, the gestural onsets of the C and H are displaced equally in opposite directions in

time from the gestural onset of the V gesture.

Figure 2.13 further details the coupling graphs and gestural scores of the four Mandarin tones. According to Gao (2008), Tone1-, Tone2-, and Tone3-bearing CV syllables all displayed the aforementioned c-center effect. Tone3, in this context, preceded a non-Tone3-bearing syllable, thus consisting of one L gesture. Despite the fact that Tone2 consisted of two lexical tone gestures (L and H), the two gestures functioned as one lexical tone gesture, which behaved like a second C gesture. Specifically, instead of being initiated in a sequential fashion, the L gesture was initiated in synchrony with the H gesture. Therefore, Tone2-bearing CV syllables were similar to both Tone1- and Tone3-bearing CV syllables in terms of gestural coordination. A comparison can be drawn from CCV syllables in English like *smee*, where the second consonant [m] consists of both the bilabial closure and the velum raising gestures, corresponding to C2a and C2b in Figure 2.13(b2, B2). The C2a and C2b gestures are in-phase coupled to each other, and function as one C2 gesture. As a result, the onset of the V gesture occurs halfway between the onsets of the C1 and C2 gestures.

Gao (2008) argued that Tone4-bearing CV syllables were slightly different from the other three tone-bearing syllables in gestural coordination. Like Tone2, Tone4 consisted of two lexical tone gestures, H and L. But unlike Tone2, both lexical tone gestures of Tone4 were coupled to the articulatory oral gestures. The onset of the V gesture occurred after the midpoint between the onsets of the C gesture and the H gesture, because the L gesture functioned as an additional onset C gesture, in that it was in-phase coupled to the V gesture and anti-phase coupled to the H gesture. A comparison can be drawn from CCCV syllables in English, where the three C gestures, C1, C2 and C3, are in-phase coupled to the V gesture, and anti-phase

coupled in order, as shown in Figure 2.13(d2, D2). As a result, the onset of the V gesture occurs after the midpoint between the onsets of the C1 and C2 gestures.

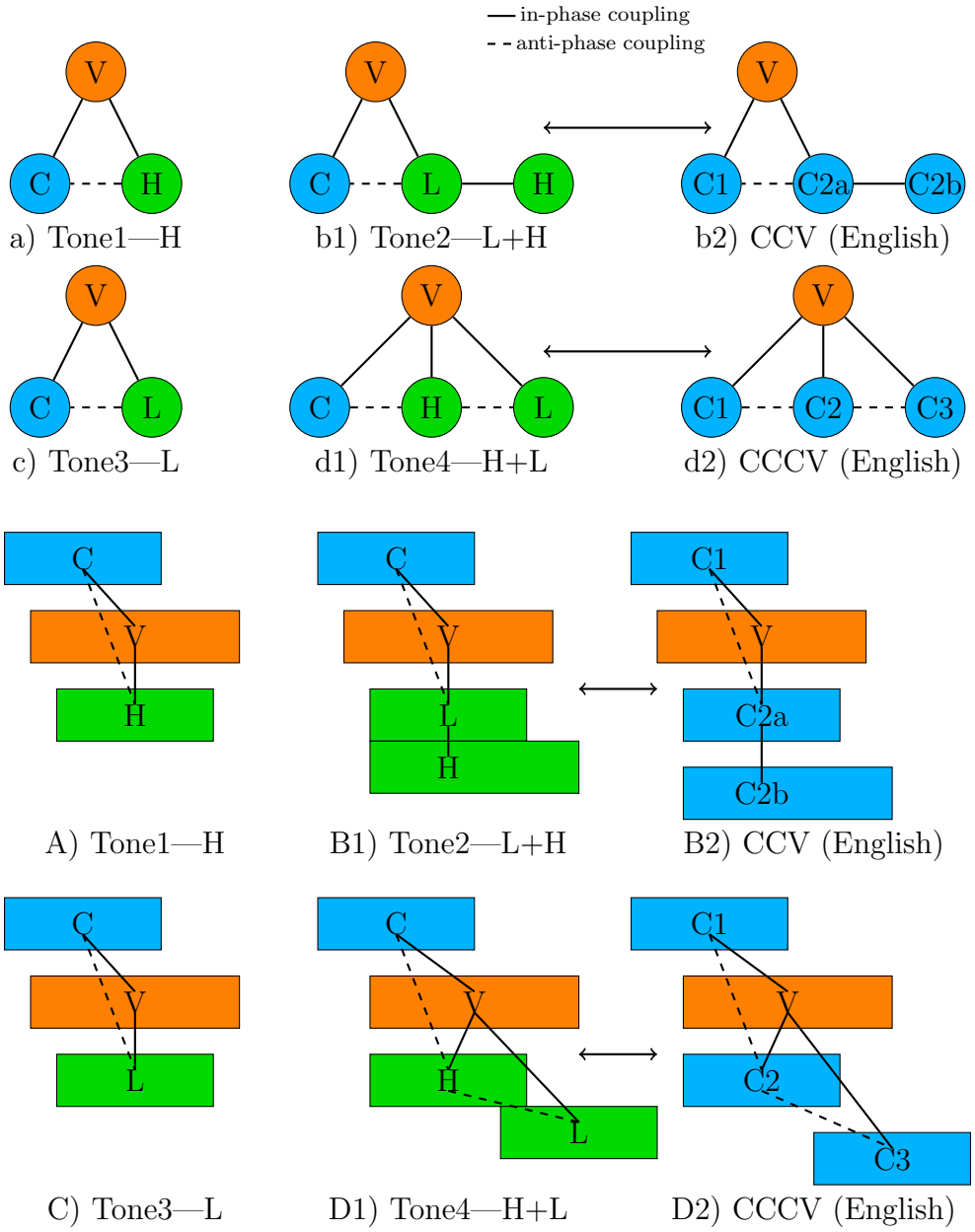


Figure 2.13: Coupling graphs (a, b1, c, d1) and gestural scores (A, B1, C, D1) of four tone-bearing syllables in Mandarin based on Gao (2008). Tone2-bearing syllable (b1, B1) resemble English CCV syllables (b2, B2), and Tone4-bearing syllables (d1, D1) resemble English CCCV syllables (d2, D2). Resemblance is indicated by double-headed arrows.

2.5 Summary of Background

The association between f_0-related events (such as lexical tones and pitch accents) and segments has been well documented in recent studies in both non-tonal and lexical tone languages. One account, Autosegmental Metrical theory, argues that f_0-related events are consistently aligned in time to segments ("anchoring sites") by way of association lines, which is the central claim of the segmental anchoring hypothesis. Note that the AM account relies solely on acoustic data. An alternative account, Articulatory Phonology, approaches tone-to-segment alignment by making reference to the speech planning mechanism. AP argues the timing control is accomplished via gestural coordination between f_0 (tone) gestures and oral articulatory gestures. Note that f_0 control is modeled as gestures that are coordinated with oral articulatory gestures.

It is against the backdrop of the gestural model of f_0 control, two experiments, i.e., Experiment 1 and Experiment 2, were carried out. Specifically, Experiment 1 (Section 3) investigates the speaker motor control of relative timing of lexical tones in an imitation study. Experiment 2 (Section 4) examines the interaction between lexical tones and intonation with a gestural approach. This dissertation can contribute to an improved understanding of f_0 control in lexical tone languages.

CHAPTER 3

EXPERIMENT 1

3.1 Introduction

The main objective is to investigate speaker motor control of relative timing between lexical tones and segments in Mandarin. Past research has shown that there exist consistent patterns of relative timing between f_0 turning points and acoustic landmarks (Xu, 1997, 1998, 1999). It is unknown how the relative timing patterns are controlled by native speakers, and to what extent the control over relative timing is influenced by perceptual or motoric categories.

The second objective is to further investigate the relationship between the relative timing and fundamental frequency of f_0 turning points. Previous research has suggested that phenomena like tonal crowding, where the previous tone undershoots its target due to the proximity of the following tone, can be accounted for under a gestural framework of tone production. However, more evidence is still needed to provide a comprehensive account of the relationship. Furthermore, the majority of previous studies approach this question in non-tonal languages by way of intonation. This experiment can further contribute to the understanding of the gestural model of f_0 control by supplementing evidence from a lexical tone language, Mandarin Chinese.

Thirty (30) native Mandarin speakers participated in an hour-long experiment comprised of a AX discrimination task, an imitation task, and a second AX discrimination task. The synthesized stimuli, all of which were disyllabic, bi-tonal sequences [ma2 ma2], varied parametrically both in the relative timing of f_0 turn-

ing points with respect to segment boundaries and in the fundamental frequency of f_0 turning points. Each bi-tonal tonal sequence had a rising-rising f_0 profile, and the parametric variation either occurs at the Low-to-High transition within the first syllable (Experiment 1A), or at the High-to-Low transition across syllable boundaries (Experiment 1B). In both experiments, the manipulated phonetic variation was applied to only one transition, i.e., Low-to-High transition in Experiment 1A and High-to-Low transition in Experiment 1B, while the other transition and the rest part of f_0 contour were kept constant.

In the first AX discrimination task, participants were asked to judge whether two tonal sequences were the same. In the imitation task, participants were instructed to imitate a bi-tonal sequence as accurately as possible. Then, participants repeated the AX discrimination task. Each participant participated in one of the two experiments, i.e., each participant was exposed to the variation in either the Low-to-High transition or the High-to-Low transition.

By asking speakers to imitate the tonal sequences with parametric variation in relative timing and fundamental frequency of the f_0 turning points, the current experiment probes the control mechanisms that coordinate lexical tone gestures and oral articulatory gestures during speech planning phase. Understanding of such a planning mechanism for lexical tone production can contribute to current models of speech motor planning.

3.2 Background

3.2.1 Imitation of Intonational Gestures in English (Tilsen et al., 2013)

Tilsen et al. (2013) investigated how native speakers of English imitated parametrically varied intonational gestures to determine the aspects of intonation contours most directly controlled by speakers. The experimenters asked speakers to imitate a synthesized name [manima] upon which a parametrically varied rise-fall f_0 contour was imposed. Specifically, the peak f_0, timing of the peak, and onset or offset f_0 were varied such that the range and the velocity of f_0 fall or rise were dissociated from the peak f_0 in subsets of the stimuli. Speakers were required to imitate the stimuli in a carrier phrase as accurately as they could. The latency between the f_0 turning points and the acoustic landmarks was measured.

The authors found that the alignment of the H tone was primarily associated with the onset of the word-initial [m], with the H onset (the start of the f_0 rise) occurring approximately 20 ms after the landmark. Moreover, this mode of association was not affected by stimulus peak timing. There was a secondary, albeit less prominent, mode of association between the H tone and the second syllable, with the H onset occurring 50-75 ms before the onset of [n]. The secondary mode was more pronounced as the f_0 peak occurred later in the stimulus. The results suggested that most speakers adopted the primary coordinative pattern, associating the H gesture with the first syllable [ma], while some speakers coordinated the H gesture with the second syllable [ni] in order to imitate stimulus peak timing.

The alignment of the L tone exhibited a more pronounced bimodal distribution. In the first three stimulus peak timing conditions, the L onset was consistently aligned 10-20 ms before the onset of [n] in the second syllable. In the last two stimulus peak timing conditions, the L onset was shifted later in time and was aligned 40-50 ms after the onset of the second [ma], the third syllable. Later analysis further suggested that the first syllable, i.e., the first [ma], was also the L anchor site for some speakers on the basis that the interval between the L onset and the first syllable was the least variable interval. The results suggested that the general distribution of the L gesture anchor shifted from the first syllable to the second or the third syllable as the peak timing increased in the stimulus.

The results, despite being acoustic in nature, spoke to a gestural model of f_0 control rather than the Segmental Anchoring Hypothesis. The Segmental Anchoring Hypothesis entails that the alignment between the target of tones and some segmental boundary is consistent such that the target is achieved at some point proximate to the segmental boundary (though not necessarily at the segmental boundary). However, it was shown earlier the L gesture was anchored to the first syllable, at least for some speakers, despite that the L onset occurred several hundred milliseconds later. Therefore, The segmental anchoring hypothesis was not entirely in line with the observed alignment patterns.

In conclusion, Tilsen et al. (2013) argued that there were several modes of coordination between intonational tone gestures and oral articulatory gestures. The differences in coordination modes reflected the categorical changes in the interactions among gestures rather than gradient variation in tone-to-segment anchoring. This is consistent with a model that governs timing of intonational tones through gestural coordination.

3.2.2 Tonal Crowding in Greek (Arvaniti et al., 2006)

Another essential premise of segmental anchoring hypothesis is that tonal targets are always achieved at some point relative to the boundary irrespective of the tonal environment. In other words, the next tonal target has no influence on the achievement of the previous tonal target under any circumstances. However, this contradict with findings related to tonal crowding. Arvaniti et al. (2006) found that under tonal crowding, a condition in which two or more tonal targets have to co-occur within the same host syllable or other tone-bearing units, Standard Greek undershot the relevant tones, resulting in adjustment of the scaling and the shifts in alignment. For example, in the tone sequence L* H- L%, the "target" of the H- tone followed by an L% tone was lower when both tones fell on the last syllable than when they fell on different syllables.

This potentially confounding issue can be solved with an alternative perspective from AP: gestural overlap (c.f. Tilsen et al., 2013). The scaling of the tonal "target" and the shifts in alignment arise because the following tone gesture is activated before the previous tone gesture reaches its target. In the case of Greek L* H- L%, the surface f_0 maximum related to the H- tone is lower under tonal crowding not because the actual target is lowered but because the L% tone is initiated before the target achievement of the H- tone. Only when the L% tone is initiated precisely after the H- tone achieves its target does the f_0 turning point correspond to the target achievement of the H- tone. Otherwise, the f_0 turning point corresponds to the initiation of the following L% tone.

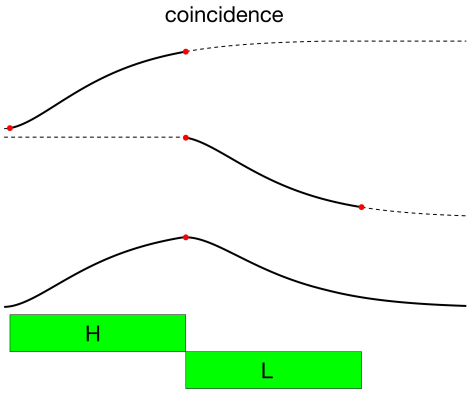
The relationship between the target achievement of the H tone and the gestural onset of the L tone in an H+L sequence is further illustrated in the left column of

Figure 3.1. In Figure 3.1(a), the H tone reaches the target at the same time as the L tone is initiated. As a result of the coincidence, the f_0 turning point corresponds to both the target achievement of the H tone and the gestural activation of the L tone. In Figure 3.1(c), the L tone is initiated before the H tone reaches the target. Because of the gestural overlap, the f_0 turning point only corresponds to the initiation of the L tone but not the target achievement or the gestural deactivation of the H tone. Moreover, the gestural overlap results in the undershoot of the H tone, i.e., f_0 maximum lower than the "real" H target, since the real H target is not achieved.

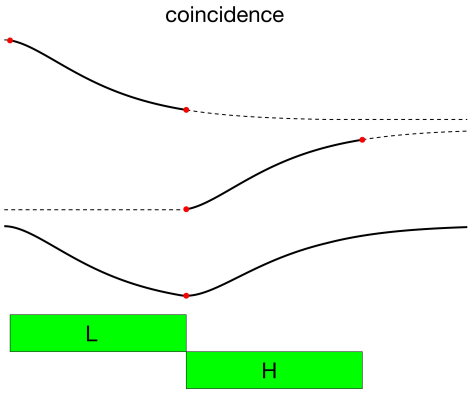
The opposite of gestural overlap is gestural underlap, in which the L tone is initiated after the H tone reaches the target. As stated in Section 2.4, the articulatory gesture is modeled as a critically damped system. In such a system, gestural underlap only leads to the delayed activation of the L tone, but not the overshoot of the H tone—the H tone remains at the target position after the target achievement. However, in an underdamped oscillating system, the gestural underlap results in a mirror image of gestural overlap: the overshoot of the H tone and the delayed initiation of the L tone. This is illustrated in Figure 3.1(e): after the H tone reaches the intended target, the f_0 contour continues to rise until the gestural onset of the L tone. Therefore, the f_0 turning point in the gestural underlap also only corresponds to the initiation of the L tone but not the target achievement or the gestural deactivation of the H tone.

Similarly, the right column of Figure 3.1 further illustrates the coincidence, gestural overlap, and gestural underlap in an L+H sequence. In the instance of coincidence (Figure 3.1(b)), the f_0 turning point corresponds to the target achievement of the L tone and the gestural activation of the H tone. The f_0 turning point only

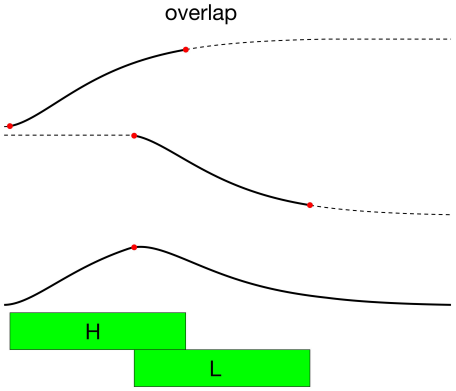
corresponds to the gestural activation of the H tone but not the target achievement of the L tone in both the gestural overlap and gestural underlap. As a result of the gestural overlap (Figure 3.1(d)), the undershoot of the L tone results a higher and earlier f_0 minimum than in the neutral condition—coincidence. As to in the case of the gestural underlap in an undamped oscillating system (Figure 3.1(e)), the L tone overshoots, yielding a lower and later f_0 minimum than in the coincidence condition.



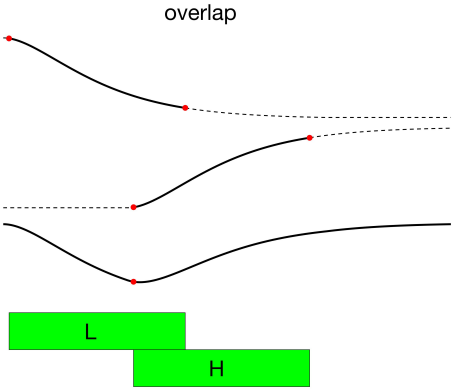
(a) Coincidence - H+L



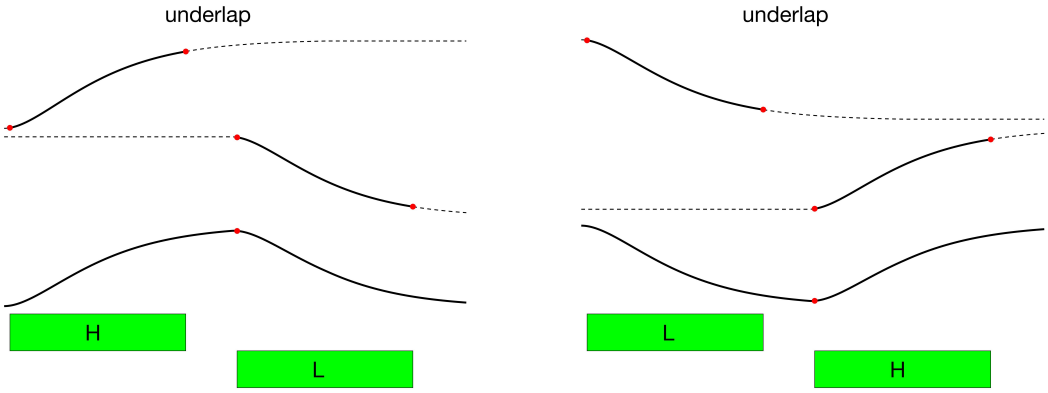
(b) Coincidence - L+H



(c) Gestural overlap - H+L



(d) Gestural overlap - L+H



(e) Gestural underlap - H+L (f) Gestural underlap - L+H

Figure 3.1: Illustration of the relationship among the f_0 turning points, target achievement, and gestural activation in an H+L sequence (left) and an L+H sequence (right) in coincidence (top), gestural overlap (middle), and gestural underlap (bottom). The rectangle represents the gestural score, the time course during which the tonal gesture is active. Note that for gestural underlap, gestures are modeled as undamped oscillating systems, as opposed to critically damped oscillating systems for coincidence and gestural overlap.

Tone sequence	H+L		L+H	
	f_0 max.	Timing	f_0 min.	Timing
Coincidence	-	-	-	-
Gestural overlap	$\downarrow$	$\leftarrow$	$\uparrow$	$\leftarrow$
Gestural underlap	$\uparrow$	$\rightarrow$	$\downarrow$	$\rightarrow$

Table 3.1: Summary of f_0 turning points fundamental frequency and relative timing in the coincidence, gestural overlap, and gestural underlap. $\uparrow$: f_0 turning point higher than the intended target; $\downarrow$: f_0 turning point lower than the intended target; $\leftarrow$: f_0 turning point earlier than the intended target achievement time; $\rightarrow$: f_0 turning point later than the intended target achievement time.

Table 3.1 summarizes the fundamental frequency and the relative timing of the f_0 turning point in the gestural overlap and underlap with coincidence as the baseline. For the fundamental frequency of the f_0 turning point, the up arrow $\uparrow$ indicates that f_0 turning point is higher than the intended target, and the down arrow $\downarrow$ indicates lower. As with the relative timing of the f_0 turning point, the left arrow $\leftarrow$ indicates f_0 turning point occurs earlier than the intended target

achievement time, and the right arrow $\rightarrow$ indicates later.

3.3 Hypotheses and Predictions

The current experiment investigates the timing control in tone-to-segment alignment by examining how faithfully speakers imitate synthesized Tone2+Tone2 (rising-rising) stimuli that vary parametrically in the relative timing and fundamental frequency of f_0 turning points. The difference between the stimuli in Experiment 1A and II should be further noted: the stimuli vary in the relative timing of the Low-to-High f_0 transition in Experiment 1A and in the High-to-Low f_0 transition in Experiment 1B. Under the gestural model of f_0 control, the Low-to-High f_0 transition within the first tone-bearing syllable indicates the activation of the H gesture of the first Tone2, i.e., the second lexical tone gesture of a rising tone (L+H). Similarly, the High-to-Low f_0 transition after the first syllable offset indicates the activation of the L gesture of the second Tone2, i.e., the first lexical tone gesture of a rising tone (L+H).

The gestural model of f_0 control argues that the timing of lexical tones is governed through coordinative interactions among lexical tone gestures and articulatory gestures. It has also been established that the alignment between f_0 turning points and syllable boundaries is indicative of the coordination (specifically the coordination of gestural activation) between the tone gestures and the articulatory gestures associated with segments. The acoustic landmarks (such as the acoustic onset of consonants) are an approximate representation of gestural landmarks.

Therefore, the stability of gestural coordination between lexical tone gestures and oral articulatory gestures will determine whether the alignment of f_0 turning

points to segmental boundaries remains relatively constant in response to variation in stimulus. If the coordinative patterns between lexical tone gestures and oral articulatory gestures are categorically distinct, the relative timing of f_0 turning points in imitation should be nonlinearly related to variation in stimulus Otherwise, variation in stimulus f_0 timing should be faithfully reflected in imitation. It is also likely that the alignment patterns differ between different positions—within a syllable in Experiment 1A and across syllable boundaries in Experiment 1B.

The specific hypotheses and predictions are delineated below:

Hypothesis A1: The coordination between lexical tone gestures and oral articulatory gestures is categorical in nature.

Prediction A1: In imitation, variation in stimulus f_0 peak/valley timing will not be linearly reflected in imitation; speakers will conform to one or two modes of relative timing despite variation in stimulus. Moreover, discrimination performance will be non-linearly related to stimulus differences.

Hypothesis A2: Speakers control the coordination between lexical tone gestures and oral articulatory gestures through some mechanism that allows for gradient specification of relative timing.

Prediction A2: In imitation, variation in stimulus f_0 peak/valley timing will be linearly reflected in imitation, resulting in linear changes in imitation f_0 peak/valley timing. Moreover, discrimination performance will be consistently high across the stimulus continuum.

A relevant question is how the temporal location of f_0 turning points varies with the fundamental frequency. The gestural model of f_0 control argues that in a

bi-tonal sequence, the f_0 turning point is indicative of the activation of the second tone gesture rather than the target achievement of the first tone gesture. Therefore, the achievement of the first tone target is modulated by the shifts in the alignment of the second tone gesture.

Hypothesis B1: The target achievement of a tone is influenced by gestural overlap or underlap .

Prediction B1: The fundamental frequency of f_0 turning points varies with the temporal location. For the Low-to-High turning point, the earlier/later the alignment, the higher/lower the fundamental frequency. For the High-to-Low turning point, the earlier/later the alignment, the lower/higher the fundamental frequency.

Hypothesis B2: The tonal target is always achieved at some point relative to the boundary irrespective of the following tone.

Prediction B2: The fundamental frequency and the temporal location of f_0 turning points are independent of one another.

3.4 Methodology

3.4.1 Participant Statistics

30 (Table 3.2) native Mandarin speakers participated in this experiment in the sound booth in the Cornell Phonetics Lab at Cornell University. Among the thirty participants, 23 were female and six were male. The disproportionality of sex was

not by design, it was believed that it would not have an effect on the results. All the participants were students from Cornell University at the time of the experiment. None of the participants reported hearing or speech problems. All the participants were financially compensated.

Sex	Experiment 1A	Experiment 1B	Total
Female	12	11	23
Male	3	3	6
Total	15	14	29

Table 3.2: Participants statistics of Experiment 1.

3.4.2 Stimuli Construction

In both experiments, the stimulus was [ma2 ma2]—a disyllabic, bi-tonal sequence with a rising + rising f_0 profile. As mentioned in Xu (1997), f_0 fell slightly into the vowel of the first [ma] before rising, reaching its peak after the acoustic offset of the first [ma]. As a result, the second rising tone, nominally speaking, did not start until well into the latter half of the second [ma], which is usually referred to as the carryover effect. The segmental durations in the disyllabic sequence were averaged based on productions of a native Mandarin speaker.

Stimuli were constructed using the diphone-based MBROLA speech synthesizer, which allows parametric specifications of duration and f_0 parameters separately via PSOLA. American English voices *us1* and *us* were used to synthesized female and male stimuli, respectively. This is because of 1) the lack of Mandarin Chinese voice databases (especially male) for the MBROLA synthesizer (compared to those of American English) and 2) the respective near-identicalness between American English [m] and [a] (the open central unrounded vowel) and Mandarin

[m] and [a]. Two sets of gender-specific stimuli were synthesized for each experiment, resulting in four sets of 80 stimuli (20 different [ma2 ma2] in each set).

Duration Parameters A female native Mandarin speaker produced [ma2 ma2] in isolation 50 times. The average duration of each segment in the disyllabic sequence is listed in Table 3.3:

σ	Syllable 1 (σ_1)		Syllable 2 (σ_2)	
Segment	[m]	[a]	[m]	[a]
Duration	125 ms	187 ms	125 ms	300 ms
Total (acc.)	125 ms	312 ms	437 ms	737 ms

Table 3.3: Duration parameters of a synthesized stimulus

Note that the segmental durations were kept constant across all synthesized stimuli.

f_0 **Parameters** Similar to duration parameters, female f_0 parameters were also averaged based on the productions of the model female speaker. Moreover, male f_0 parameters were modified by shifting female f_0 parameters down by 90 Hz and by reducing the f_0 range by half.

f_0-related landmarks are illustrated in Figure 3.2. Specifically, in Experiment 1A, parametric variation was only applied to the Low-to-High transition (*f0 TP1* in Figure 3.2) within the first Tone2-bearing syllable, while the onset, *f0 TP2*, f0 TP3, and offset were kept constant within the set. Similarly, in Experiment 1B, parametric variation was only applied to the High-to-Low transition (*f0 TP2* in Figure 3.2) across the syllable boundary, while the onset, *f0 TP1*, f0 TP3, and offset were kept constant within the set.

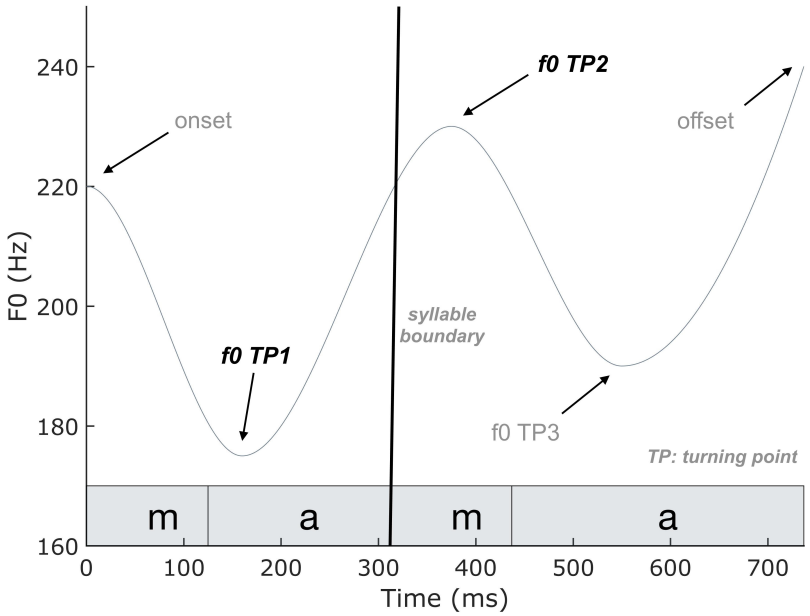


Figure 3.2: f_0-related acoustic landmarks. *f0 TP1* and *f0 TP2* were parametrically varied in Experiment 1A and 1B, respectively. The other acoustic landmarks were kept constant.

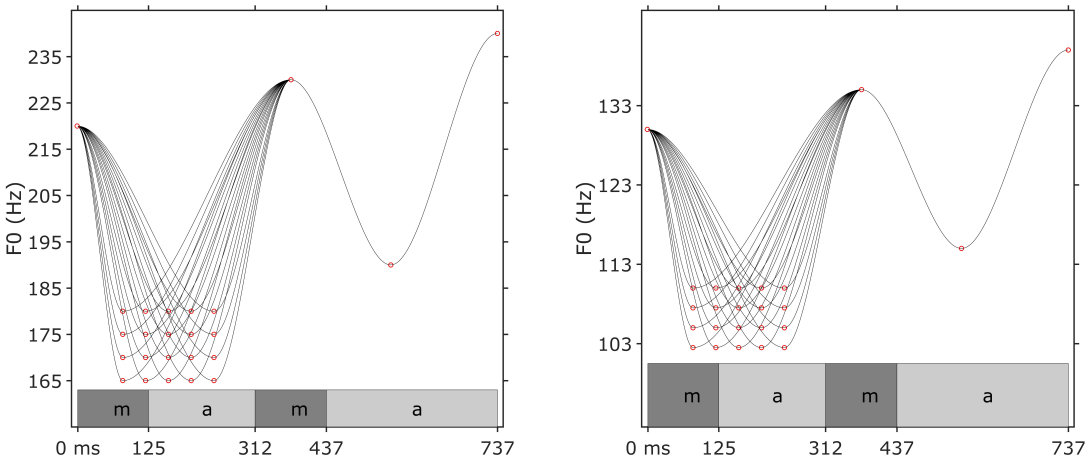
In Experiment 1A, parametric variation of f_0 consisted of two parts: the relative timing of f_0 turning points (TP[1]) to the acoustic landmark (the acoustic onset of first [ma1]) and the fundamental frequency. For the female stimuli in Experiment 1A, the relative timing of TP1 to the acoustic onset of the first [ma1] ranged from 80 ms to 240 ms, and the fundamental frequency of TP1 ranged from 165 Hz to 180 Hz. Therefore, the segmental sequence [ma ma] was synthesized with 20 f_0 contours (5 steps of relative timing $\times$ 4 steps of fundamental frequency). The male stimuli were derived by shifting and shrinking the female f_0 parameters (only fundamental frequency): the baseline f_0 was shifted downwards by 90 Hz, and the f_0 range was reduced to half. As a result, the relative timing continuum was identical to that of female stimuli, whereas the fundamental frequency ranged from 102 Hz to 110 Hz. There were also 20 synthesized male stimuli in Experiment 1A. Both female and male f_0 parameters in Experiment 1A are listed in Table 3.4.

[1] "Turning point" and "TP" are used interchangeably in the rest of the text.

	Onset		TP1		TP2		TP3		Offset	
Sex	Time	F0	Time (ms)	F0 (Hz)	Time	F0	Time	F0	Time	F0
Female	0	220	80, 120, 160, 200, 240	165, 170, 175, 180	375	230	550	190	737	240
Male	0	130	80, 120, 160, 200, 240	102, 105, 108, 110	375	135	550	115	737	140

Table 3.4: f_0 parameters in Experiment 1A: TP1 (first turning point, shaded columns) vary in five steps of relative timing and four steps of fundamental frequency; other landmarks were kept constant.

The f_0 contours were generated by evaluating the piecewise polynomial forms of the cubic spline interpolants between any two adjacent landmarks (*spline* and *ppval* in MATLAB). The synthesized female and male f_0 contours in Experiment 1A are shown in Figure 3.3(a) and Figure 3.3(b), respectively.



(a) female stimuli

(b) male stimuli

Figure 3.3: Twenty (20) synthesized stimuli for female (a) and male (b) participants in Experiment 1A. Red circles represent landmarks.

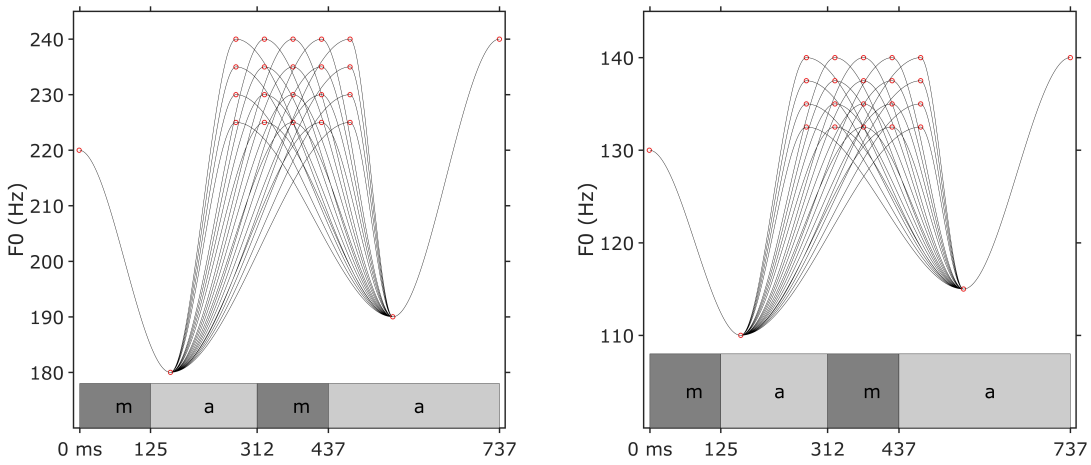
The stimuli in Experiment 1B were constructed in a similar way. For the female stimuli in Experiment 1B, the relative timing of TP2 to the acoustic onset of the first [ma1] ranged from 275 ms to 475 ms, and the fundamental frequency of TP2

ranged from 225 Hz to 240 Hz. The time step between any two adjacent TP2 of the same fundamental frequency was increased to 50 ms (40 ms in Experiment 1A). This is because the second [ma2] was approximately 1.3 times longer than the first [ma2] thanks to final lengthening. Similarly, the male stimuli were derived by shifting and shrinking the female f_0 parameters (only fundamental frequency): the baseline f_0 was shifted downwards by 90 Hz, and the f_0 range was reduced to half. As a result, the relative timing continuum was identical to that of female stimuli, whereas the fundamental frequency ranged from 132 Hz to 140 Hz. Therefore, there were also 20 synthesized female and male stimuli in Experiment 1B. Both female and male f_0 parameters in Experiment 1B are listed in Table 3.5.

	Onset		TP1		TP2		TP3		Offset	
Sex	Time	F0	Time	F0	Time (ms)	F0 (Hz)	Time	F0	Time	F0
female	0	220	160	180	275, 325, 375, 425, 475	225, 230, 235, 240	550	190	737	240
male	0	130	160	110	275, 325, 375, 425, 475	132, 135, 138, 140	550	115	737	140

Table 3.5: f_0 parameters in Experiment 1B: TP2 (second turning point, shaded columns) vary in five steps of relative timing and four steps of fundamental frequency; other landmarks were kept constant.

The f_0 contours were generated the same way as in Experiment 1A. The synthesized female and male f_0 contours in Experiment 1B are shown in Figure 3.4(a) and Figure 3.4(b), respectively.



(a) female stimuli

(b) male stimuli

Figure 3.4: Twenty (20) synthesized stimuli for female (a) and male (b) participants in Experiment 1B. Red circles represent landmarks.

3.4.3 Experiment Sessions

Each speaker participated in either Experiment 1A or Experiment 1B and therefore was only exposed to one set of stimuli that matched the speaker's gender throughout the experiment. The instructions were given in Mandarin Chinese and it was emphasized that the whole experiment would be conducted solely in Mandarin Chinese. Each experiment lasted about an hour, comprised of three sessions: an AX discrimination task, an imitation task, and a second AX discrimination task. The two AX discrimination tasks were identical to each other except that the order in which the stimulus pairs occurred was random, and each took about eight minutes. The imitation task lasted for about forty-five minutes.

Session I: First AX discrimination task The first session was the first AX discrimination task, which was conducted using the E-Prime software (Scheider

et al., 2012). Participants, who were seated in front of a computer monitor throughout the experiment, responded to a series of stimulus pairs that differed in turning point timing but not in fundamental frequency. They were instructed to press "1" for two stimuli perceived as the same, and "2" for different. Two stimuli with different turning point frequencies were not used in the AX discrimination task because of both time constraints and relatively high degree of discrimination between two stimuli with different turning point frequencies in the pilot experiments. Every stimulus pair occurred twice. Therefore, the AX discrimination task consisted of 80 pairs (= 5 steps of relative timing × (5-1) × 4 steps of fundamental frequency) of stimuli occurring in random order. Participants were instructed to make a decision as soon as possible provided that it was not a random guess.

Session II: Imitation task The second session was the imitation task, which was conducted in MATLAB (The MathWorks, Inc., 2016). The imitation task was divided into blocks. Each block consisted of 20 (= 5 steps of relative timing × 4 steps of fundamental frequency) trials, meaning that each block would exhaust all stimuli for that experiment. Depending on the participant, between 15 and 18 blocks of imitation were completed in this session.

Within in each block, the order of the 20 stimuli was randomized. In each trial, the monitor screen started out in grey. While the screen remained grey, a stimulus was played. The participant imitated the stimulus after the screen turned into green. The screen remained green for two seconds before it changed back into grey, indicating the end of the current trial and the beginning of the next trial. Using screen colors to indicate the start of the trial and the imitation serves to mitigate the negative effects brought out by the repetitive nature of the task. The participant was instructed to imitate the disyllabic, bi-tonal sequence as accurately as

possible once the screen turned green. During the familiarization phase, if a participant was judged to be pausing between the two syllables or speaking unnaturally, the experimenter would demonstrate how to imitate the stimulus without pause in a natural and/or colloquial way until the participant could perform the intended task.

For every two trials after the first five trials, participants received feedback (an accuracy score) on the accuracy of his/her imitation of that particular trial. The higher the score, the closer resemblance between the imitation and the stimulus. The accuracy score appeared on the upper left corner after the screen turned grey from green. It then disappeared after one second while the screen stayed grey, indicating the start of the next trial. Participants were encouraged to maintain or modify their imitation based on the accuracy scores. Monetary incentives were added to the base compensation for scoring high (>65 out of 100).

The accuracy feedback algorithm consists of two parts: detecting the onset and the offset of [ma2 ma2] in the imitation, and evaluating the distance between the imitation and the stimulus. Following (Tilsen and Arvaniti, 2013), the audio signal, collected at a sampling rate of 22 kHz, was first sent to a fourth-order Butterworth IIR band-pass filter that allowed frequencies between 100 Hz and 4000 Hz to pass. This frequency range served to preserve both consonantal and vocalic energy, therefore was desirable to detect the alternation between the actual signal and background noise. The effect of the 100 Hz high-pass was to admit the strong low frequency resonance (about 200 Hz) of [m], and also to emphasize to some extent the contribution of f_0 and thereby to increase the presence of voicing in the signal. Then, the magnitude of the signal (containing both consonants and vocalic nuclei), the output of the bandpass filter passing the original audio, was

sent to a fourth-order low-pass filter with a cut-off frequency of 10 Hz. This was done to ensure that the detected region was no shorter than 100 ms. Given that the stimuli duration was set at 737 ms, this extra step guaranteed a shorter region, such as the first [ma2], would not be likely to be treated as the whole simulation.

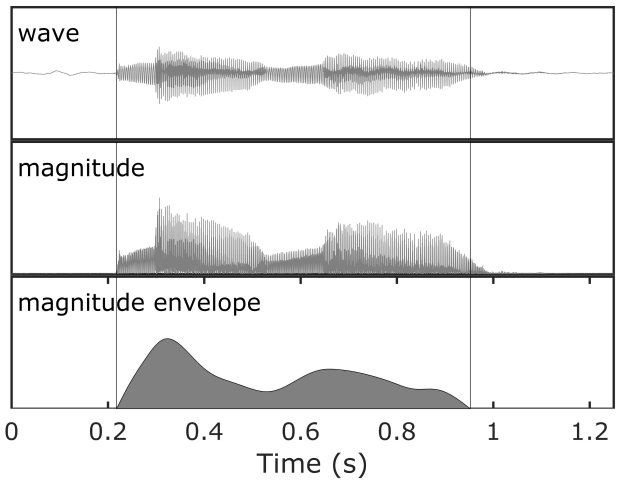


Figure 3.5: Detecting the acoustic onset and offset by passing the audio to a band-pass filter and subsequently a low-pass filter. Top: Example waveform of a [ma2 ma2] imitation; middle: Magnitude of the signal after coming out of a fourth-order filter with a passband of [100 Hz, 4000 Hz]; bottom: Magnitude envelope, which is the result of low-pass filtering the magnitude. The vertical lines denote the acoustic onset and offset according to the zero-crossings of the magnitude envelope.

In Figure 3.5, the top panel shows the original waveform of a [ma2 ma2] imitation. After filtering the waveform through a fourth-order Butterworth filter with a passband of [100 Hz, 4000 Hz], the magnitude of the filtered waveform is shown in the middle panel. The bottom panel shows the envelope of the magnitude that comes out of the a fourth-order low-pass Butterworth filter with a cut-off frequency of 10 Hz. The vertical line on the left and on the right, coinciding with the zero-crossings of the magnitude envelope, indicate the acoustic onset and offset of [ma2 ma2], respectively. Due to the relative simplicity of the speech materials, the accuracy of acoustic boundary detection was high.

The disyllable [ma2 ma2] was extracted from both a given stimulus and the corresponding imitation. The distance between stimulus and imitation was represented by the maximal cross-correlation. The maximal correlation score of this imitation and all the previous maximal correlation scores were further standardized. The accuracy score of one imitation was derived as the value of the standard normal cumulative distribution function at the standardized maximal correlation score of the imitation. The bonus compensation was tied to the accuracy score for every imitation. Hypothetically, if a participant did not change his/her imitation strategy throughout the experiment, he/she would receive 50 for each and every trial, which would not meet the bonus cut-off (65). Therefore, the accuracy score algorithm was set up to encourage participants to actively modify imitation strategies to strive for higher bonus compensation.

Session III: Second AX discrimination task The third session was the second AX discrimination task, which is identical to the first AX discrimination task in the first session. The only difference between the two sessions was the order with which the stimuli pairs occur. This session was set up to evaluate if participants adapted to the stimulus distribution after slightly less than one hour of intensive and exclusive exposure to [ma2 ma2].

3.4.4 Data Processing

Segmentation: Forced Alignment The Hidden Markov Model Toolkit (HTK; Young et al., 2006) was used to carry out segmentation for each imitation. Mel Frequency Cepstral Coefficients (MFCCs—16 were used in the current case), derived from Fast Fourier Transform of the log spectra, were extracted for each imitation.

A mini dictionary that contained *m1*, *a1*, *m2*, *a2*, *sil* ('silence') was created. For each speaker, between five to ten imitations (roughly 2%) were manually labeled. Relying on MFCCs, monophone Hidden Markov Models (HMMs) were created for each label. Moreover, HTK provides speaker adaptation with *HRest* and *HVite* so that with a small amount of enrollment data, the HMMs can be customized to speaker-specific characteristics (Young et al., 2006).

The alignment accuracy was ideal, as seen in Figure 3.6.

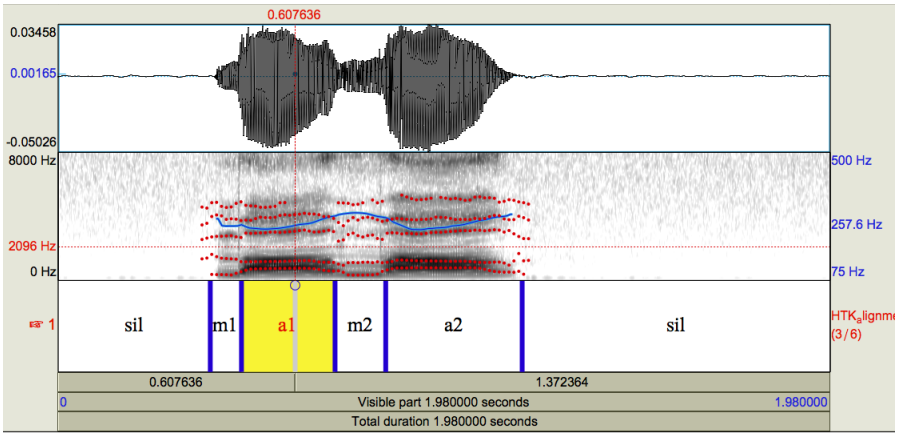
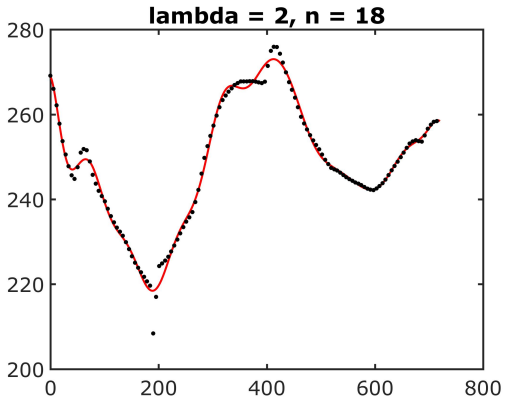


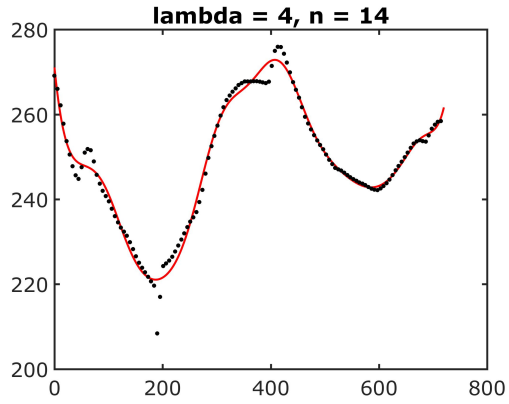
Figure 3.6: Example of using HTK to segment an imitation.

f_0 Tracking For each imitation, real-time f_0 in Hz was converted directly from vocal periods using *ProsodyPro*, a Praat-based, large-scale prosody analysis tool (Xu, 2013). The vocal periods were largely marked by Praat, with the exception of some manual corrections due to pitch-tracking anomalies. Raw f_0 was further processed using MATLAB-based f_0-processing algorithms. Imitations in which there were gaps longer than 20 ms were thrown out. Linear interpolation was then applied to the f_0 contour with 5-ms intervals. Outlier pitch track values, i.e., values that were more than three standard deviations away from the distribution mean, or values that were more than 15 Hz away from their preceding or following value, were excluded.

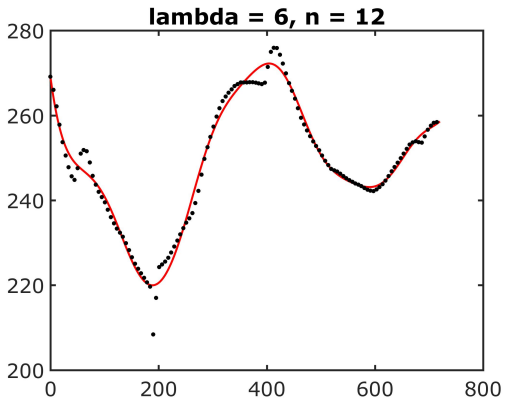
The coarsely processed f_0 contours were further smoothed by adopting *B-spline* (short for basis spline) smoothing (de Boor, 2001). A B-spline basis is a series of polynomial functions. By assigning each B-spline an appropriate weight and summing all the weighted B-splines, the resulting curve is a continuous function that approximates the f_0 contour in question. Two parameters determined the degree of smoothing: the number of B-splines n and the preference for overfitting over underfitting λ. The larger the n, the more details on the f_0 contour. The larger the λ, the more smoothing of the f_0 contour. To strike a balance between overfitting and underfitting is crucial in curve-fitting because the former results in micro-prosodic details while the latter risks losing important phonetic details such as f_0 turning points. The two parameters were chosen based on Gubian et al. (2015), with the help of visual inspection. As can be seen in Figure 3.7, as λ increased and as n decreased, the less and less phonetic details were preserved of an raw f_0 sequence. The parameters $(\lambda = 6, n = 12)$ in Figure 3.7(c) was chosen as a good compromise to fit all the f_0 curves.



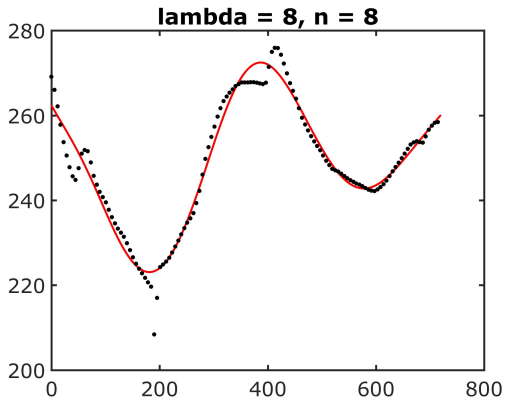
(a) $\lambda = 2, n = 18$



(b) $\lambda = 4, n = 14$



(c) $\lambda = 6, n = 12$



(d) $\lambda = 8, n = 8$

Figure 3.7: B-spline fiitting an f_0 contour with four sets of parameters. (a) and (b): overfitting; (d): slightly underfitting; (c) a good compromise.

3.4.5 Data Analysis

In general, the data analyses are structured along two dimensions: TP relative timing and TP fundamental frequency. The notional conventions are summarized below:

Dur (ms): the duration of the tone-bearing syllable;

RelTiming (ms): the latency between the acoustic onset of the tone-bearing syllable and the corresponding TP (TP1 for the first [ma2], and TP2 for the second [ma2]);

RelTimingDist (ms): the distance in RelTiming between the stimulus and the imitation;

CRelTiming (ms): the centralized RelTiming, derived by subtracting RelTiming by the mean RelTiming for each participant;

CRelTimingDist (ms): the distance in CRelTiming between the stimulus and the imitation;

F0 (Hz):[2] the fundamental frequency at the TP;

F0Dist (Hz): the distance in F0 between the stimulus and the imitation;

CF0 (Hz): the centralized F0, derived by subtracting F0 by the mean F0 for each participant.

CF0Dist (Hz): the distance in CF0 between the stimulus and the imitation;

Moreover, suffixes "-A" and "-B" indicate Experiment 1A and 1B, respectively; "-s" and "-i" stand for the stimulus and imitation, respectively. Therefore, RelTiming-A-i represents the latency (in ms) of TP1 in imitations, whereas CRelTiming-B-s stands for the centralized latency (in ms) of TP2 in stimuli.

Two more notational conventions are used exclusively for stimuli:

TimingStep: the stimulus relative timing in step, with smaller steps number indicating earlier TP;

F0Step: the stimulus fundamental frequency in step, with smaller steps number indicating lower TP.

Table 3.6 and 3.7 illustrate the parametric variation in the TP relative timing and fundamental frequency in the aforementioned conventions in both Experiment 1A and 1B.

[2]Note that "f_0" refers to the contour in general, whereas "F0" refers exclusively to the fundamental frequency at the TP.

	Experiment 1A	Experiment 1B
TimingStep	RelTiming-A-s	RelTiming-B-s
1	80 (80) ms	-37 (275) ms
2	120 (120) ms	13 (325) ms
3	160 (160) ms	63 (375) ms
4	200 (200) ms	113 (425) ms
5	240 (240) ms	163 (475) ms

Table 3.6: Parametric variation in the TP relative timing illustrated in TimingStep-s and RelTiming-s (followed by the latency in ms between the TP and the acoustic onset of the stimulus, i.e., the acoustic onset of the first [ma2]), in Experiment 1A and 1B.

	Experiment 1A		Experiment 1B	
F0Step	F0-A-s		F0-B-s	
	Female	Male	Female	Male
1	165 Hz	102 Hz	225 Hz	132 Hz
2	170 Hz	105 Hz	230 Hz	135 Hz
3	175 Hz	108 Hz	235 Hz	138 Hz
4	180 Hz	110 Hz	240 Hz	140 Hz

Table 3.7: Parametric variation in the TP fundamental frequency illustrated in F0Step-s and F0-s in Experiment 1A and 1B.

3.5 Results

Section 3.5.1 presents the AX discrimination results; Section 3.5.2 presents the imitation results, which are further structured along the line of TP relative timing and TP F0; Section 3.5.3 illustrates the link between discrimination and imitation; Section 3.5.4 explores the relationship between TP RelTiming-i and TP F0-i in imitation; Section 3.5.5 examines the speaker-induced variation in imitation.

3.5.1 AX Discrimination

Session and Participant Overall, participants in Experiment 1A perform worse than in Experiment 1B. The mean discrimination success is below 50% for the former, and above 50% for the latter. Individual participants differ significantly in the change of the performance from Session I (before the imitation session) to Session II (after the imitation session): some participants exhibit an increase in discrimination success, while other exhibit a decrease (Figure 3.8 and Table 3.8).

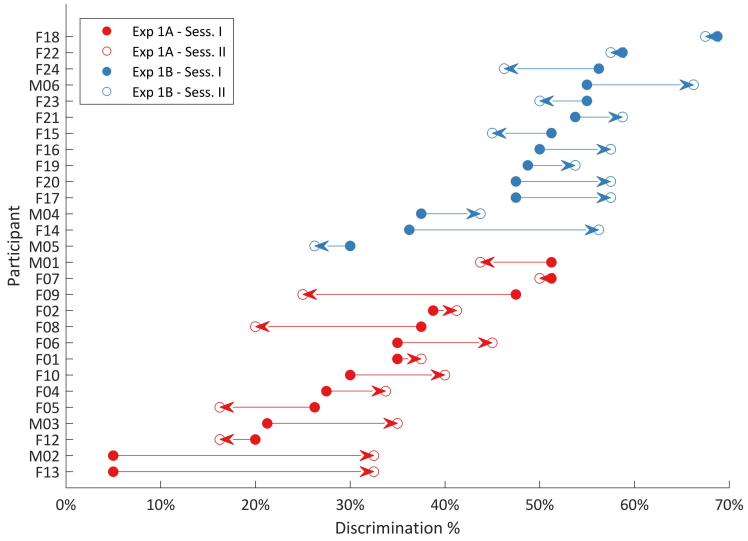


Figure 3.8: Changes in discrimination performance (in %) from Session I to Session II. Participants are sorted by discrimination performance in Session I for Experiment 1A and 1B, respectively. Red represents Experiment 1A and blue represents Experiment 1B.

Exp. 1A				Exp. 1B			
Par.	Sess. I	Sess. II	Δ	Par.	Sess. I	Sess II	Δ
F13	5	32.5	27.5	M05	30	26.25	-3.75
M02	5	32.5	27.5	F14	36.25	56.25	20
F12	20	16.25	-3.75	M04	37.5	43.75	6.25

M03	21.25	35	13.75	F17	47.5	57.5	10
F05	26.25	16.25	-10	F20	47.5	57.5	10
F04	27.5	33.75	6.25	F19	48.75	53.75	5
F10	30	40	10	F16	48.75	53.75	7.5
F01	35	37.5	2.5	F15	51.25	45	-6.25
F06	35	45	10	F21	53.75	58.75	5
F08	37.5	20	-17.5	F23	55	50	-5
F02	38.5	41.25	2.5	M06	55	66.26	11.25
F09	47.5	25	-22.5	F24	56.25	46.26	-10
F07	51.25	50	-1.25	F22	58.75	57.5	-1.25
M01	51.25	43.75	-7.5	F18	68.75	67.5	-1.25

Table 3.8: Discrimination performance (in %) by individual participant in Session I and Session for Experiment 1A (left) and Experiment 1B (right). Participants are sorted by discrimination performance in Session I for Experiment 1A and 1B, respectively.

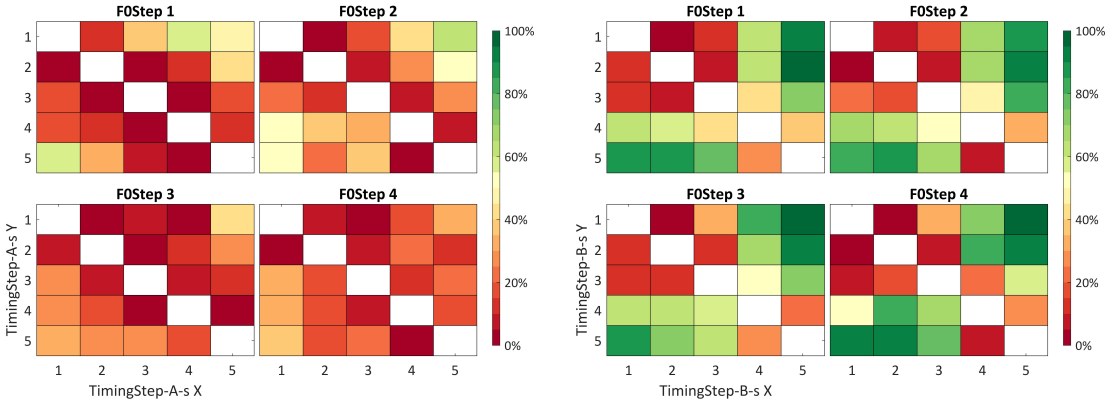
Factor	Coeff.	t	d.f.	p-value
Experiment 1B	18.9	4.47	52	< 0.001
Session II	2.68	0.87	52	0.39
Experiment 1B : Session II	0.71	0.16	52	0.87

Table 3.9: Linear mixed effect model on discrimination performance (in %). Only fixed terms, i.e., Experiment, Session, and their interaction, are shown. The random term Participant is not shown.

A linear mixed effect regression model is then fitted to the above discrimination data, with Experiment, Session, and their interaction as the fixed terms, and Participant as the random term, as shown in Table 3.9. The results show that only the effect of Experiment is significant on the discrimination performance—the

discrimination success in Experiment 1B is 18.9% higher than that in Experiment 1A ($t(52) = 4.47, p < 0.001$). Neither the effect of Session nor the interaction between Experiment and Session is significant: for Session, $t(52) = 0.87, p > 0.05$; for the interaction between Experiment and Session, $t(52) = 0.16, p > 0.05$. Since the discrimination performance is not influenced by Session, subsequent analyses will pool together the discrimination results of both sessions.

F0Step-s and DTimingStep-s The discrimination performance is further broken down by stimulus pair, as shown in Figure 3.9. Figure 3.9(a) shows the discrimination performance in Experiment 1A, and Figure 3.9(b) shows that in Experiment 1B. Since the perception tasks only asked the participants to discriminate two stimuli with the same TP F0, the results are shown in four small heatmaps, each of which corresponds to one F0Step-s.



(a) Experiment 1A (b) Experiment 1B

Figure 3.9: Heatmaps showing discrimination performance for each F0Step-A-s in Experiment 1A (a) and Experiment 1B. Each of the four smaller heatmaps represents one F0Step-A-s. Each cell represents a pair of stimuli (with the same F0Step-A-s) that differ in TimingStep-A-s. The X coordinate represents the stimulus that occurs first in the pair, while Y coordinate represents the stimulus that occurs second. Green indicates high discrimination success, and red indicates low.

The overall color of Figure 3.9(a) is redder than that of Figure 3.9(b), which is line with the previous result that the participants in Experiment 1B have overall higher discrimination performance. In Experiment 1A, the color of the cells ranges from dark orange to dark red, indicating that the discrimination is well below chance. The discrimination performance only reaches above chance in the case of differentiating TimingStep-A-s 1 from 5 at F0Step-A-s 1 or 2. This suggests, only when the difference between the two stimuli is large enough, and when TP F0 is low enough, the participants in Experiment 1A exhibit improved discrimination performance. In Experiment 1B, the overall color of the heatmaps are much greener. The discrimination performance that involves TimingStep-B-s 4 and 5 in general achieves well above chance.

A linear mixed effect regression model is fitted to the above discrimination data for each experiment. The fixed terms are F0Step-s, and the distance in TimingStep-s between two stimuli (DTimingStep-s) and their interaction. DTimingStep-s is used because the order of stimuli in the perception task does not affect the discrimination outcome. This is supported by the near symmetry of both Figure 3.9(a) and 3.9(b). The random term is Participant. The regression results are shown in Table 3.10.

	Factor	Estimate	tStat	d.f.	pValue
Exp. 1A	F0Step-A-s 2	-0.45%	-0.07	208	0.94
	F0Step-A-s 3	-6.25%	-0.99	208	0.33
	F0Step-A-s 4	-10.71%	-1.69	208	0.09
	DTimingStep-A-s 2	9.38%	1.48	208	0.14
	DTimingStep-A-s 3	24.55%	3.87	208	< 0.001
	DTimingStep-A-s 4	36.16%	5.71	208	< 0.001
Exp. 1B	F0Step-B-s 2	1.79%	0.36	208	0.72

F0Step-B-s 3	-1.33%	-0.27	208	0.79
F0Step-B-s 4	0%	0	208	1
DTimingStep-B-s 2	26.79%	5.38	208	< 0.001
DTimingStep-B-s 3	52.68%	10.58	208	< 0.001
DTimingStep-B-s 4	66.07%	13.27	208	< 0.001

Table 3.10: Linear mixed effect model on discrimination performance (in %) for Experiment 1A (a) and Experiment 1B (b). Only fixed terms, i.e., F0Step-s, DTimingStep-s are shown. The fixed interaction term and the random term Participant are not shown.

The regression results show that in both Experiment 1A and Experiment 1B, only DTimingStep-s affects the overall discrimination performance. In Experiment 1A, when two stimuli are 3 and 4 TimingStep-A-s away, the discrimination performance increases by 24.55% and 36.16%, respectively, compared to the baseline discrimination performance for F0Step-A-s 1 and DTimingStep-A-s 1. In Experiment 1B, the effect size of DTimingStep-s is larger: when two stimuli are 2, 3 and 4 TimingStep-A-s away, the discrimination performance increases by 26.79%, 52.68%, and 66.07%, respectively, compared to the baseline discrimination performance for F0Step-B-s 1 and DTimingStep-B-s 1.

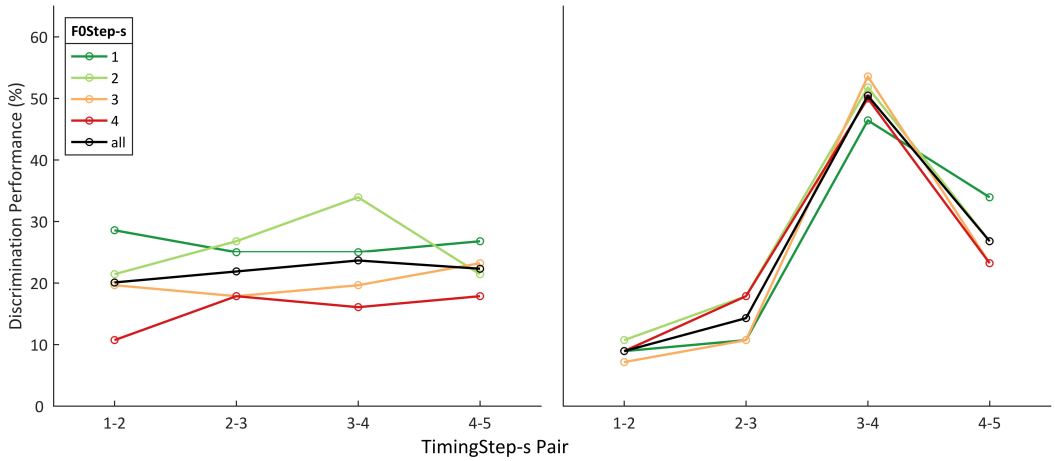


Figure 3.10: Discrimination of RelTiming-s continuum for Experiment 1A (left) and Experiment 1B (right). Each non-black line represents the discrimination function for one F0Step-s. The black line represents the overall discrimination function.

Discrimination Function The standard discrimination function is plotted for each experiment in Figure 3.10. The data points are the overall discrimination success of any two adjacent stimuli on the RelTiming-s continuum: TimingStep-s 1 and 2, 2 and 3, 3 and 4, and 4 and 5. In Experiment 1A, the discrimination performance remains roughly flat across the RelTiming-A-s continuum, fluctuating around 20%. In Experiment 1B, the discrimination performance is around 15% for the first two pairs of TimingStep-B-s (1 and 2, and 2 and 3), increases significantly to around 50% for the pair of TimingStep-B-s 3 and 4, and falls back to around 20% for the last pair of TimingStep-B-s 4 and 5. For both experiments, F0Step-s does not affect the overall shape of the discrimination function.

The discrimination results can be summarized as: 1) the overall discrimination performance does not change from Session I to Session II; 2) the participants exhibit higher discrimination performance in Experiment 1B than in Experiment 1A; 3) regardless of F0Step-s, the distance on the stimulus relative timing continuum is positively correlated with the discrimination performance; 4) for Experiment 1B,

there is a discrimination boundary between TimingStep-B-s 3 and TimingStep-B-s 4, at which the discrimination performance increases significantly, whereas for Experiment 1A, the discrimination performance remains flat across the relative timing continuum.

3.5.2 Imitation of TP

3.5.2.1 An Overview

The distribution of the imitations is illustrated in the following heatmaps. Each imitation is represented by the TP relative timing (RelTiming-i) and the centralized TP F_0 (CF0-i). Note that the centralized F_0 is used to generalize a global pattern because male and female participants have vastly different f_0. The imitations are then binned into a 20-by-20 grid of equally spaced cells. The instances that fall in each cell are counted. The more observations the cell contains, the more red it is.

Figure 3.11 shows the overall distribution of the imitations in Experiment 1A. Visual inspection shows that there is only one distinct distribution that encompasses all the imitations. This suggests the imitations in Experiment 1A do not exhibit large variation overall. Figure 3.12 further breaks down the overall distribution into five TimingStep-A-s. For the sake of comparison, each subplot in Figure 3.12 is plotted with the same scale as Figure 3.11. There are no significant differences in the subplot distributions for TimingStep-A-s 1-4. For TimingStep-A-s 5, the subplot distribution moves slightly to the right of the previous four subplot distributions.

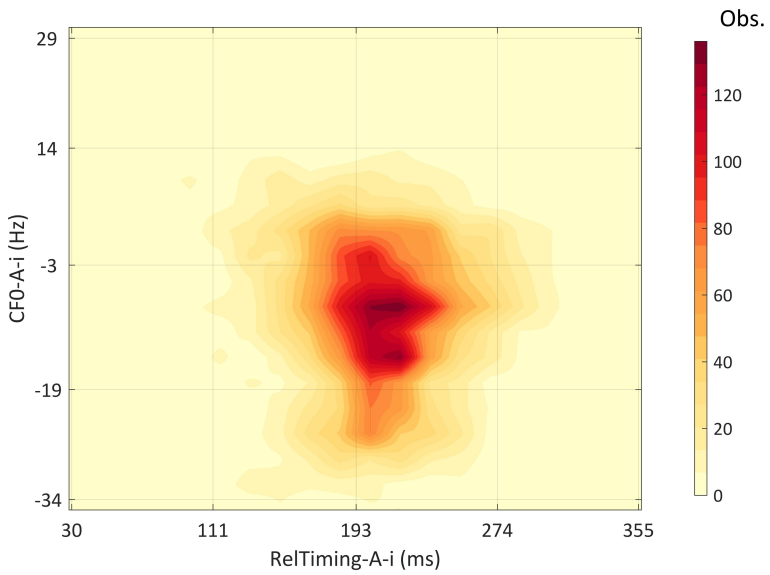


Figure 3.11: Heatmap illustrating the distribution of imitation in Experiment 1A. X axis: RelTiming-A-i; Y axis: CF_0-A-i.

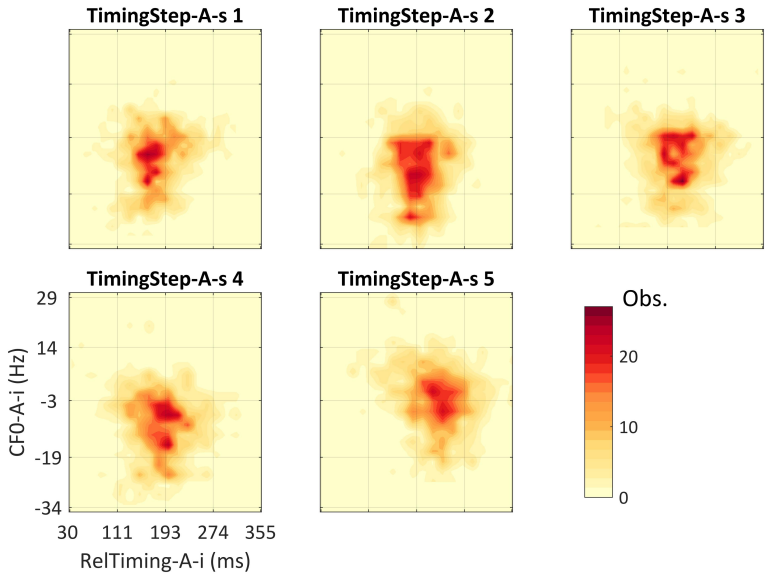


Figure 3.12: Heatmap illustrating the distribution of imitation at each TimingStep-A-s in Experiment 1A. X axis: RelTiming-A-i; Y axis: CF_0-A-i.

Figure 3.13 shows the overall distribution of the imitations in Experiment 1B. Unlike in Experiment 1A, there are two distinct distributions that fall on the diagonal line from the lower left to upper right. A majority of the imitations belong

to the distribution closer to the lower left corner. Figure 3.14 further breaks down the overall distribution into five TimingStep-B-s. The subplot distributions for the first three TimingStep-B-s mostly overlap with the lower left distribution in Figure 3.13, while the subplot distributions for the last two TimingStep-B-s mostly overlap with the upper right distribution.

To sum up, the imitations belong to one distribution in Experiment 1A but two distinct distributions in Experiment 1B. Specifically in Experiment 1B, the imitations for TimingStep-B-s 1-3 are associated with one distribution, whereas the imitations for TimingStep-B-s 4 and 5 are associated with the other.

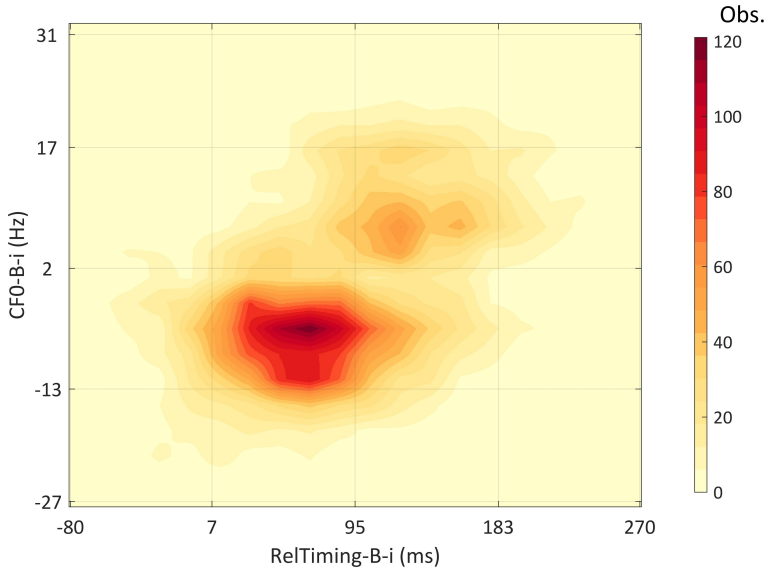


Figure 3.13: Heatmap illustrating the distribution of imitation in Experiment 1B. X axis: RelTiming-B-i; Y axis: CF_0-B-i.

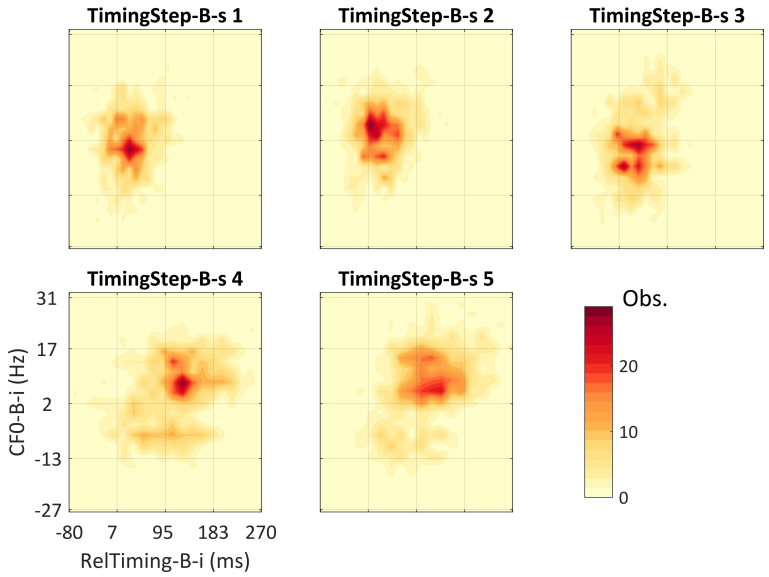


Figure 3.14: Heatmap illustrating the distribution of imitation at each TimingStep-B-s in Experiment 1B. X axis: RelTiming-B-i; Y axis: CF0-B-i.

3.5.2.2 Imitation of TP RelTiming

RelTiming-A In Experiment 1A, the distributions of RelTiming-A-i at the five TimingStep-A-s, represented by five colors, all exhibit a single peak at around 160 ms, indicating that the primary mode of association of TP1 occurs 160 ms after the onset of the first [m]. The differences in the distribution of RelTiming-A-i among the five TimingStep-A-s are rather small, with the pooled mean RelTiming-A-i ranging from 155.67 ms to 178.31 ms, in the ascending order of TimingStep-A-s (RelTiming-A-s). The pooled mean RelTiming-A-i increases as TP1 advances in the first [ma2] in the stimulus. However, judging from the error bars, the increases in the mean RelTiming-A-i are not always significant.

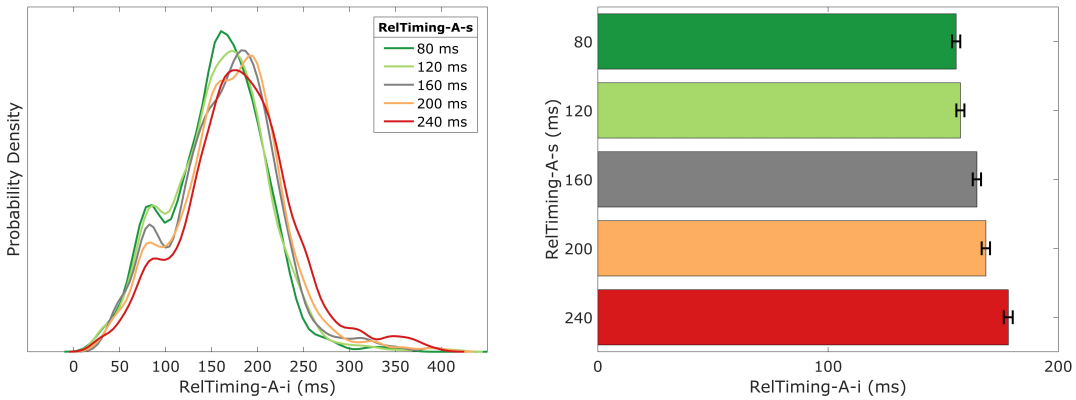
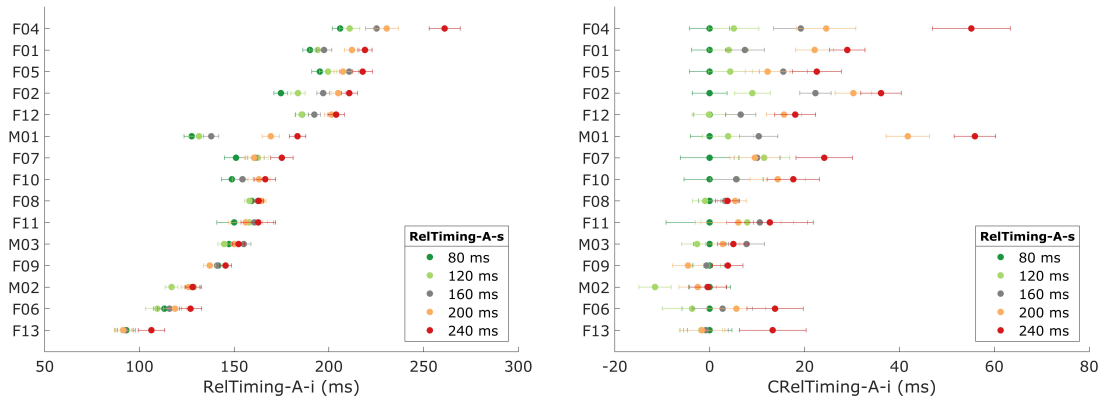


Figure 3.15: *Left*: kernel smoothing estimate (bandwidth = 10 ms) of probability density of the RelTiming-A-i at each TimingStep-A-s (RelTiming-A-s) for all participants in Experiment 1A. *Right*: bar plots displaying the pooled mean RelTiming-A-i at each TimingStep-A-s (RelTiming-A-s) for all participants in Experiment 1A. Error bars of ± 1 *s.e.* are also plotted. Each color represents one of the five TimingStep-A-s.

The mean RelTiming-A-i is further investigated at the individual participant level. Figure 3.16(a) shows the mean RelTiming-A-i at each TimingStep-A-s for each participant; Figure 3.16(b) shows the same data centered at TimingStep-A-s 3 (RelTiming-A-s = 160 ms). At the individual participant level, the mean RelTiming-A-i ranges between 90 ms to 270 ms. It appears that each participant has a preferred RelTiming-A-i value towards which the imitation of RelTiming-A-s is biased. As the RelTiming-A-s increases, the mean RelTiming-A-i also increases, despite a handful exceptions. Judging from the error bars, not all the increases are significant. However, It is safe to conclude that the further apart the two stimuli on the relative timing continuum, the more likely the difference in the mean RelTiming-A-i is significant.



(a) RelTiming-A-i (b) RelTiming-A-i centered at 160 ms

Figure 3.16: Scatter plots displaying mean RelTiming-A-i at each TimingStep-A-s (RelTiming-A-s) for each participant in Experiment 1A (in the descending order of mean RelTiming-A-i at TimingStep-A-s 5, RelTiming-A-s = 240 ms). Each color represents one of the five TimingStep-A-s. Error bars of ± 1 *s.e.* are also plotted in the corresponding colors. (a) RelTiming-A-i; (b) Centralized RelTiming-A-i (centered at TimingStep-A-s 3, RelTiming-A-s = 160 ms).

Table 3.11 shows ANOVA results of RelTiming-A-i. One-way ANOVA (Table 3.11B) shows that the effect of TimingStep-A-s on RelTiming-A-i is significant ($F(4, 4918) = 25.25, p < 0.001$). Taking into consideration the random effects of participants and the interaction between TimingStep-A-s and participants (Table 3.11C), more variation in the RelTiming-A-i is explained. The effect of the random term Participant is significant. This means that individual participants differ from one another in imitating the TP relative timing by a random value. That is, each participant's regression line (RelTiming-A-i as a function of TimingStep-A-s) is shifted up or down by a random amount. Moreover, the interaction between TimingStep-A-s and Participant is also significant, which means that participants in Experiment 1A also differ from one another in the random slope in TimingStep-A-s. If a participant has a positive random effect, the TP in the imitations is more delayed than average when the TP in the stimuli comes later. If a participant has a negative random effect, the TP in the imitations is less delayed than average

when the TP in the stimuli comes later.

(A)	TimingStep-A-s	1	2	3	4	5
	RelTiming-A-s (ms)	80	120	160	200	240
	RelTiming-A-i: mean (ms)	155.67	157.51	164.69	168. 57	178.31
	RelTiming-A-i: sem (ms)	1.79	1.75	1.80	1.81	1.94
(B)	**TimingStep-A-s (fixed)**	F(4, 4918) = 25.25, p = 9.87e-21***				
(C)	**TimingStep-A-s (fixed)**	F(4, 4848) = 21.08, p = 1.13-10***				
	Participant (random)	F(14, 4848) = 98.26, p =4.05e-34***				
	TimingStep-A-s * Participant (random)	F(56, 4848) = 1.91, p = 5.46e-5***				

Table 3.11: (A) Mean and standard error of RelTiming-A-i at five TimingStep-A-s in Experiment 1A. (B) One-way ANOVA; fixed term: TimingStep-A-s. (C) Two-way ANOVA; fixed term: TimingStep-A-s, random terms: Participant, interaction between TimingStep-A-s and Participant. Statistical significant terms are in bold.

TimingStep-A-s	1	2	3	4	5
1					
2					
3					
4					
5					

Table 3.12: Comparisons of RelTiming-A-i between pairs of TimingStep-A-s in Experiment 1A. Green indicates statistical significance; red indicates statistical non-significance.

With the random effects being controlled for, comparisons of RelTiming-A-i are conducted between pairs of TimingStep-A-s, as shown in Table 3.12. The differences in RelTiming-A-i between any of the two TimingStep-A-s are significant except for between TimingStep-A-s 1 and 2, and between 3 and 4. Note that both pairs differ by one step (40 ms) on the TimingStep-A-s (RelTiming-A-s) continuum.

To sum up the results regarding the imitation of TP relative timing for Experiment 1A: 1) the mean RelTiming-A-i increases as TP1 progresses in the stimulus; 2) speaker-specific variation (random effects) can account for a large part of the

variation in RelTiming-A-i; 3) with the random effects being controlled for, the effect of TimingStep-A-s on RelTiming-A-i is statistically significant despite the fact that the differences across different TimingStep-A-s are small: RelTiming-A-i at TimingStep-A-s 1 and 2 is significantly different from that at TimingStep-A-s 3 and 4, which in turn is significantly different that at TimingStep-A-s 5.

RelTiming-B In Experiment 1B, the distributions of RelTiming-B-i show more variation than in Experiment 1A. At the first three TimingStep-B-s, the distributions of RelTiming-B-i exhibit one single peak at around 40 ms, indicating that the primary mode of association of TP2 occurs 40 ms after the onset of the second [m]. As TP2 occurs later in the stimulus, the primary mode of association shifts to occurring after 110 ms following the onset of the second [m]. There appears to be a secondary mode of association occurs around 40 ms after the onset of the first [m], which suggests that some participants still retain this alignment pattern at the later TimingStep-B-s.

Consequently, the differences in the pooled mean RelTiming-B-i among the five TimingStep-B-s are much larger than in Experiment 1A, with the pooled mean RelTiming-B-i ranging from 38.18 ms to 133.88 ms, in the ascending order of TimingStep-B-s (RelTiming-B-s). The pooled mean RelTiming-B-i increases as TP2 advances in the second [ma2] in the stimuli. The pairwise comparisons, judging from the error bars, are mostly significant.

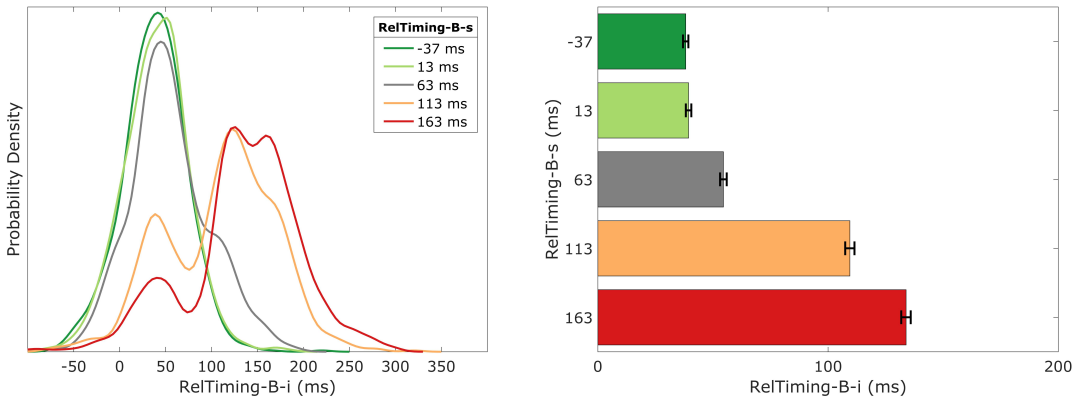
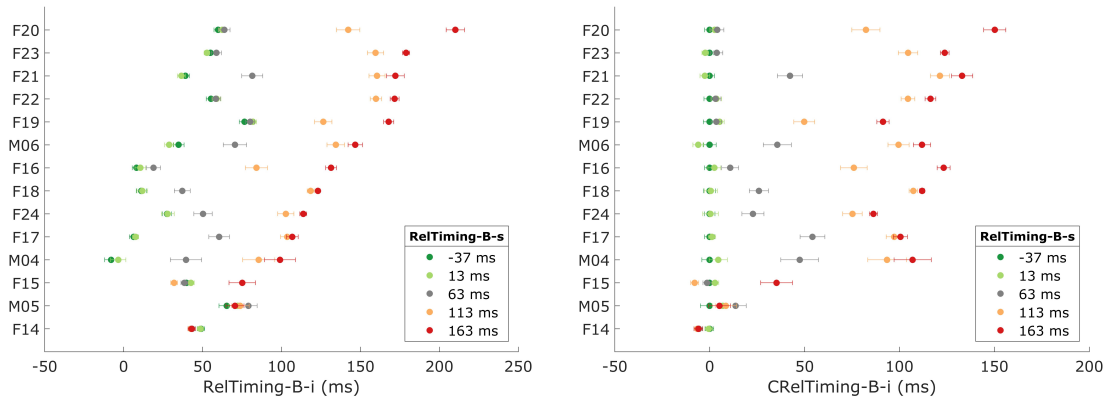


Figure 3.17: *Left*: kernel smoothing estimate (bandwidth = 10 ms) of probability density of RelTiming-B-i at each TimingStep-B-s (RelTiming-B-s) for all participants in Experiment 1B. *Right*: bar plots displaying the pooled mean RelTiming-B-i at each TimingStep-B-s (RelTiming-B-s) for all participants in Experiment 1B. Error bars of ± 1 *s.e.* are also plotted. Each color represents one of the five TimingStep-B-s.

The mean RelTiming-B-i is further investigated at the individual participant level. Figure 3.16(a) shows the mean RelTiming-B-i at each TimingStep-B-s for each participant in Experiment 1B. Figure 3.16(b) shows the same data centered at TimingStep-B-s 3 (RelTiming-B-i = 63 ms). Like in Experiment 1A, each participant has a preferred RelTiming-B-i value towards which the imitation of RelTiming-B-s is biased. However, the random effects induced by speaker-specific preferences are less prominent in Experiment 1B than in Experiment 1A. As the TimingStep-B-s increase, the mean RelTiming-B-s also increases. Judging from the error bars, the mean RelTiming-B-i at TimingStep-B-s 4 and 5 are noticeably distant from the rest, for all but two participants (F14 and M05).



(a) RelTiming-B-i (b) RelTiming-B-i centered at 63 ms

Figure 3.18: Scatter plots displaying mean RelTiming-B-i at each TimingStep-B-s (RelTiming-B-s) for each participant in Experiment 1B (in the descending order of mean RelTiming-B-i at TimingStep-B-s 5, RelTiming-B-s = 163 ms). Each color represents one of the five TimingStep-B-s. Error bars of ±1 *s.e.* are also plotted in the corresponding colors. (a) RelTiming-B-i; (b) Centralized RelTiming-B-i (centered at TimingStep-B-s 3, RelTiming-A-s = 63 ms).

Table 3.13 shows ANOVA results of RelTiming-B-i. One-way ANOVA (Table 3.13B) shows that the effect of TimingStep-B-s on RelTiming-B-i is significant ($F(4, 4481) = 727.09, p < 0.001$). This indicates the differences in RelTiming-B-i among TimingStep-B-s are highly significant. Furthermore, the random effects of Participant and the interaction between TimingStep-B-s and Participant are also highly significant (Table 3.13C).

(A)	TimingStep-B-s	1	2	3	4	5
	RelTiming-B-s (ms)	-37	13	63	113	163
	RelTiming-B-i: mean (ms)	38.18	39.43	54.56	109.48	133.88
	RelTiming-B-i: sem (ms)	1.16	1.19	1.44	2.01	2.06
(B)	**TimingStep-B-s (fixed)**	F(4, 4481) = 727.09, p = 0***				
(C)	**TimingStep-B-s (fixed)**	F(4, 4416) = 22.70, p = 1.40e-9***				
	Participant (random)	F(13, 4416) = 4.58, p =3.69e-5***				
	TimingStep-B-s * Participant (random)	F(52, 4416) = 32.54, p = 3.02e-267***				

Table 3.13: (A) Mean and standard error of RelTiming-B-i at five TimingStep-B-s in Experiment 1B. (B) One-way ANOVA; fixed term: TimingStep-B-s. (C) Two-way ANOVA; fixed term: TimingStep-B-s, random terms: Participant, interaction between TimingStep-B-s and Participant. Statistical significant terms are in bold.

With the random effects being controlled for, comparisons of RelTiming-B-i are conducted between pairs of TimingStep-B-s, as shown in Table 3.14. The differences in RelTiming-B-i between any of the two TimingStep-B-s are significant except for between TimingStep-B-s 1 and 2. This is in line with the relatively large differences in the pooled mean RelTiming-B-i across different TimingStep-B-s.

TimingStep-B-s	1	2	3	4	5
1					
2					
3					
4					
5					

Table 3.14: Comparisons of RelTiming-B-i between pairs of TimingStep-B-s in Experiment 1B. Green indicates statistical significance; red indicates statistical non-significance.

To sum up the results regarding the imitation of TP relative timing for Experiment 1A: 1) the mean RelTiming-B-i increases as TP2 progresses in the stimulus; 2) speaker-specific variation (random effects) can account for a part of the variation in RelTiming-B-i; 3) with the random effects being controlled for, the effect

of TimingStep-B-s on RelTiming-B-i is statistically significant: RelTiming-B-i between any of the two TimingStep-B-s are significantly different except for between TimingStep-B-s 1 and 2.

3.5.2.3 Imitation of TP F0

Turning to TP F0, the results will be presented in two parts: female and male.

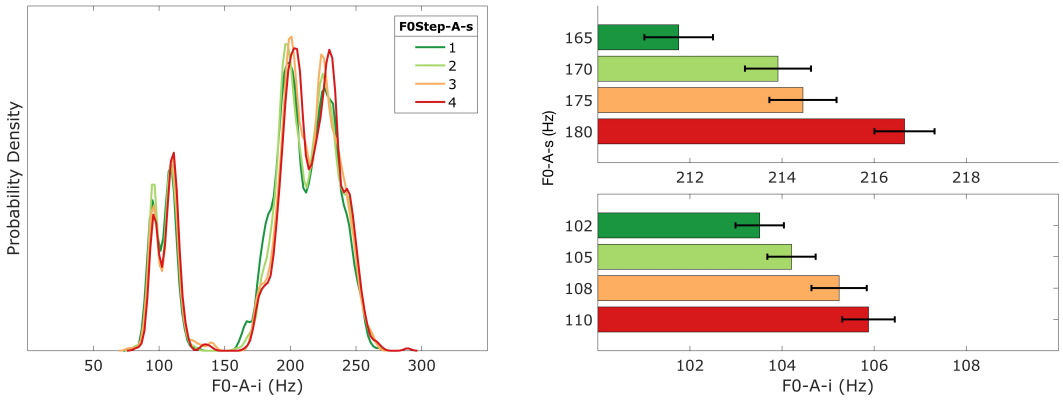
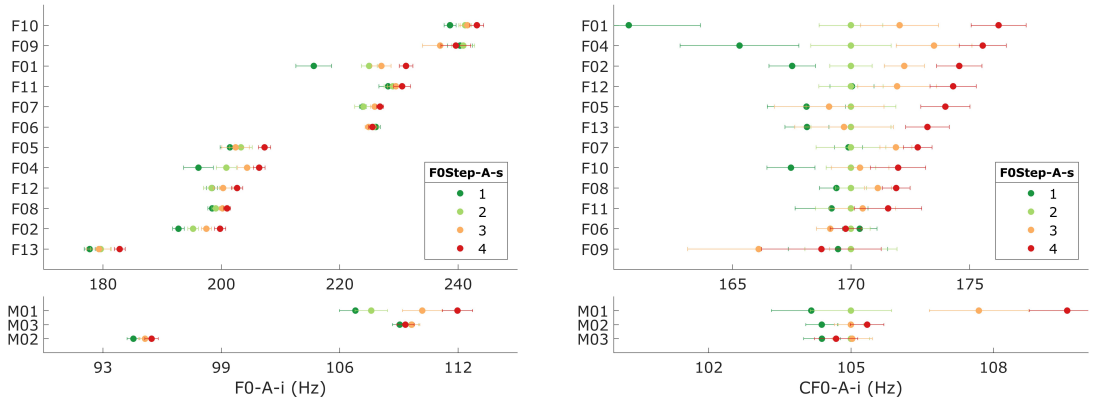


Figure 3.19: *Left*: kernel smoothing estimate (bandwidth = 5 Hz) of probability density of F0-A-i at each F0Step-A-s (F0-A-s) for all participants in Experiment 1B. *Right*: bar plots displaying mean F0-A-i at each F0Step-A-s (F0-A-s) for all participants in Experiment 1A. Female and male participants are shown in the top and bottom panel, respectively. Error bars of ±1 *s.e.* are also plotted. Each color represents one of the four F0Step-A-s.

F0-A-i In Experiment 1A, the bi-modal distributions of F0-A-i reflect the differences in F0-A-i between male and female participants. For female participants, the distributions of F0-A-i also exhibit two modes: one below 200 Hz and the other at around 230 Hz. The bi-modality in the distributions of F0-A-i also holds for male participants, albeit with a shorter distance between the two modes. For both female and male participants, the differences in the distribution of F0-A-i among the four F0Step-A-s, represented by four colors, are rather small. The pooled mean

F0-A-i ranges from 212 Hz to 217 Hz for female participants, and from 103.5 Hz to 106 Hz for male participants, in the ascending order of F0Step-A-s (F0-A-s). Given that the TP in the stimuli ranges from 165 Hz to 180 Hz for female, and from 102 Hz to 110 Hz for male, the increases in the mean F0-A-i are rather small. The pooled mean F0-A-i increases as F0-A-s increases. However, judging from the error bars, none all the increases in the mean F0-A-i are significant.



(a) F0-A-i (b) F0-A-i centered at F0Step-A-s 2

Figure 3.20: Scatter plots displaying mean F0-A-i at each F0Step-A-s (F0-A-s) for each participant in Experiment 1A (in the descending order of mean F0-A-i at F0Step-A-s 4, F0-A-s = 180 Hz for female and 110 Hz for male). Female and male participants are shown in the top and bottom panel, respectively. Each color represents one of the four F0Step-A-s. Error bars of ±1 *s.e.* are also plotted in the corresponding colors. (a) F0-A-i; (b) Centralized F0-A-i (centered at F0Step-A-s 2, F0-A-s = 170 Hz for female and 105 Hz for male).

The mean F0-A-i is further investigated at the individual level for both female and male participants. Figure 3.20(a) shows the mean F0-A-i at each F0Step-A-s for each participant; Figure 3.20(b) shows the same data centered at F0Step-A-s 2 (F0-A-s = 170 Hz for female, and 105 Hz for male). A great deal of variation in F0-A-i can be accounted for by taking into consideration the pre-determined F0 range of each participant. Each participant has a preferred F0-A-i value towards which the imitation of F0-A-s is biased. It holds that in general, as the F0Step-A-s

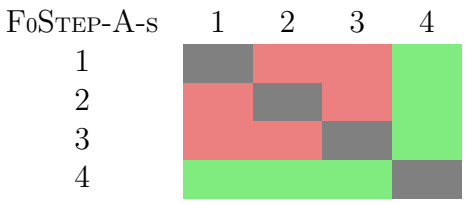
increases, the mean F0-A-i also increases, despite that the increases may not reach statistical significance. However, judging from the error bars, the mean F0-A-i at F0Step-A-s 4 (F0-A-s = 180 Hz for female and 110 Hz for male) is significantly different from the rest for most participants, female or male.

(A-f)	F0Step-A-s	1	2	3	4
	F0-A-s (Hz)	165	170	175	180
	F0-A-i: mean (Hz)	211.76	213.91	214.46	216.66
	F0-A-i: sem (Hz)	0.75	0.72	0.73	0.65
(B-f)	**F0Step-A-s (fixed)**	F(3, 3973) = 8.01, p = 2.55e-5***			
(C-f)	**F0Step-A-s (fixed)**	F(3, 3929) = 11.32, p = 2.86e-5***			
	Participant (random)	F(11, 3929) = 337.12, p =1.37e-30***			
	F0Step-A-s * Participant (random)	F(33, 3929) = 2.17, p = 1.30e-4***			
(A-m)	F0Step-A-s	1	2	3	4
	F0-A-s (Hz)	102	105	108	110
	F0-A-i: mean (Hz)	103.52	104.21	105.24	105.88
	F0-A-i: sem (Hz)	0.53	0.52	0.60	0.57
(B-m)	**F0Step-A-s (fixed)**	F(3, 942) = 3.60, p = 0.01**			
(C-m)	F0Step-A-s (fixed)	F(3, 934) = 1.55, p = 0.30			
	Participant (random)	F(2, 934) = 132.23, p = 1.09e-5***			
	F0Step-A-s * Participant (random)	F(6, 934) = 4.61, p = 1.25e-4***			

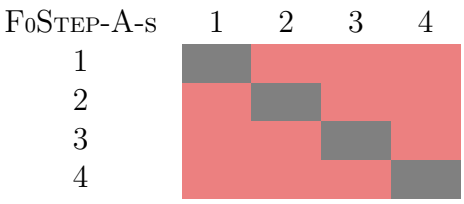
Table 3.15: (A) Mean and standard error of F0-A-i at four F0Step-A-s for female participants in Experiment 1A. (B) One-way ANOVA; fixed term: F0Step-A-s. (C) Two-way ANOVA; fixed term: F0Step-A-s, random terms: Participant, interaction between F0Step-A-s and Participant. "-f" denotes female (top panel); "-m" denotes male (bottom panel). Statistical significant terms are in bold.

Table 3.15 shows ANOVA results of F0-A-i (top panel: female; bottom panel: male). For both female and male participants, the effect of F0Step-A-s on F0-A-i is significant in one-way ANOVA (Table 3.15(B-f), $F(3, 3973) = 8.01, p < 0.001$ for female; Table 3.15(B-m) $F(3, 942) = 3.06, p = 0.01$ for male). However, after factoring in the random effects of Participant and the interaction between Participant and F0Step-A-s, the effect of F0Step-A-s is significant only for female participants but not for male participants (Table 3.15(C-f) and (C-m)). Meanwhile, the effects of both random terms are significant, which is in line with the speaker-specific variation in Figure 3.20.

Table 3.16 shows the multiple comparisons of F0-A-i between pairs of F0Step-A-s, respectively for female (Table 3.16(a)) and male (Table 3.16(b)), after the random effects are controlled for. For female participants, the differences in F0-A-i are only significant between F0Step-A-s 1 and 4, between 2 and 4, and between 3 and 4. For male participants, none of the pairwise comparisons reach statistical significance. These results are in line with ANOVA in Table 3.15(C).



(a) Female participants



(b) Male participants

Table 3.16: Comparisons of F0-A-i between pairs of F0Step-A-s female (a) and male (b) participants in Experiment 1A. Green indicates statistical significance; red indicates statistical non-significance.

F0-B-i In Experiment 1B, the distributions of F0-B-i are also bi-modal, reflecting the differences in F0-B-i between male and female participants. The distributions of F0-B-i are primarily unimodal: 250 Hz for female participants and 140 Hz for male participants. For both female and male participants, the differences in the distribution of F0-B-i among the four F0Step-A-s are even smaller among the four F0Step-B-s. The pooled mean F0-B-i ranges from 257 Hz to 260 Hz for female participants, and from 141.5 Hz to 144 Hz. The pooled mean F0-B-i increases as F0-B-s increases, despite that not all the increases reach statistical significance. This holds for both genders, but more so for male participants than for female participants.

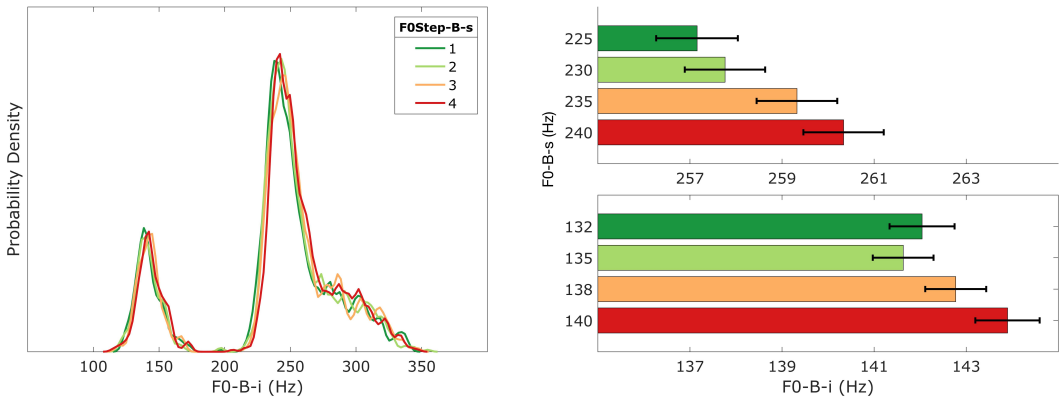
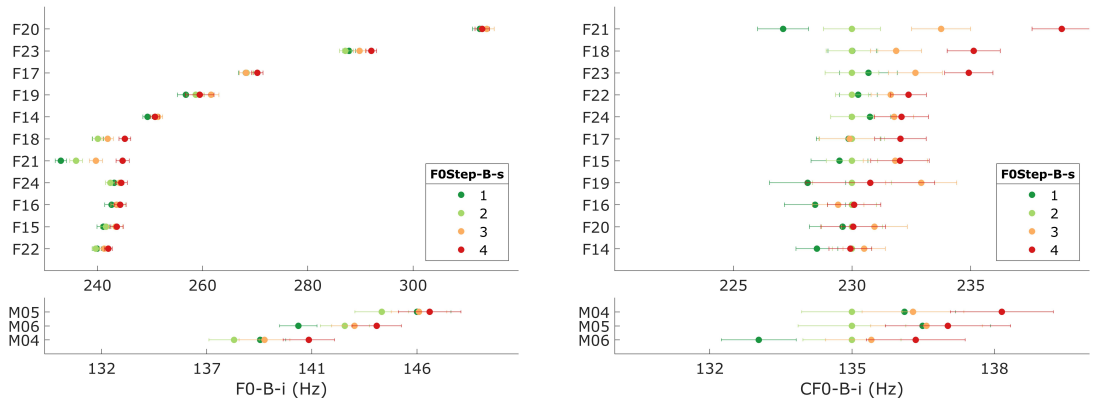


Figure 3.21: *Left*: kernel smoothing estimate (bandwidth = 2.5 Hz) of probability density of F0-B-i at each F0Step-B-s (F0-B-s) for all participants in Experiment 1B. *Right*: bar plots displaying mean F0-B-i at each F0Step-B-s (F0-B-s) for all participants in Experiment 1B. Female and male participants are shown in the top and bottom panel, respectively. Error bars of ± 1 *s.e.* are also plotted. Each color represents one of the four F0Step-B-s.

The mean F0-B-i is further investigated at the individual level for both female and male participants. Figure 3.22(a) shows the mean F0-B-i at each F0Step-B-s for each participant; Figure 3.22(b) shows the same data centered at F0Step-B-s 2 (F0-B-s = 230 Hz for female, and 135 Hz for male). Each participant has a preferred F0-B-i value towards which the imitation of F0-B-s is biased. Most participants have a fairly narrow range of F0-B-i. For these participant, the effect of F0Step-B-s on F0-B-i is little to none. For participants with a wider range of F0-B-i, F0-B-i increases significantly as F0Step-B-s increases (most notably are F21 and F18).



(a) F0-B-i (b) F0-B-i centered at F0Step-B-s 2

Figure 3.22: Scatter plots displaying mean F0-B-i at each F0Step-B-s (F0-B-s) for each participant in Experiment 1B (in the descending order of mean F0-B-i at F0Step-B-s 4, F0-B-s = 240 Hz for female and 134 Hz for male). Female and male participants are shown in the top and bottom panel, respectively. Each color represents one of the four F0Step-B-s. Error bars of ± 1 *s.e.* are also plotted in the corresponding colors. (a) F0-B-i; (b) Centralized F0-B-i (centered at F0Step-B-s 2, F0-B-s = 230 Hz for female and 135 Hz for male).

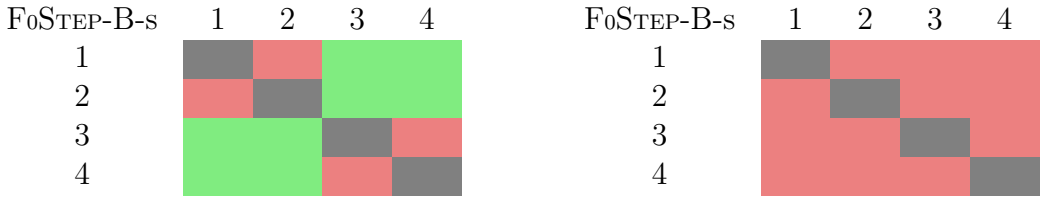
Table 3.17 shows ANOVA results of F0-B-i (top panel: female; bottom panel: male). In one-way ANOVA, the effect of F0Step-B-s on F0-B-i is only significant for female participants (Table 3.17(B-f), $F(3, 3666) = 2.74, p = 0.04$), but non-significant for male participants (Table 3.17(B-m), $F(3, 812) = 2.10, p = 0.10$). Even for female participants, the p-value is just below 0.05, which indicates that the effect size of F0Step-B-s is fairly small. After taking into consideration of the random effects of Participant and the interaction between Participant and F0Step-B-s, the effect of F0Step-B-s on F0-B-i is much more significant with a much lower p-value (Table 3.17(C-f), $F(3, 3626) = 10, 50, p < 0.001$). For male participants, the fixed effect of F0Step-B-s remains non-significant, albeit with a lower p-value (Table 3.17(C-m), $F(3, 804) = 4.00, p = 0.07$). The random effect of Participant is highly significant for female and male participants, whereas the interaction term is only significant for female participants.

(A-f)	F0Step-B-s	1	2	3	4
	F0-B-s (Hz)	225	230	235	240
	F0-B-i: mean (Hz)	257.15	257.76	259.33	260.34
	F0-B-i: sem (Hz)	0.89	0.87	0.88	0.87
(B-f)	**F0Step-B-s (fixed)**	F(3, 3666) = 2.74, p = 0.04*			
(C-f)	**F0Step-B-s (fixed)**	F(3, 3626) = 10.50, p = 6.28e-5***			
	Participant (random)	F(10, 3626) = 1164.87, p = 5.34e-6***			
	F0Step-B-s * Participant (random)	F(30, 3626) = 1.52, p = 0.03*			
(A-m)	F0Step-B-s	1	2	3	4
	F0-B-s (Hz)	132	135	138	140
	F0-B-i: mean (Hz)	142.04	141.63	142.77	143.9
	F0-B-i: sem (Hz)	0.71	0.66	0.66	0.70
(B-m)	F0Step-B-s (fixed)	F(3, 812) = 2.10, p = 0.10			
(C-m)	F0Step-B-s (fixed)	F(3, 804) = 4.00, p = 0.07			
	Participant (random)	F(2, 804) = 56.17, p = 1.29e-4***			
	F0Step-B-s * Participant (random)	F(6, 804) = 0.54, p = 0.77			

Table 3.17: (A) Mean and standard error of F0-B-i at four F0Step-B-s for female participants in Experiment 1B. (B) One-way ANOVA; fixed term: F0Step-B-s. (C) Two-way ANOVA; fixed term: F0Step-B-s, random terms: Participant, interaction between F0Step-B-s and Participant. "-f" denotes female (top panel); "-m" denotes male (bottom panel). Statistical significant terms are in bold.

Table 3.18 shows the multiple comparisons of F0-B-i between pairs of F0Step-

B-s, respectively for female (Table 3.18(a)) and male (Table 3.18(b)), after the random effects are controlled for. For female participants, the differences in F0-B-i are significant between any of the two F0Step-B-s except for between F0Step-B-s 1 and 2, and between 3 and 4. For male participants, none of the pairwise comparisons reaches statistical significance. These results are in line with ANOVA in Table 3.17(C).



(a) Female Participants (b) Male Participants

Table 3.18: Comparisons of F0-B-i between pairs of F0Step-B-s for female (a) and male (b) participants in Experiment 1B. Green indicates statistical significance; red indicates statistical non-significance.

To sum up the results regarding the imitation of TP F0 for both Experiment 1A and 1B: 1) the differences in F0-i across F0Step-s are fairly small to the extent that most of the pairwise comparisons do not reach statistical significance; 2) speaker-specific variation (random effects) can account for a large part of the variation in F0-i; 3) with the random effects being controlled for, the effect of F0Step-s on F0-i is significant only for female participants due in large part to the relatively high F0-i at F0Step-s 4.

3.5.3 Perception-Production Correlation

The correlation between discrimination and imitation is examined in Figure 3.23. There is a strong positive correlation between the discrimination performance and

the distance in the mean RelTiming-i between two TimingStep-s with the same F0Step-s. The correlation is high for both experiments: for Experiment 1A, $\rho = 0.75, p < 0.001$; for Experiment 1B, $\rho = 0.98, p < 0.001$. This suggests the better the participants differentiate two stimuli (with different TP relative timing but the same F0) in the discrimination tasks, the more distinct their imitations of those two stimuli would become.

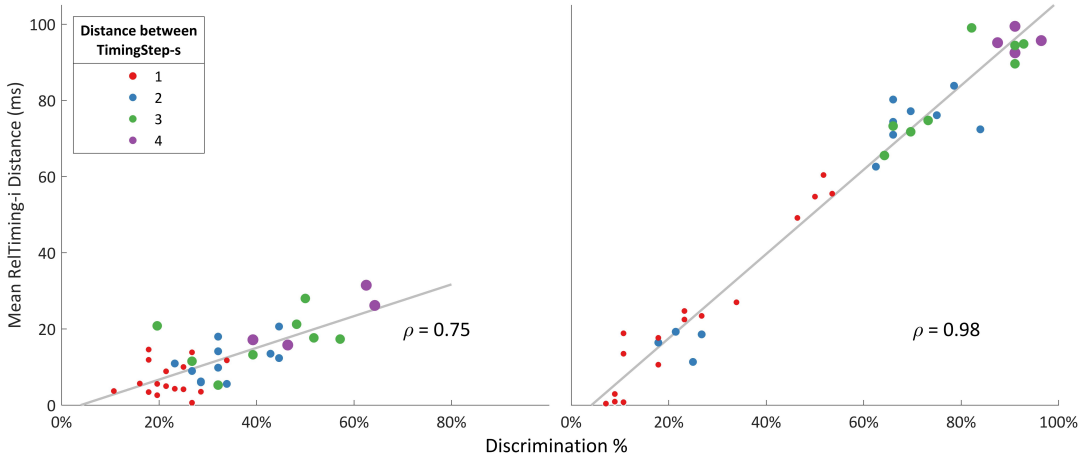


Figure 3.23: Scatter plots showing the correlation between the discrimination performance (in %) and the distance in the mean RelTiming-i between two TimingStep-s with the same F0Step-s for Experiment 1A (left) and Experiment 1B (right). The least squares regression lines are also plotted.

3.5.4 Bivariate Analysis of TP RelTiming and TP F0

While the previous analyses investigate the respective effects of RelTiming-s on RelTiming-i and F0-s on F0-i, it leaves unanswered that whether TP RelTiming-i is correlated with TP F0-i. If there exist some relationship between the two variables, the obvious questions are: 1) Do changes in RelTiming-s and F0-s have significant effects on the relationship between RelTiming-i and F0-i? 2) If so, what is the size and direction of the effects of RelTiming-s, F0-s and the interaction?

To address these question, we first look at the correlations between RelTiming-i and F0-i for each stimulus condition (a combination of TimingStep-s and F0-s). Figure 3.24 and 3.25 display the correlations between CRelTiming-A-i and CF0-A-i in Experiment 1A, and between CRelTiming-B-i and CF0-B-i in Experiment 1B, respectively. Each subplot represents a stimulus condition. The correlation coefficients are stored in Table 3.19 and 3.20.

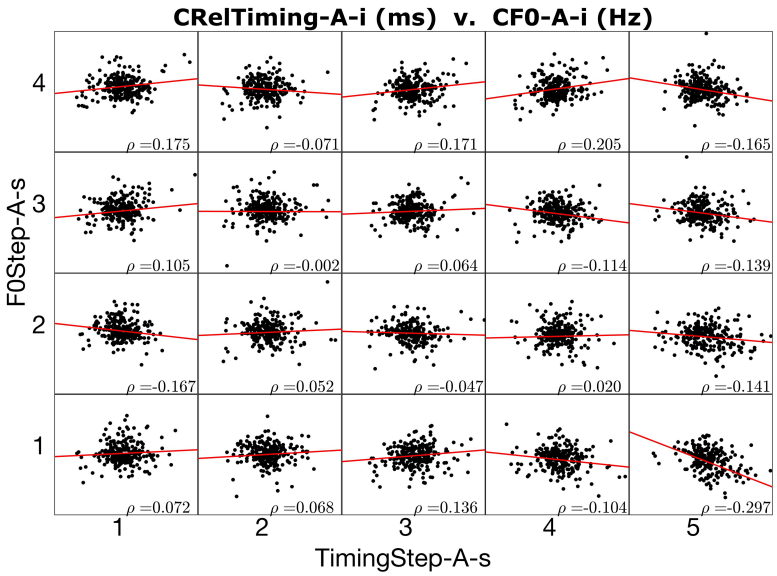


Figure 3.24: Scatter plots displaying correlation between CRelTiming-A-i and CF0-A-i in Experiment 1A. X axis: TimingStep-A-s; Y axis: F0Step-A-s. Each coordinate (subplot) represents one stimulus condition of TimingStep-A-s and F0Step-A-s. The least-squares regression lines are shown in red.

TimingStep-A-s	1	2	3	4	5
F0Step-A-s					
4	0.17**	-0.07	0.17**	0.20**	-0.17**
3	0.11·	0.00	0.06	-0.11·	-0.14*
2	-0.17**	0.05	-0.05	0.02	-0.14*
1	0.07	0.07	0.14*	-0.10	-0.30***

Table 3.19: Correlation coefficients between CRelTiming-A-i and CF0-A-i in Experiment 1A. Each cell represents one stimulus condition of TimingStep-A-s and F0Step-A-s. Statistical significance: · 0.1; * 0.05; ** 0.01; *** 0.001.

It appears that the relationship between CRelTiming-A-i and CF0-A-i varies from one stimulus condition to another without having a consistent trend. For the first four TimingStep-A-s, the correlation tends change sign between two adjacent F0Step-A-s. However, because the majority of these correlations are non-significant and the correlation coefficients are generally close to zero, the irregular changes in the correlation sign might arise from speaker variation. Most notable is the last column in Figure 3.24 and Table 3.19, which shows that the correlation between CRelTiming-A-i and CF0-A-i are consistently negative. More importantly, the correlation between the two dependent variables reaches a highly significant -0.30 as TP1 occurs the latest in the stimulus with the lowest F0. This suggests that at least for this particular stimulus condition, i.e., TimingStep-A-s 5 and F0Step-s 1, the earlier the turning point occurs in the imitation, the higher the F0, and vice versa.

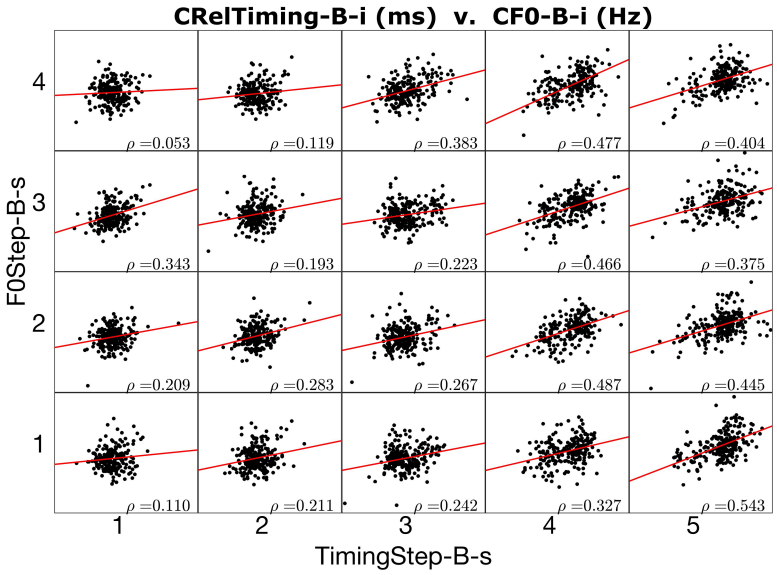


Figure 3.25: Scatter plots displaying correlation between CRelTiming-B-i and CF0-B-i in Experiment 1B. X axis: TimingStep-B-s; Y axis: F0Step-B-s. Each coordinate (subplot) represents one stimulus condition of TimingStep-B-s and F0Step-B-s. The least-squares regression lines are shown in red.

TimingStep-B-s	1	2	3	4	5
F0Step-B-s					
4	0.05	0.12·	0.38***	0.48***	0.40**
3	0.34***	0.19**	0.22***	0.47***	0.37*
2	0.21**	0.28***·	0.27***	0.49***	0.45***
1	0.11	0.21**	0.24***	0.33***	0.54***

Table 3.20: Correlation between CRelTiming-B-i and CF0-B-i in Experiment 1B. Each cell represents one stimulus condition of TimingStep-B-s and F0Step-B-s. Statistical significance: · 0.1; * 0.05; ** 0.01; *** 0.001.

The relationship between CRelTiming-B-i and CF0-B-i is more straightforward and more consistent across stimulus conditions. In general, the two variables have a positive correlation: as the turning point occurs later in the imitation, it reaches a higher F0 value. However, the relationship is complicated by the stimulus condition. Most notably, as TP2 occurs later in the stimulus, the positive correlation between CRelTiming-B-i and CF0-B-i becomes more prominent and exhibits a steeper regression line.

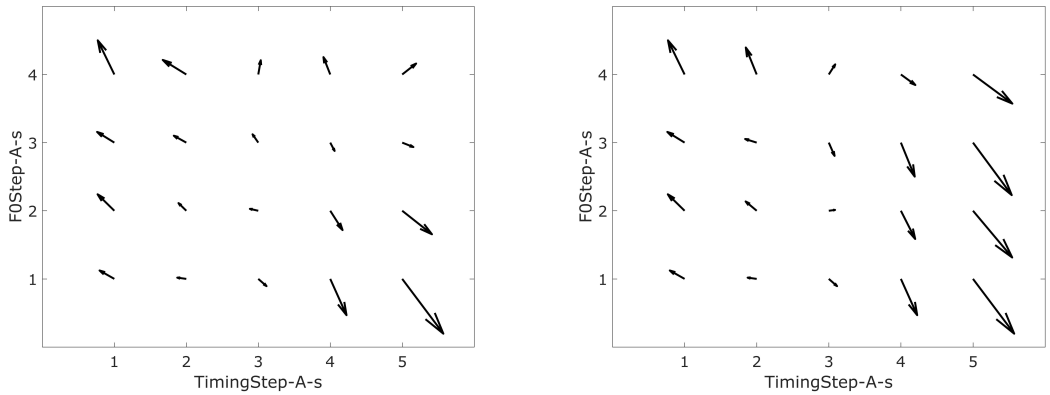
We conduct Multivariate Analysis of Variance (MANOVAs) to confirm these observations and to answer the two questions brought up in the first paragraph of the current section. MANOVA is generalized from ANOVA, and therefore is much like ANOVA in that it is also a procedure of comparing sample means by testing the statistical significance of the mean differences. Unlike ANOVA, where there is only one dependent variable, MANOVA is a multivariate procedure in that there are more than one dependent variables being compared simultaneously. In the current experiment, the two dependent variables are CRelTiming-i and CF0-i, and the two independent variables are TimingStep-s and F0Step-s. Note that CRelTiming-i and CF0-i are centralized values of RelTiming-i and F0-i. The re-centering does not affect the outcome of MANOVA. Moreover, as will be seen in the subsequent analysis, it provides a clearer picture of the effect size of the independent variables.

	F	dF1	dF2	p-value
F0Step-A-s	5.55	30	8710	0***
TimingStep-A-s	9.45	32	8767	0***
F0Step-A-s* TimingStep-A-s	2.40	24	8493	0.0015**

Table 3.21: Bivariate ANOVA of CF0-A-i and CRelTiming-A-i in Experiment 1A. The full model includes the main effects of F0Step-A-s and TimingStep-A-s and the interaction. Statistical significant terms are in bold.

Table 3.21 presents the MANOVA results in Experiment 1A. The model tests the effects of the three factors, i.e., F0Step-A-s, TimingStep-A-s, and their interaction, on the joint patterning of the two dependent variables CRelTiming-A-i and CF0-A-i. The effects of all three factors have a significant effect on the relationship between the two response variables. For F0Step-A-s, $F(30, 8710) = 5.55$, $p < 0.001$; for TimingStep-A-s, $F(32, 8767) = 9.45$, $p < 0.001$; for the interaction between F0Step-A-s and TimingStep-A-s, $F(24, 8493) = 2.40$, $p < 0.01$.

The coefficients of the main and interaction effects are reconstructed and plotted in Figure 3.26(a). Each pointing arrow has two coordinates: the X coordinate represents the reconstructed effects of the stimulus condition, including both the main effects and the interaction, on CRelTiming-A-i; the Y coordinate represents the reconstructed effects of the stimulus condition on CF0-A-i. A positive X coordinate value indicates that RelTiming-A-i is higher than the median value, meaning TP1 occurs relatively late in the stimulus; a negative X coordinate value therefore indicates that TP1 occurs relatively early in the stimulus. Similarly, a positive Y coordinate value indicates that F0-A-i is higher than the median value, meaning TP1 has a relatively high F0 minimum; a negative Y coordinate value therefore indicates that TP1 has a relatively low F0 minimum. This is further illustrated in Figure 3.27.



(a) Full Model (b) Main Effects

Figure 3.26: Quiver plots displaying the effects of F0Step-A-s and TimingStep-A-s on the two dependent variables, i.e., CF0-A-i and CRelTiming-A-i. (a): Full model including interaction effects; (b): Only main effects. For each pointing arrow, the X coordinate denotes the effects on CRelTiming-A-i, and the Y coordinate denotes the effects on CF0-A-i.

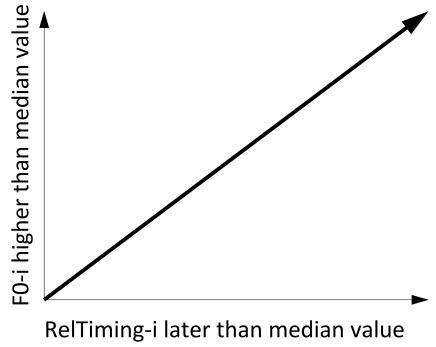


Figure 3.27: Illustration of the point arrow in the quiver plots.

The coefficients of the main and interaction effects are reconstructed and plotted in Figure 3.26(a). Each pointing arrow has two coordinates: the X coordinate represents the reconstructed effects of the stimulus condition, including both the main effects and the interaction, on CRelTiming-A-i; the Y coordinate represents the reconstructed effects of the stimulus condition on CF0-A-i. A positive X coordinate value indicates that RelTiming-A-i is higher than the median value, meaning

TP1 occurs relatively late in the stimulus; a negative X coordinate value therefore indicates that TP1 occurs relatively early in the stimulus. Similarly, a positive Y coordinate value indicates that F0-A-i is higher than the median value, meaning TP1 has a relatively high F0 minimum; a negative Y coordinate value therefore indicates that TP1 has a relatively low F0 minimum. This is further illustrated in Figure

It appears that the effects of the stimulus condition on CRelTiming-A-i and CF0-A-i are the most prominent when TP1 occurs late in the first [ma] with a low F0 minimum, i.e., the lower right corner. On average, in response to a stimulus at TimingStep-A-s 5 and F0Step-A-s 1, the imitation TP1 occurs 24 ms later than the median RelTiming-A-i with the F0 minimum 7 Hz lower than the median F0-A-i. With smaller effect sizes, the effects of the stimulus condition on the upper left corner are in a perfect opposite direction. The most notable case would be that in response to a stimulus at TimingStep-A-s 1 and F0Step-A-s 4, the imitation TP1 occurs about 10 ms earlier than the median RelTiming-A-i with the F0 minimum 4 Hz higher than the median F0-A-i.

Figure 3.26(b) and Table 3.22, display only the main effects of TimingStep-A-s and F0Step-A-s on CF0-A-i and CRelTiming-A-i. The main effect of F0Step-A-s is modest: CF0-A-i increases by an amount of 3 Hz from F0Step-A-s 1 to 4, whereas the changes in CRelTiming-A-i are little to none (about -1 ms) across F0Step-A-s. On the other hand, the main effect of TimingStep-A-s has a larger size: CF0-A-i decreases by as much as 8 Hz, and CRelTiming-A-i increases by 32 ms from TimingStep-A-s 1 to 5.

Comparing Figure 3.26(b) to Figure 3.26(a), it becomes apparent that the interaction between TimingStep-A-s and F0Step-A-s has a significant effect on the

F0Step-A-s	TimingStep-A-s	CF0-A-i (Hz)	CRelTiming-A-i (ms)
1	1	1.0	-8.5
2	0	1.0	-1.1
3	0	0.3	-1.4
4	0	3.1	-1.1
0	2	-0.9	3.3
0	3	-1.9	13.2
0	4	-5.4	17.9
0	5	-7.7	32.2

Table 3.22: Main effects of F0Step-A-s and TimingStep-A-s on CF0-A-i and CRelTiming-A-i in Experiment 1A.

relationship between CRelTiming-A-i and CF0-A-i. Specifically, the interaction effect of the stimulus conditions on the upper right corner cancels out the main effects. As a result, in the full model, only for the stimuli on the lower right corner, the aggregate effects on the joint distribution of the two dependent variables are prominent.

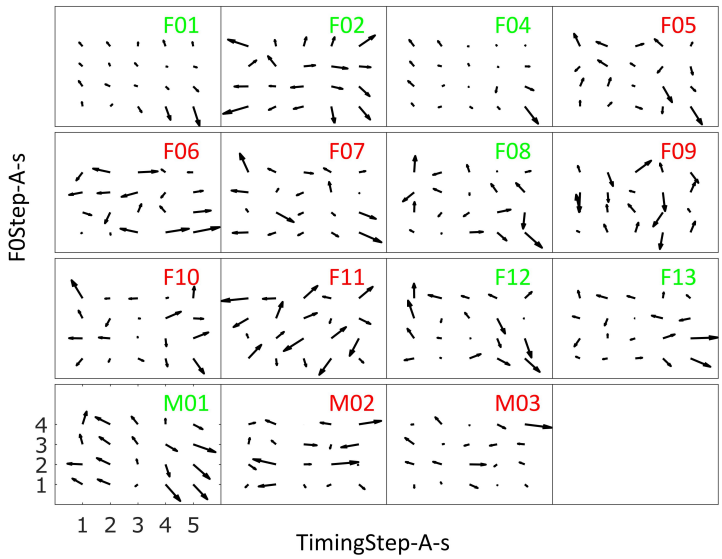


Figure 3.28: Quiver plots displaying the effects of F0Step-A-s and TimingStep-A-s on CF0-A-i and CRelTiming-A-i in the full model for each participant in Experiment 1A. Participants are identified on the upper right corner: red indicates neither the main nor interaction effects are significant; green indicates at least one is significant.

A closer investigation into the individual participant behavior reveals that out of the 15 participants in Experiment 1A, the majority of them—eight—show neither main effects nor interaction effects, as indicated by red on the upper right corner of each subplot in Figure 3.28. This is in line with the observation that the correlation coefficients between CRelTiming-A-i and CF0-A-i do not exhibit a consistent trend across stimulus conditions. For the other seven participants, at least the main effects are significant (the interaction effect is significant for five participants), as indicated by green on the upper right corner of each subplot. Despite the speaker-induced variation, there is a consistent pattern across these seven participants reminiscent of all participants: the arrows on the lower right corner point to the same general direction. More notably, the quiver plots of Participant F01 and F04 almost mirror that of all participants in Figure 3.26(a).

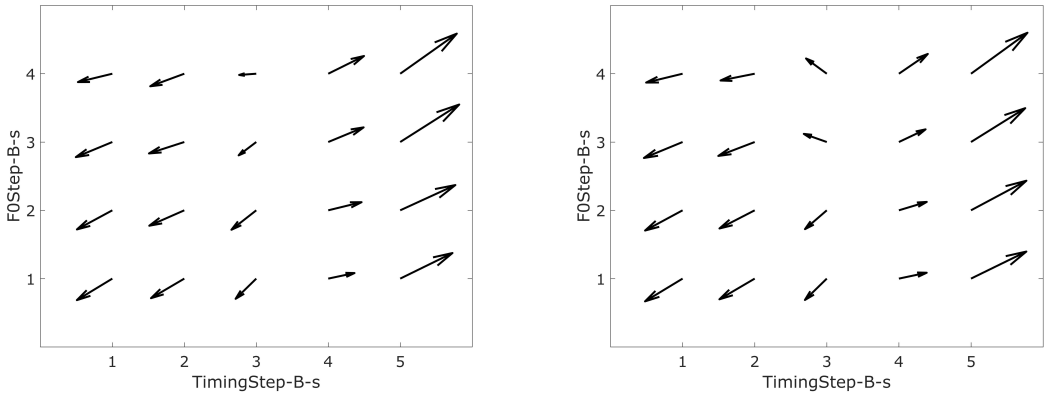
Table 3.23 shows the MANOVA results in Experiment 1B. The main effects of F0Step-B-s and TimingStep-B-s are both significant on the joint patterning of the two dependent variables CRelTiming-B-i and CF0-B-i. For F0Step-B-s, $F(30, 7933) = 2.82$, $p < 0.001$; for TimingStep-B-s, $F(32, 7985) = 123.53$, $p < 0.001$. However, the effect of the interaction between F0Step-B-s and TimingStep-B-s does not reach statistical significance; $F(24, 7735) = 0.91$, $p = 0.59$.

	F	dF1	dF2	p-value
F0Step-B-s	2.82	30	7933	4.56e-07***
TimingStep-B-s	123.53	32	7985	0***
F0Step-B-s* TimingStep-B-s	0.91	24	7735	0.59

Table 3.23: Bivariate ANOVA of CF0-B-i and CRelTiming-B-i in Experiment 1B. The full model includes the main effects of F0Step-B-s and TimingStep-B-s and the interaction. Statistical significant terms are in bold.

The reconstructed coefficients of the full model are plotted in Figure 3.29(a).

When TP2 occurs early in the second [ma] in the stimulus, both CRelTiming-B-i and CF0-B-i are negative, indicating that TP2 in the imitations occurs earlier than the median RelTiming-B-i with the F0 maximum lower than the median F0-B-i. For the early TP2 stimulus conditions (the first three), the higher F0Step-B-s, the smaller the aggregate effects on CF0-B-i, whereas the effect size on CRelTiming-B-i does not change across F0Step-B-s. When TP2 occurs late in the second [ma] in the stimulus, both CRelTiming-B-i and CF0-B-i are positive, indicating that TP2 in the imitations occurs later than the median RelTiming-B-i with the F0 maximum higher than the median F0-B-i. For these late TP2 stimulus conditions (the last two), the higher F0Step-B-s, the larger the aggregate effects on CF0-B-i, whereas the effect size on CRelTiming-B-i does not vary with F0Step-B-s.



(a) Full Model (b) Main Effects

Figure 3.29: Quiver plot displaying the effects of F0Step-B-s and TimingStep-B-s on the two dependent variables, i.e., CF0-B-i and CRelTiming-B-i. (a): Full model including interaction effects; (b): Only main effects. For each pointing arrow, the X coordinate denotes the effects on CRelTiming-B-i, and the Y coordinate denotes the effects on CF0-B-i.

Figure 3.29(b) and Table 3.24, display only the main effects of TimingStep-B-s and F0Step-B-s on CF0-B-i and CRelTiming-B-i. The main effect of F0Step-B-s is small: CF0-B-i increases by 3 Hz, while CRelTiming-B-i does not change much

F0Step-B-s	TimingStep-B-s	CF0-B-i (Hz)	CRelTiming-B-i (ms)
1	1	-4.8	-37.2
2	0	0.5	0.1
3	0	1.4	-1.1
4	0	2.9	1.1
0	2	0.4	2.3
0	3	0.3	15.5
0	4	6.0	64.8
0	5	10.4	92.1

Table 3.24: Main effects of F0Step-B-s and TimingStep-B-s on CF0-B-i and CRelTiming-B-i in Experiment 1B.

across F0Step-B-s. The main effect of TimingStep-B-s is relatively large: CF0-B-i rises by as much as 10 Hz, and CRelTiming-B-i increases by 92 ms from from TimingStep-B-s 1 to 5.

The comparison between the full model in Figure 3.29(a) and the main effects model in Figure 3.29(b) confirms that the interaction between F0Step-B-s and TimingStep-B-s does not have a significant effect on the joint patterning of CRelTiming-B-i and CF0-B-i.

In Experiment 1B, there is less speaker-specific variation in the patterning of CRelTiming-i and CF0-i than in Experiment 1A, as shown in Figure 3.30. Out of 14 participants, only two participants show neither main effects nor interaction effects. The other 12 participants show at least the main effect of TimingStep-B-s if not both TimingStep-B-s and F0Step-B-s. The main effect of TimingStep-B-s on the joint patterning of CRelTiming-B-i and CF0-B-i is consistent across these participants, which is in line with the relatively large F-statistics in the MANOVA results ($F(32, 7985) = 123.53$, $p < 0.001$). Moreover, the quiver plots of these 12 participants resemble that of all paticipants in Figure 3.29(a).

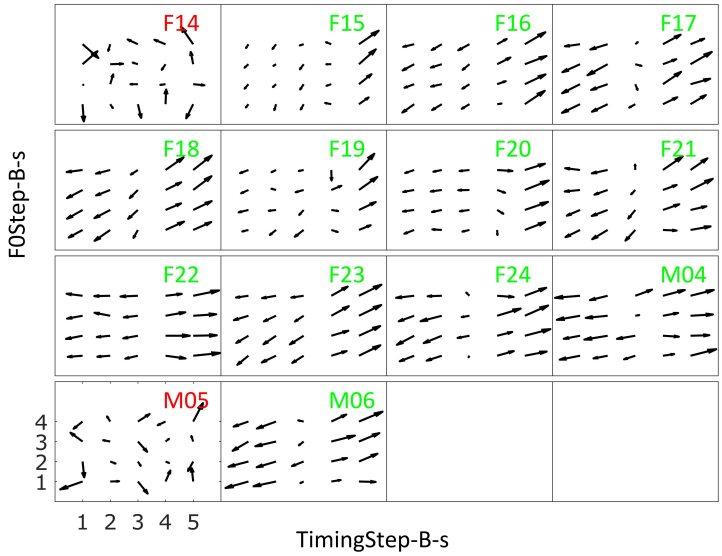


Figure 3.30: Quiver plots displaying the effects of F0Step-B-s and TimingStep-B-s on CF0-B-i and CRelTiming-B-i in the full model for each participant in Experiment 1B. Participants are identified on the upper right corner: red indicates neither the main nor interaction effects are significant; green indicates at least one is significant.

To sum up the findings of the bivariate MANOVA: 1) in Experiment 1A, the changes in both TimingStep-A-s and F0Step-A-s, as well as their interaction, have significant effects on the joint patterning of CRelTiming-A-i and CF0-A-i; 2) in Experiment 1B, only the main effects of TimingStep-B-s and F0Step-B-s, but not their interaction, are significant on the joint patterning of CRelTiming-B-i and CF0-B-i; 3) the effect size of F0Step-s is smaller than that of TimingStep-s in both Experiment 1A and 1B; 4) for the main effect of F0Step-s, the F0Step-s-induced changes in CF0-i are in the same direction (positive) for both experiments, whereas the F0Step-s-induced changes in CRelTiming-i are negligible; 5) for the main effect of TimingStep-s, while the TimingStep-s-induced changes in CRelTiming-i are in the same direction (positive) for both experiments, the TimingStep-s-induced changes in CF0-i are in the opposite directions for the two experiments—negative

in Experiment 1A but positive in Experiment 1B; 6) the interaction effect is only significant in Experiment 1A, leaving the aggregate effects of TimingStep-A-s and F0Step-A-s on the joint distribution of CRelTiming-A-i and CF0-A-i only prominent when TP1 occurs relatively late in the stimulus with a relatively low F0 minimum; 7) there is more speaker-specific variation in the MANOVA results in Experiment 1A than in Experiment 1B.

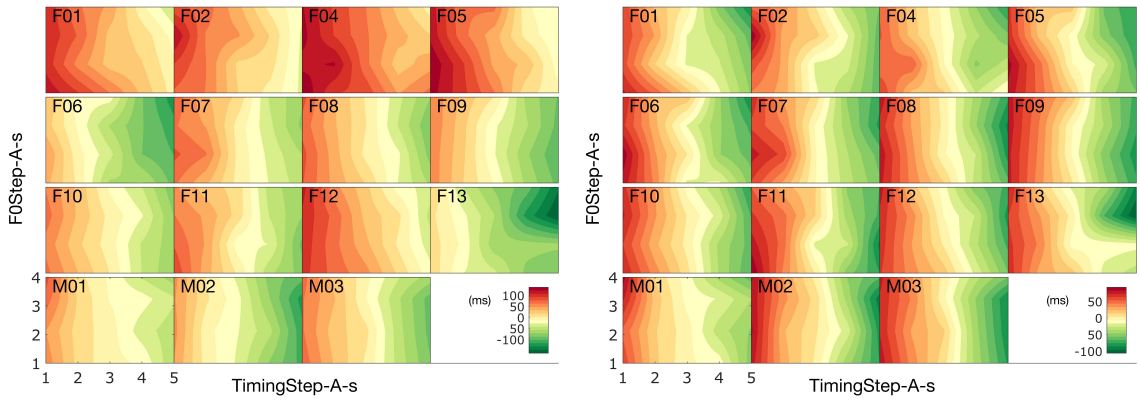
3.5.5 Speaker Adaptation

Individual participants exhibit a great amount of variation in their imitations of TP relative timing and F0. In this section, how individual participants adapt to the stimuli in the imitations is investigated by conducting analyses on the distance in TP relative timing and F0 between the stimuli and the imitations.

The RelTimingDist and CRelTimingDist for each participant are plotted in two separate heatmaps. In the heatmaps, the mean RelTimingDist values are organized along two dimensions: TimingStep-s and F0Step-s. Red indicates that the imitation TP is delayed compared to the stimulus TP; green indicates the imitation TP is initiated earlier than the stimulus TP. The closer the imitation is to the stimulus in terms of the TP relative timing, the lighter the color in the heatmaps. Note that the difference between RelTimingDist heatmap and CRelTimingDist heatmap is that the former compares the latency in raw TP relative timing between the imitations and the stimuli, whereas the latter compares the latency in centralized TP relative timing, i.e., the distance after re-centering both stimulus and imitation data.

Figure 3.31(a) shows that in Experiment 1A, the behaviors of timing imitation

vary across participants with a consistent pattern: as the stimulus TP progresses in the tone-bearing syllable, the stimulus-imitation latency decreases, thus early stimulus TP timing induces greater stimulus-imitation latency than late stimulus TP timing. Note that here, "greater" refers to the mathematical value of a number, but not the absolute value. Thus, a positive latency is greater than zero, which is in turn greater than a negative latency. The general pattern is indicated by the warmer color on the left end of the heatmaps and the colder color on the right end of the heatmaps.



(a) RelTimingDist-A: Experiment 1A (b) CRelTimingDist-A: Experiment 1A

Figure 3.31: Smoothed heatmaps displaying RelTimingDist-A (a) and CRelTimingDist-A (b) for each participant in Experiment 1A. X axis: TimingStep-s; Y axis: F0Step-s.

On the other hand, Figure 3.31(b) shows all participants in Experiment 1A exhibit one pattern after RelTimingDist is re-centered by speaker: early stimulus TP timing induces positive stimulus-imitation latency, whereas late stimulus TP timing induces negative latency. When the stimulus TP occurs early in the tone-bearing syllable, the imitation TP is delayed compared to stimulus TP, as indicated by red in the heatmap. As the stimulus TP progresses, the imitation TP is initiated closer in time to the stimulus TP, as indicated by light yellow in the heatmap.

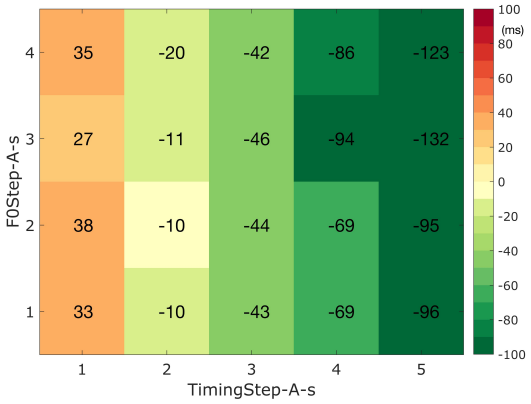
When the stimulus TP occurs late in the tone-bearing syllable, the imitation TP is initiated earlier in time than the imitation TP.

Three groups of imitation patterns can be identified from the RelTimingDist heatmaps in Figure 3.31(a): early boundary, mid boundary, and late boundary. This is further illustrated in Figure 3.32. Heatmaps and quiver plots of RelTimingDist (averaged across participants within the group) are plotted for each group. In the quiver plots, a right and left arrow indicate that RelTiming-i is larger and smaller than RelTiming-s, respectively; the length of the arrow is indicative of the absolute value of RelTimingDist.

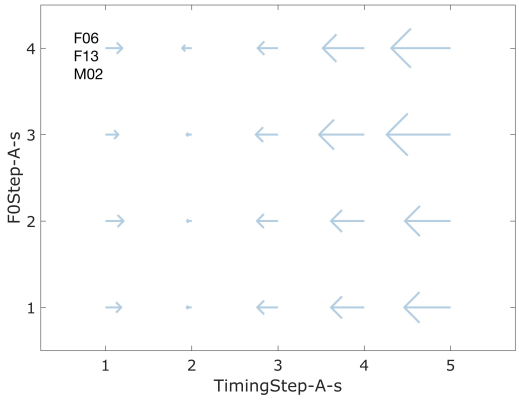
An imitation boundary is defined as a point on the stimulus continuum around which the stimulus TP is most faithfully imitated in terms of relative timing, or F_0, or both. Thus, it corresponds to the light yellow area in the RelTimingDist heatmap, and the area where the lengths of the arrows are the shortest in the quiver plot.

In Experiment 1A, the early boundary group (including F06, F13, and M02) has a RelTiming imitation boundary at TimingStep-A-s 2, which corresponds to 120 ms into the first [ma2]. The stimulus-imitation latency increases in the negative direction as the TP progresses in the tone-bearing syllable. The mid boundary group (including F07, F08, F09, F10, F11, M01, and M03) has an RelTiming imitation boundary at TimingStep-A-s 3, which corresponds to 160 ms into the first [ma2]. The stimulus-imitation latency increases in both directions as the TP occurs more distant from the middle TP stimulus timing: the latency is positive to the left of the RelTiming imitation boundary and negative to the right of the boundary. The late boundary group (including F01, F02, F04, F04, F05, and F12) is a mirror image of the early boundary group with an RelTiming imitation

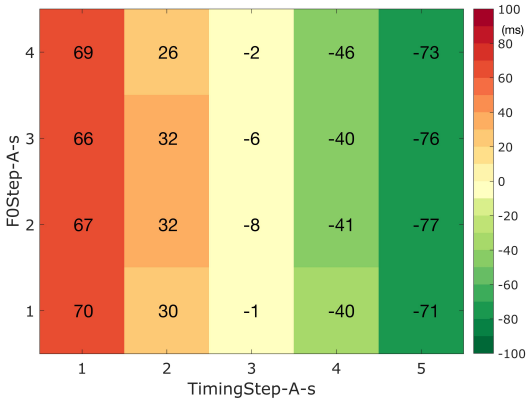
boundary somewhere between TimingStep-A-s 4 and 5. The farther left the TP is from the boundary, the stimulus-imitation latency grows larger in the positive direction.



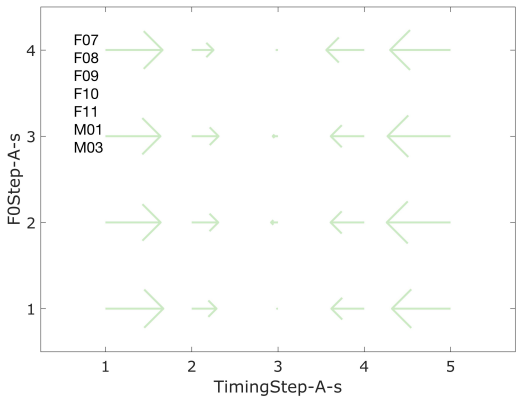
(a) Early boundary: heatmap



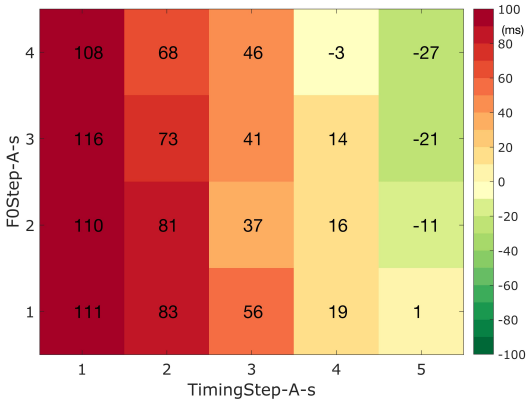
(b) Early boundary: quiver plot



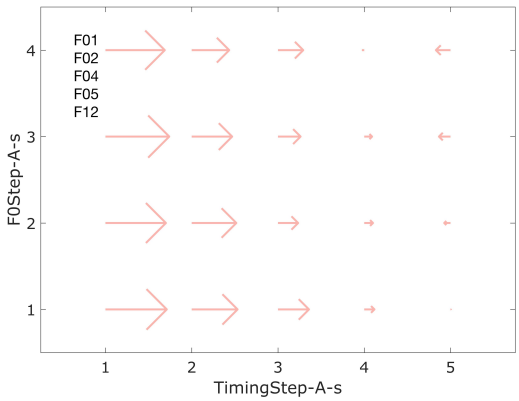
(c) Mid boundary: heatmap



(d) Mid boundary: quiver plot



(e) Late boundary: Heatmap



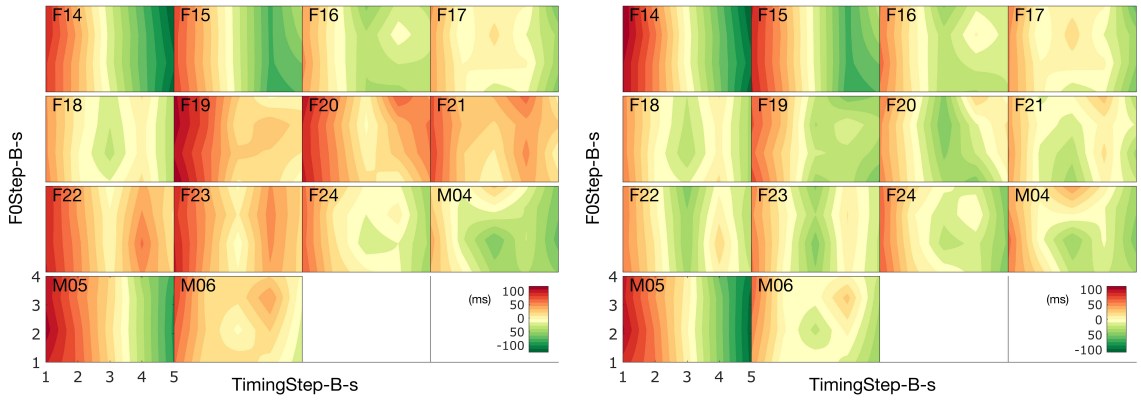
(f) Late boundary: quiver plot

Figure 3.32: Three groups of RelTiming imitation patterns in Experiment 1A

Note that all participants in Experiment 1A have a similar CRelTiming imitation boundary, approximately corresponding to TimingStep-A-s 3.

In Experiment 1B, the participant behaviors of timing imitation vary with more noise than in Experiment 1A, as shown in Figure 3.33(a). Three participants—F14, F15, and M05—exhibit imitation patterns similar to Experiment 1A with early stimulus TP inducing greater stimulus-imitation latency. Specifically, the three participants can be regarded as falling into the mid boundary group in Experiment 1A. The other participants differ significantly in terms of the overall impression of the heatmap in shades: for F16 and M04, the heatmap color is close to green; for

F17, F18, and F24, the heatmap color is close to light yellow; for F19, F20, F21, F22, F23, and M06, the heatmap color is close to red.



(a) RelTimingDist-B: Experiment 1B (b) CRelTimingDist-B: Experiment 1B

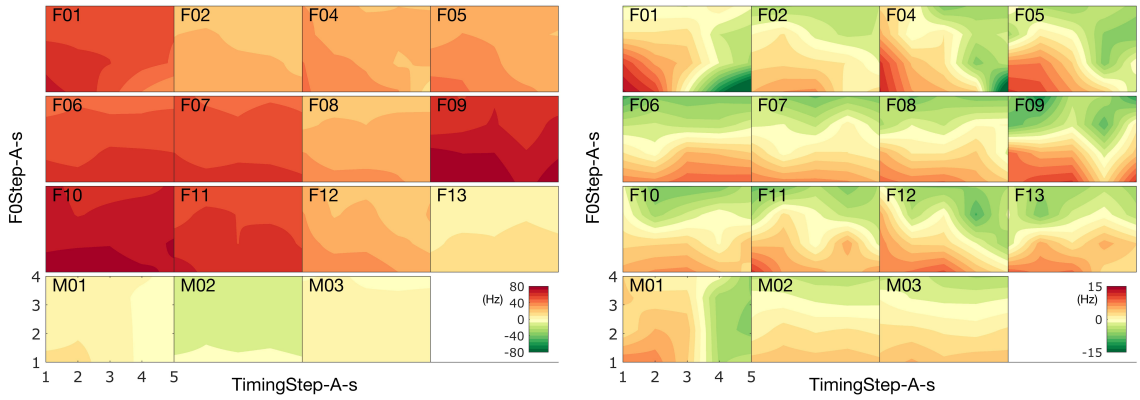
Figure 3.33: Smoothed heatmaps displaying RelTimingDist-B (a) and CRelTimingDist-B (b) for each participant in Experiment 1B. X axis: TimingStep-s; Y axis: F0Step-s.

After re-centering the data, all participants barring F14, F15, and M05, appear to fall under one category with two CRelTiming imitation boundaries: one occurs approximately at TimingStep-A-s 2 and the other at TimingStep-A-s 4. Similarly, three groups of imitation patterns—early, mid, and late—can also be identified within these participants based on the overall RelTimingDist heatmaps. The early, mid, and late timing group corresponds to the green, light yellow, and red overall color in the heatmap, respectively. The early timing group includes F16 and M04; the mid timing group includes F17, F18, and F24; the late timing group includes the rest.

Turning to F_0, the F0Dist and CF0Dist for each participant are plotted in two separate heatmaps in which the mean F0Dist values are organized along two dimensions: TimingStep-s and F0Step-s. Red indicates that the imitation TP is higher than the stimulus TP in the stimulus; green indicates lower. The closer the

imitation is to the stimulus in terms of the TP F0, the lighter the color in the heatmaps.

Figure 3.35(a) shows that most participants in Experiment 1A produce much higher TP F0 in the imitations than in the stimuli. Moreover, there is little variation in the imitation TP F0 within each participant. This is due to the relative low ratio of the within-speaker variation to the range of the imitation TP F0 across all participants.

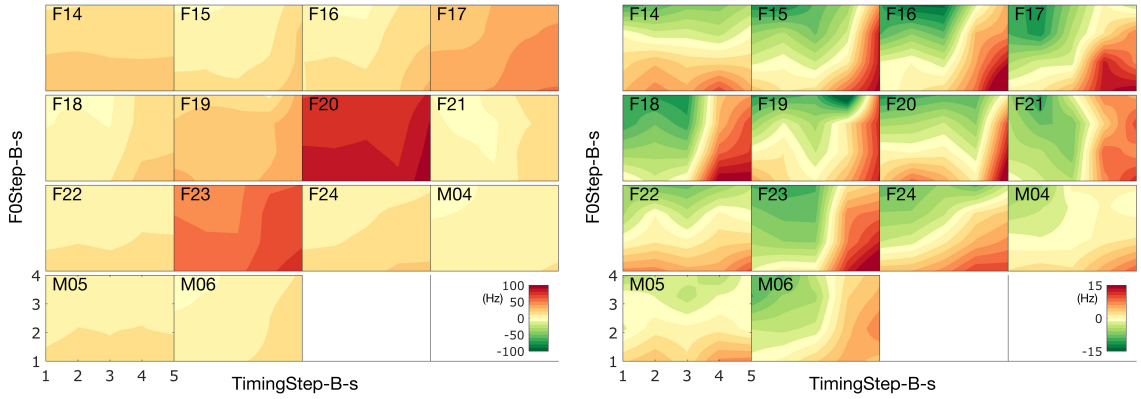


(a) F0Dist-A: Experiment 1A (b) CF0Dist-A: Experiment 1A

Figure 3.34: Smoothed heatmaps displaying F0Dist-A (a) and CF0Dist-A (b) for each participant in Experiment 1A. X axis: TimingStep-s; Y axis: F0Step-s.

The re-centered data show a much clearer picture of stimulus-induced variation in F0Dist across stimulus conditions (Figure 3.35(b)). in general, low stimulus TP CF0 induces greater stimulus-imitation distance. "Greater" here is used in a way similar to previously in that a positive distance is greater than zero, which is in turn greater than a negative distance. To spell it out, when the stimulus TP CF0 is high, the participants produce lower imitation TP CF0 , thus negative stimulus-imitation distance; when the stimulus TP CF0 is low, the participants produce higher imitation TP CF0, thus positive stimulus-imitation distance. Therefore, an CF0 imitation boundary can be identified between F0Step-2 and 3.

However, at least five participants—F01, F04, F05, F12, and M01—exhibit a more complex CF0 imitation pattern. For these participants, the aforementioned pattern still holds that low stimulus TP CF0 induces greater stimulus-imitation distance. More importantly, the stimulus TP timing also influences CF0Dist-A: as the TP progresses in the tone-bearing syllable, the stimulus-imitation distance (CF0Dist-A) grows smaller. That is, when the stimulus TP occurs early, the imitation TP CF0 is higher than the stimulus TP, whereas when the stimulus TP occurs late, the imitation TP CF0 is lower than the stimulus TP. As a result, the CF0 imitation boundary is, instead of horizontal, either vertical (F01 and M01), approximately corresponding to TimingStep-A-s 3, or diagonal from upper left to lower right (F04 and F05, and F12).



(a) F0Dist-B: Experiment 1B (b) CF0Dist-B: Experiment 1B

Figure 3.35: Smoothed heatmaps displaying F0Dist-B (a) and CF0Dist -B (b) for each participant in Experiment 1B. X axis: TimingStep-s; Y axis: F0Step-s.

Similar to Experiment 1A, participants in Experiment 1B produce higher TP F0 in the imitations with little variation. However, the overall impression of the RelTimingDist-B heatmaps in shades is closer to light yellow than the RelTimingDist-A heatmaps (the two heatmaps have similar F0 ranges). This indicates participants' imitation of F0-B-s are closer than that of F0-A-s.

The stimulus-induced variation in F0Dist-B is better illustrated in Figure 3.35(b). The same general pattern holds that high stimulus TP CF0 induces greater stimulus-imitation distance. Consequently, the lower part of the CF0Dist-B heatmaps is warmer in color than the upper part. At least two participants, F14 and M05, exhibit the CF0 imitation pattern solely influenced by the TP stimulus CF0. Therefore, the CF0 imitation boundary can be identified between F0Step-2 and 3.

More participants in Experiment 1B show the complex CF0 imitation pattern. For these participants, in addition to the variation induced by stimulus TP CF0, the variation in CF0Dist-B arises due in large part to changes in stimulus TP timing. As the TP progresses in the tone-bearing syllable, the stimulus-imitation distance (CF0Dist-B) grows larger: when the stimulus TP occurs early, the imitation TP CF0 is lower than the stimulus TP, whereas when the stimulus TP occurs late, the imitation TP CF0 is higher than the stimulus TP. Note that this is the mirror image of the less dominant CF0 imitation pattern in Experiment 1A. As a result, the CF0 imitation boundary is either vertical (F14, F21, and F23), or diagonal from upper right to lower left (F17 and F24), or a combination of the two (F15, 16, 19, 20, M04, and M06).

To sum up, in both Experiment 1A and 1B, the participants imitate the TP relative timing and F0 by making stimulus-induced adjustments to the preferred values. The speaker adaptation is most readily observed in the imitation behaviors of RelTiming-A-s in Experiment 1A: three distinct groups—early, mid, and late—fall under one category after re-centering. In Experiment 1B, the imitation behaviors of RelTiming-B-s fall into two categories. Within the more dominant category, three groups corresponding to early, mid, and late timing, are also iden-

tified. With regard to the F0 imitation, despite having higher overall f_0 than the stimuli, the participants also adapt to the stimulus TP F0, as evidenced by the fact that most participants converge to one CF0 imitation pattern.

3.6 Discussion

The main findings of this study can be summarized as:

1. The discrimination performance is low and there is no clear-cut discrimination boundary on the stimulus timing continuum in Experiment 1A, whereas the participants exhibit high discrimination performance in Experiment 1B, in which there is a discrimination boundary between TimingStep-B-s 3 and 4.
2. Most imitations in Experiment 1A belong to one distribution. The imitations in Experiment 1B belong to two distinct distributions. Specifically, the imitations for TimingStep-B-s 1-3 are associated with one distribution and those for TimingStep-B-s 4-5 the other.
3. In both experiments, the effect of TimingStep-s on RelTiming-i is significant: as TP progresses in the stimulus, RelTiming-i increases; the effect magnitude is lower in Experiment 1A than in Experiment 1B.
4. In both experiments, the effect of F0Step-s on F0-i is only significant for female participants.
5. For two stimuli with different TP relative timing but the same TP F0, the discrimination performance is positively correlated with the difference in the imitations.

6. In both experiments, the effects of TimingStep-s and F0Step-s on the joint patterning of CRelTiming-i and CF0-i are significant: while the TimingStep-s-induced changes in CRelTiming-i are both for both experiments, the TimingStep-s-induced changes in CF0-i are negative for Experiment 1A and positive for Experiment 1B; the F0Step-s-induced changes in CF0-i are in the same direction, and the F0Step-s-induced changes in CRelTiming-i are negligible.
7. There is a great deal of speaker-specific variation in the imitations: participants have their preferred values of relative timing and F0, which their imitations are adjusted to.

The findings show that speakers encode stimulus variation in a non-linear fashion, and that the parameter space is structured by perceptual categories. Therefore, it is argued that the coordination between the tone gestures and the oral articulatory gestures is categorical. Specifically, in Experiment 1A, most imitations exhibit one mode of gestural coordination. Naturally, this mode of gestural coordination corresponds to that of Tone2-bearing syllables in Mandarin. This captures the finding that in Experiment 1A, there is no clear-cut discrimination boundary on the stimulus timing continuum, and the imitations can be grouped into one distribution. However, in Experiment 1B, there are two modes of gestural coordination, which correspond to Tone2- and Tone4-bearing syllables, respectively. This can capture the finding that in Experiment 1B, there is a discrimination boundary between TimingStep-B-s 3 and 4, and the imitations for the first three and the last two TimingStep-B-s belong to two distinct distributions. It is also argued that the categorical patterns of tone-to-segment alignment are further subject to stimulus-induced adjustments by way of adjusting coupling strengths between lexical tone gestures and oral articulatory gestures. Last but not the least, it is also shown

that the TP relative timing and the TP F0 are influenced by the degree of gestural overlap between two tone gestures, which is in keeping with the gestural model of f_0 control.

3.6.1 Categorical Modes of Coordination

It is argued that the relative timing pattern is governed via categorical coordination between lexical tone gestures and articulatory gestures, which is in line with Hypothesis A1.

In both Experiment 1A and 1B, as TP progresses in the stimulus, the mean RelTiming-i increases. Moreover, the effect of TimingStep-s on RelTiming-i is significant. At first glance, Hypothesis A2 seems to be supported by these results. However, a close inspection of the results show that the mean differences in RelTiming-i are relatively small except for between later stimulus TP timing in Experiment 1B, as shown in Table 3.25. More importantly, the magnitude of variation in imitation TP relative timing is not on the order of the magnitude of variation in stimulus.

TimingStep-s: TimingStep-s	1:2	2:3	3:4	4:5
Exp. 1A (40 ms)	1.84 ms	**7.18 ms**	3.88 ms	**9.78 ms**
Exp. 1B (50 ms)	1.25 ms	**15.13 ms**	**54.82 ms**	**24.40 ms**

Table 3.25: Differences in mean RelTiming-i between any two adjacent TimingStep-s in Experiment 1A and 1B. Statistically significant differences are underlined and in bold.

Experiment 1A In Experiment 1A, multiple comparisons of RelTiming-A-i confirm that statistically, RelTiming-A-i at TimingStep-A-s 1 and 2 is different

from RelTiming-A-i at TimingStep-A-s 3 and 4, which is in turn different from RelTiming-A-i at TimingStep-A-s 5. The difference in the mean RelTiming-A-i is 7.18 ms between TimingStep-A-s 2 and 3, and 9.74 ms between TimingStep-A-s 4 and 5. Both are much smaller than the difference between two adjacent TimingStep-A-s—40 ms.

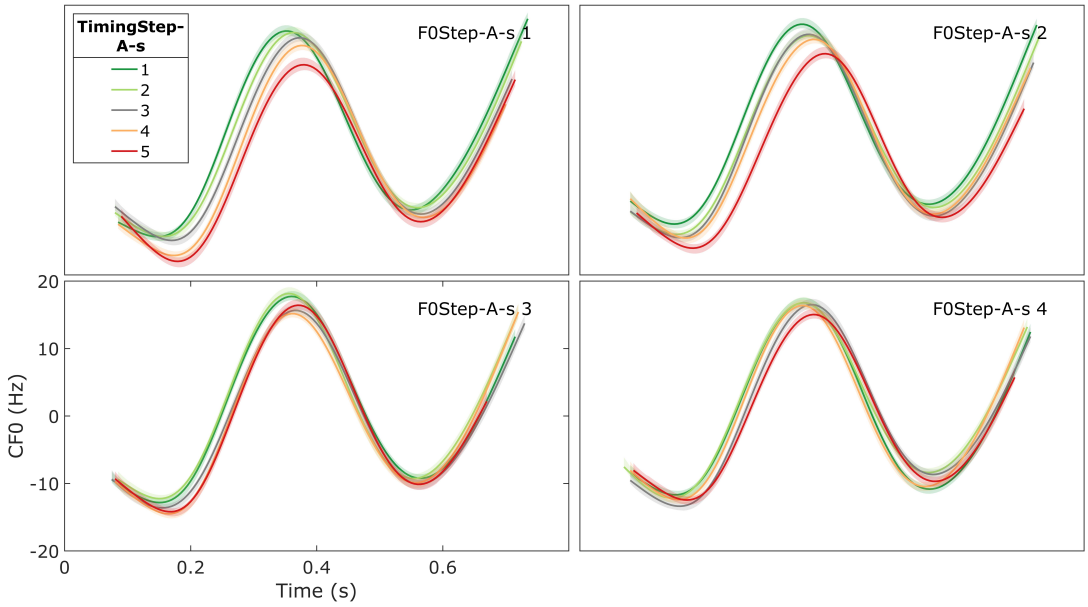


Figure 3.36: Mean f_0 contour at each TimingStep-A-s, grouped by F0Step-A-s, in Experiment 1A. Each color represents one TimingStep-A-s.

The differences in RelTiming-A-i should not be regarded as reflecting the differences in the alignment patterns. That is, these differences do not result from categorical changes in the coordinative pattern between lexical tone gestures and oral articulatory gestures. It would be difficult to account for the alignment phonologically, because they involve the assumption that native speakers of Mandarin consciously distinguish between two alignment patterns within 10 ms. Figure 3.36 shows that the average f_0 contour does not vary significantly with TimingStep-A-s for all F0Step-A-s. Recall that in Figure 3.15 that RelTiming-A-i is unimodally distributed for all TimingStep-A-s. The consistency in the alignment of TP1 can

only point to one mode of association between lexical tone gestures and oral articulatory gestures in the first tone-bearing syllable.

	All F0Step-A-s	F0Step-A-s 1	F0Step-A-s 4
F04	64	90	32
M01	42	47	11
F02	38	54	54
F01	34	48	14
F13	33	58	-5
F07	26	43	17
F10	23	21	15
F12	20	40	11
F06	16	30	1
F05	15	25	7
F11	13	-8	40
M03	8	2	33
F09	7	9	12
F08	7	21	4
M02	-1	6	11

Table 3.26: Differences (in ms) in the mean RelTiming-A-i between TimingStep-A-s 1 and 5 in Experiment 1A.

The difference in the mean RelTiming-A-i between TimingStep-A-s 1 and 5 amounts to 23 ms on average for all participants. Despite that it is still not comparable to the differences between the two stimulus TP timing (160 ms), a close inspection of the difference for individual participants suggests that some participants are more influenced by the stimulus TP timing than others, and might adopt a secondary association at later stimulus TP timing. Table 3.26 shows that the mean RelTiming-A-i difference between TimingStep-A-s 1 and 5 can go up as high as 64 ms for some participant, but remain negligible for others. Moreover, the timing difference is more apparent for lower F0Step-A-s than for higher F0Step-A-s. This is further supported by Figure 3.36: the differences in RelTiming-A-i (as well as F0-A-i) between TimingStep-A-s 1 and 5 are the largest for F0Step-A-s 1 and the smallest for F0Step-A-s 4.

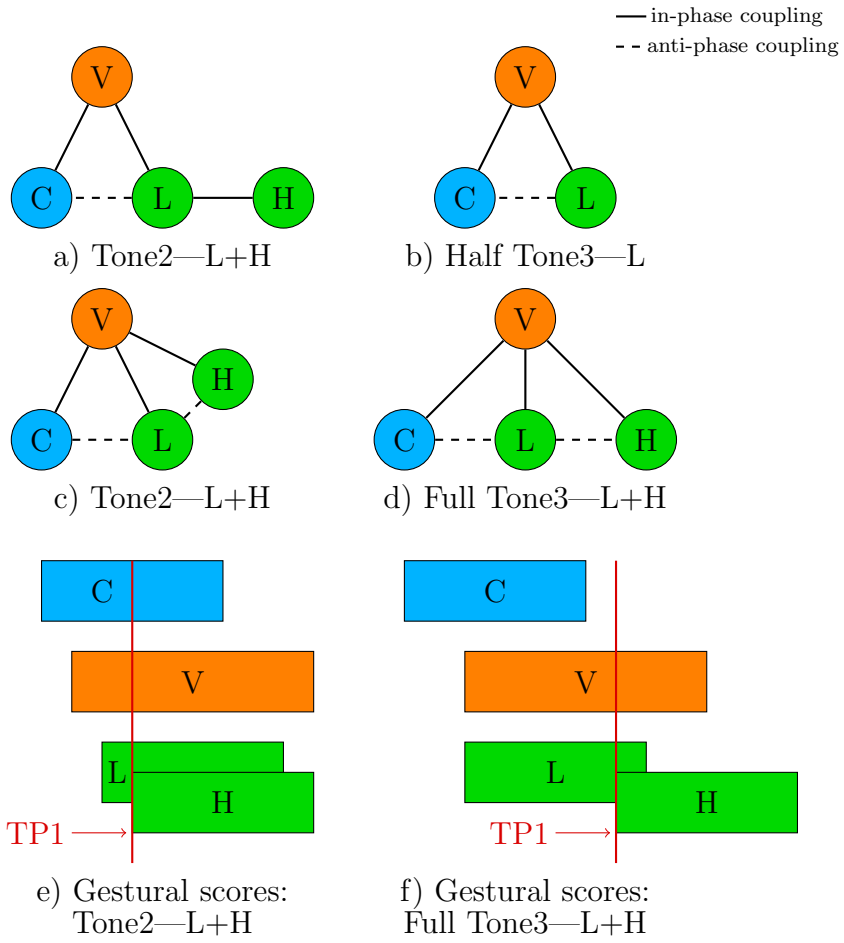


Figure 3.37: Coupling graphs of Mandarin Tone2 (a) and half Tone3 (b) based on Gao (2008). Proposed coupling graphs for Mandarin Tone2 (c) and full Tone3 (d) and their corresponding gestural scores (e-f).

The alignment patterns at later stimulus TP timing can be seen as a compromise between two categorical alignment patterns that correspond to Tone2 and Tone3, respectively. The two coordinative patterns are illustrated in Figure 3.37 (c-f). Recall Gao (2008) argued that lexical tone gestures are coordinated with oral articulatory gestures in a way that resembles onset C gestures. Taking this as the point of departure, it is proposed that the difference between Tone2 and Tone3 lies in the coupling strength between the H gesture and the C gesture: the C-H coupling strength in Tone2 is weak, while that in Tone3 is strong. As a re-

sult, the H gesture is attracted away from the C gesture in Tone3. Acoustically, the Low-to-High transition occurs late in the tone-bearing, as illustrated in Figure 3.37(e-f).

Note that this proposal differs from Gao (2008) in the coupling relations that involve the H gesture in Tone2 and in the definition of Tone3. Gao (2008) argued that the L and H gestures in Tone2 function as one C-like gesture such that the two gestures are initiated synchronously (Figure 3.37(a)). Tone3 analyzed in Gao (2008) is in fact half Tone3—Tone3 that is not phrase-final and does not precede another Tone3. A half Tone3 only has one L gesture (Figure 3.37(b)). With this analysis, the alignment patterns in Tone2- and Half Tone3-bearing syllables both exhibit the c-center effect.

However, Yi and Tilsen (2014) found that the coordinative patterns differ between Tone2 and Half Tone3 such that the V gesture is initiated later in the C-V-T sequence in Tone2-bearing syllable than in Half Tone3-bearing syllables, suggesting that the H gesture should not be seen as a "tag-along" gesture in Tone2. Instead, the H gesture introduces additional coupling relations with the V and L gestures. Specifically, the H gesture is in-phase coupled to the V gesture and anti-phase coupled to the L gesture. The V-H in-phase coupling force is weaker than the C-V and V-L in-phase coupling forces, and the L-H anti-phase coupling force is weaker than the C-L anti-phase coupling force, capturing that fact the V gesture is initiated after the midpoint between the C and the L initiations, although the latency between the V initiation and the midpoint is fairly small (Figure 3.37(c)). The Tone3 in the current analysis is Full Tone3, consisting of an L gesture and an H gesture occurring sequentially (Figure 3.37(d)). Moreover, the H gesture in Full Tone3 is on an equal footing with the C and L gestures in that the V-H in-phase

coupling force is as strong as the C-V and V-L in-phase coupling forces, and the L-H anti-phase coupling force is as strong as the C-L anti-phase coupling force. The Full Tone3 is the citation form of Half Tone3, representing the underlying structure of Tone3 in the lexicon.

The difference in gestural score between Tone2 and Full Tone3 is further illustrated in Figure 3.37(e) and (f). The V gesture is initiated closer to the L gesture in Full Tone3-bearing syllables thanks to the stronger coupling strength between the V and H gesture. This is reminiscent of the alignment pattern in Tone4-bearing syllables in which the H and L gestures are initiated sequentially, and both lexical tone gestures are coupled to the V gesture and to each other (Gao 2008, c.f. Figure 2.13(d1)). What is more relevant in the current study is that the H gesture is attracted away from the C gesture in Full Tone3-bearing syllables, compared to that in Tone2-bearing syllables, thanks to the increased L-H anti-phase coupling force. If a participant adopts the alignment pattern of a Full Tone3-bearing syllable, TP1 occurs later in the first [ma2]. In this case, the differences in RelTiming-A-i reflect the differences in the alignment pattern between Tone2- and Full Tone3-bearing syllables, and therefore can be seen as differences between two phonological categories.

To sum up, in Experiment 1A, the primary mode of tone-to-segment alignment does not vary categorically with TimingStep-A-s in most stimulus conditions. However, when TP1 occurs late in the tone-bearing syllable with low F_0, some participants can adopt a secondary mode of tone-to-segment alignment that resembles a Full Tone3-bearing syllable. The mean RelTiming-A-i at later stimulus timing, as a result, increases considerably. However, the increase is compromised by the fact that a Full Tone3 cannot occur at a non-final position, otherwise it would undergo

tone sandhi and become a Half Tone3. The mean increase in RelTiming-A-i is also mediated by the fact that the shift in the categorical mode of association is not shared by all participants. The latter compromising factor can also be contingent on the former in that native speakers tend not to shift from one category to another if the target category (Full Tone3) does not occur frequently in a certain context. In other words, most participants are biased towards the tone-to-segment alignment pattern of Tone2-bearing syllables regardless of TimingStep-A-s.

Experiment 1B Turning to Experiment 1B, RelTiming-B-i at TimingStep-B-s 1 and 2 is different from RelTiming-B-i at TimingStep-B-s 3, which is in turn different from RelTiming-B-i at TimingStep-B-s 4, and at TimingStep-B-s 5. The difference in the mean RelTiming-B-i is 15.13 ms between TimingStep-B-s 2 and 3, 54.82 ms between TimingStep-B-s 3 and 4, and 24.40 ms between TimingStep-B-s 4 and 5.

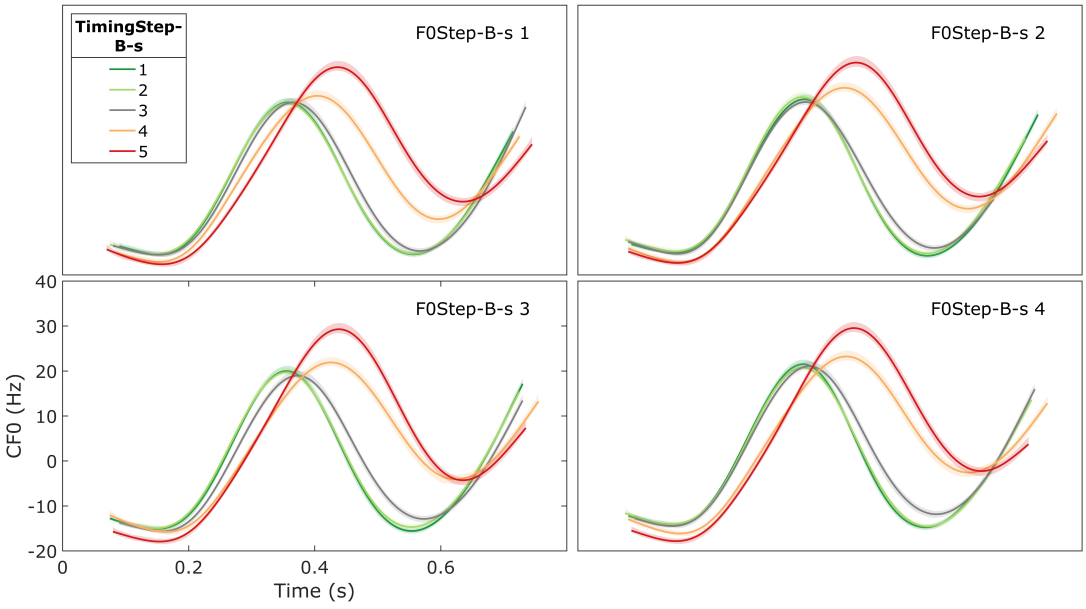


Figure 3.38: Mean f_0 contour at each TimingStep-B-s, grouped by F0Step-B-s, in Experiment 1B. Each color represents one TimingStep-B-s.

The differences in RelTiming-B-i are much larger than in RelTiming-A-i. Especially, the difference between TimingStep-B-s 3 and 4—54.82 ms—is comparable to the difference in time between two adjacent TimingStep-B-s—50 ms. Figure 3.38 also shows that the average f_0 contour at later stimulus TP timing (TimingStep-B-s 4 and 5) departs significantly from that at early stimulus TP timing. Also recall that in Figure 3.17 that the distribution of RelTiming-B-i shifts from unimodal to bimodal as TP2 progresses in the second tone-bearing syllable, suggesting that there are two modes of association between lexical tone gestures and oral articulatory gestures in Experiment 1B. Thus, the difference in RelTiming-B-i between certain pairs of TimingStep-B-s results from the categorical changes in gestural coordination pattern of the tone-bearing syllable.

Specifically, the alignment pattern that emerges at later stimulus TP timing is associated with Tone4-bearing syllables. Gao (2008) argued that Tone4-bearing syllables resemble English CCCV syllables in terms of gestural coordination, and that the V gesture is initiated after the midpoint between the gestural onsets of the C and the H gestures. When TP2 occurs late in the second [ma2], the tone-to-segment alignment is perceived as resembling that of a Tone4-bearing syllable. Accordingly, a coordinative system that resembled that of Tone4-bearing syllables is adopted to imitate the late TP2. In such a coordinative system, both H and L gestures function as onset C gestures, and thus the L gesture is in-phase coupled to the V gesture, and anti-phase coupled to the H gesture (c.f. Gao, 2008), as illustrated in Figure 3.39.

To sum up, in Experiment 1B, there are two primary modes of tone-to-segment alignment: one resembles that of a Tone2-bearing syllable and the other resembles that of a Tone4-bearing syllable. As TP2 progresses in the second [ma2], nearly

all participants switch to the latter alignment pattern, i.e., Tone4, by adopting a different coordinative system. In such a system, both the H and L gestures function as onset C gestures, and thus the L gesture is in-phase coordinated with the V gesture, and anti-phase coordinated with the H gesture. As a result, the initiation of the L gesture is much delayed, and acoustically, the imitation TP2 occurs significantly later in the second [ma2], which accounts for the fact that the increases in the mean RelTiming-B-i at later stimulus TP timing are fairly large.

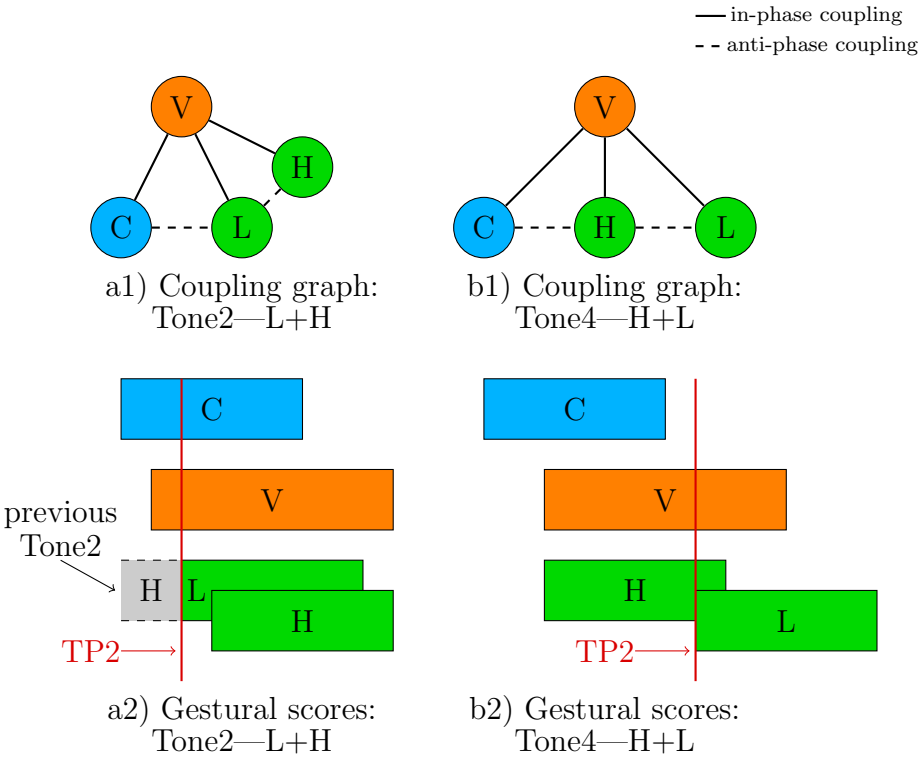


Figure 3.39: Schematic illustration of the shift in categorical mode of gestural coordination for the second tone-bearing syllable in Experiment 1B. Left: Tone2-bearing syllables; right: Tone4-bearing syllables. Top: coupling graphs; bottom: gestural scores.

The participants in Experiment 1B are more committed to the categorical changes in the coordinative pattern at later TimingStep-s than in Experiment

1A. This could be because Tone4 occurs much more frequently than Full Tone3 in Mandarin, and its occurrence is not location-specific. In other words, participants are no longer biased towards the tone-to-segment alignment of a Tone2-bearing syllable when TP2 occurs late in the second [ma2].

Perception-production Link The discrimination results further affirm that the participants perceive categorical modes of coordination of tone gestures and oral articulatory gestures. In Experiment 1A, there is no discrimination boundary on the stimulus timing continuum, indicating that the participants only perceive one category of tone-to-segment alignment, which corresponds to Tone2-bearing syllables. The categorical perception constrains the imitations so that most imitations fall into one category for the majority of the participants (Figure 3.13) Also note that for some participants, the perceived tone category shifts from Tone2 to Tone3 when the TP occurs late with low F_0, which results in larger differences in imitation. As illustrated in the left subplot of Figure 3.23, when the two stimuli to be differentiated are TimingStep-A-s 1 and 5 at F_0Step-A-s is 1 or 2, the discrimination performance is the highest, and the difference in the mean RelTiming-A-i is the largest.

In Experiment 1B, there is a discrimination boundary between TimingStep-B-s 3 and 4, indicating that the participants perceive two categories of tone-to-segmental alignment, which correspond to Tone2- and Tone4-bearing syllables, respectively. This further feeds the categorical production which gives rise to the two distinct distributions in imitation (Figure 3.13). The link between perception and production is illustrated in the right subplot of Figure 3.23: when the distance between TimingStep-B-s is higher than three, i.e., between TimingStep-B-s 1 and 4, between 2 and 5, and between 1 and 5, the discrimination success is above

chance ($> 50\%$), and the difference in the mean RelTiming-B-i is larger than 60 ms.

In sum, the high correlation between discrimination and imitation ($\rho = 0.75$ for Experiment 1A; $\rho = 0.98$ for Experiment 1B) corroborates the notion that there are categorical modes of alignment between tone gestures and oral articulatory gestures which the participants' perception and production are biased towards.

Stimulus-induced Changes The results presented here mainly support Hypothesis A1 over Hypothesis A2. Thus, the relative timing pattern is governed via categorical coordination of lexical tone gestures and articulatory gestures, and the imitations should further reflect such categorical changes in the coordination.

However, the categorical patterns of tone-to-segment alignment in the imitations are also subject to gradient stimulus-induced adjustments, which entails additional explanation of the gestural model. The gradient variation in the imitations is most readily observed in Experiment 1A, where TimingStep-A-s introduces significant increases in the mean RelTiming-A-s (e.g., between TimingStep-A-s 2 and 3, and between TimingStep-A-s 4 and 5), despite the fact that the increases are relatively small compared to the increase in the stimuli, therefore gradient. In Experiment 1B, RelTiming-B-i fall into two categories, which roughly correspond to the first three and the last two TimingStep-B-s. Within each category, an increase in TimingStep-B-s also leads to a significant increase in the mean RelTiming-B-i (e.g., between TimingStep-B-s 2 and 3, and between TimingStep-B-s 4 and 5).

This is consistent with a multi-level mental representation of the sound structures, which specifically includes sub-phonemic variation. Nielsen (2011) conducted an imitation study in which VOT on the phoneme /p/ is manipulated. She found

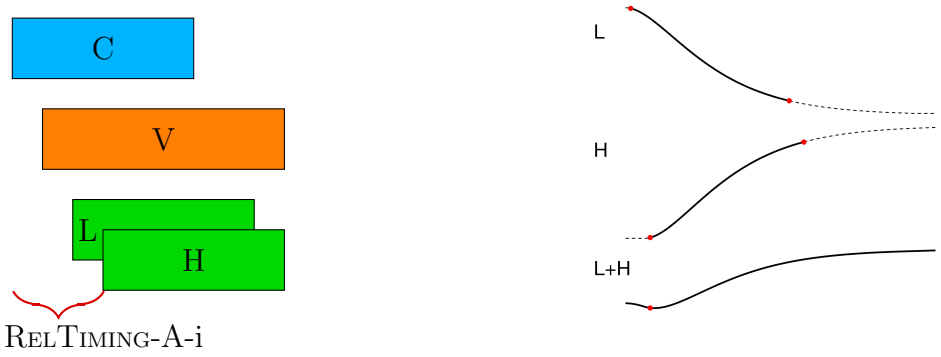
that speakers produce significantly longer VOT after being exposed to speech with extended VOTs. It is obvious that the manipulated features, i.e., VOT of /p/, should not be considered as categorical features. Instead, the VOTs are more fine-grained and gradient, because it is the degree of aspiration—not the categorical value of [spread glottis]—that is manipulated. Therefore, the evidence is in support of the claim that there exists sub-phonemic variation, and that sub-phonemic features like VOT can be imitated by speakers.

Turning to the current experiment, an apt analogy can be made: the phonemic level of representation corresponds to the categorical modes of tone-to-segment alignment, and the stimulus-induced variation in imitation timing, on the other hand, is of more gradient nature. In imitation, participants are biased towards the categorical modes of tone-to-segment but also are flexible in shifting the alignment in the direction of the perceived stimuli. Gesturally, the gradient stimulus-induced adjustments can be interpreted as the result of gradient changes in coupling strength between lexical tone gestures and oral articulatory gestures.

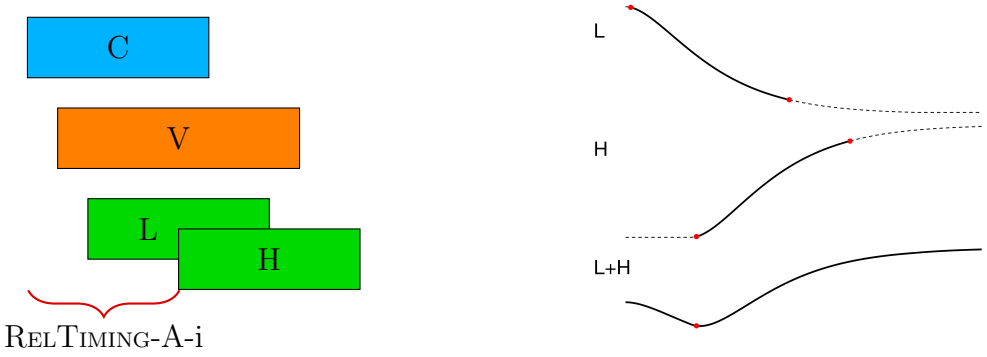
3.6.2 Relationship Between TP Timing and F_0

The results also support Hypothesis B1: the TP relative timing and the TP F_0 are influenced by the degree of gestural overlap between two tone gestures. Recall that in Experiment 1A, when the stimulus TP occurs late with low F_0, the imitation TP occurs later than the median RelTiming-A-i with F_0 lower than the median F_0-A-i, as can ben seen on the lower right corner in Figure 3.26(a). On the other hand, when the stimulus TP occurs early with high F_0, the imitation TP occurs earlier than the median RelTiming-A-i with F_0 higher than the median F_0-A-i, as shown on the upper left corner in Figure 3.26(a). These observations are in line with the

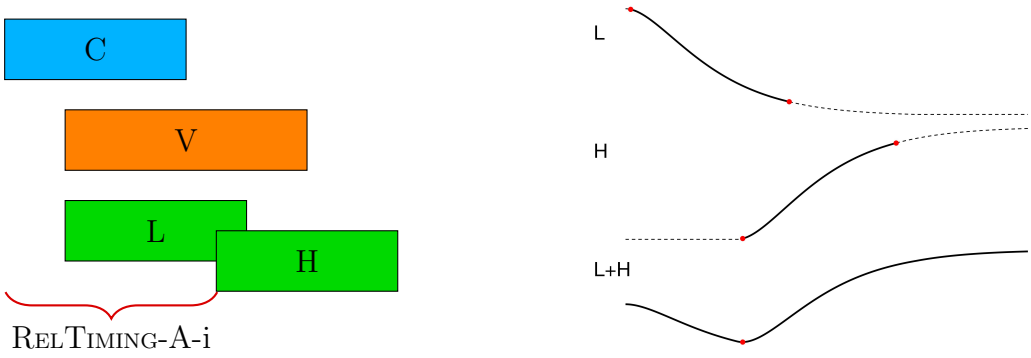
summary under "L+H" in Table 3.1. However, rather than the difference between gestural overlap and underlap, the differences in the tone-to-segment alignment result from different degrees of gestural overlap between the L and H gestures, which in turn result from gradient changes in the coupling strength between lexical tone gestures and oral articulatory gestures. This is illustrated in Table 3.27.



(A) Early TimingStep-A-s



(B) Mid TimingStep-A-s



(C) Late TimingStep-A-s

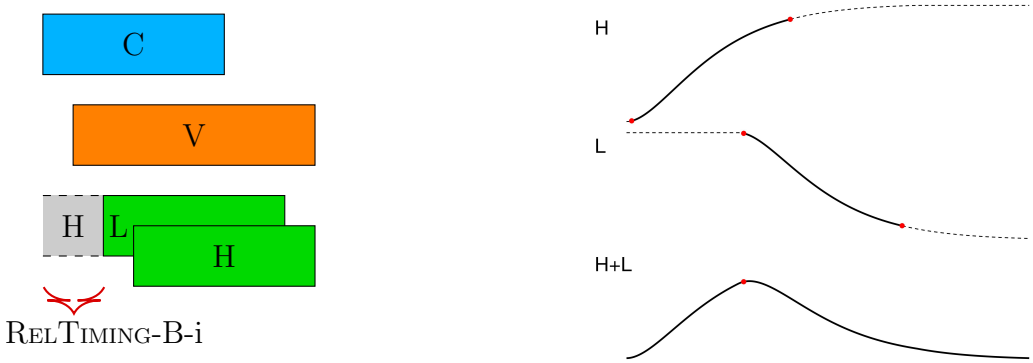
Table 3.27: Schematic illustrations of the imitations of Early (A), Mid (B), and Late (C) TimingStep-A-s in Experiment 1A.

At early TimingStep-A-s, the L and H gesture overlap greatly in the imitations to the extent that the H gesture is initiated not long after the L gesture is initiated. The duration of the downward movement of f_0 only lasts shortly before f_0 starts to rise. Therefore, the imitation TP occurs early with a relatively high F0 (Table 3.27(A)). As TP1 progresses on the timing continuum, the gestural overlap of L and H decreases. As a result, the imitation TP occurs later with lower F0 (Table 3.27(B)-(C)).

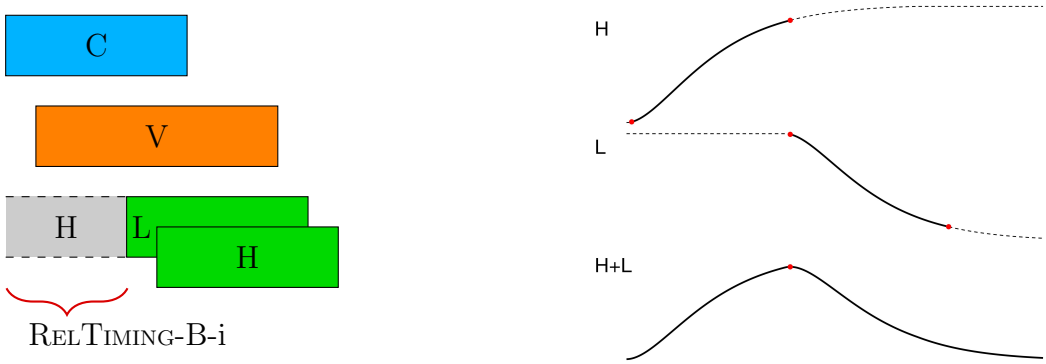
What is also worth mentioning is the other diagonal line—from upper right to lower left—in Figure 3.26(a). Along this diagonal, the effects of TimingStep-A-s and F0Step-A-s cancel out one another. On the upper right corner, the stimulus TP occurs late with high F0. However, late stimulus TP timing induces low F0. On the lower left corner, stimulus TP occurs early with low F0. However, early stimulus TP timing induces high F0. Thanks to the compromise between TimingStep-A-s and F0Step-A-s along this diagonal, the imitations do not sway significantly from the baseline production.

In Experiment 1B, the opposite pattern emerges in terms of the relationship between TP timing and F0. On one hand, when the stimulus TP occurs late with high F0, the imitation TP occurs later than the median RelTiming-B-i with F0 higher than the median F0-B-i, as shown on the upper right corner in Figure 3.29(a). On the other hand, when the stimulus TP occurs early with low F0, the imitation TP occurs earlier than the median RelTiming-B-i with F0 lower than the median F0-B-i, as can be seen on the lower left corner in Figure 3.29(a). These observations are in line with the summary under "H+L" in Table 3.1.

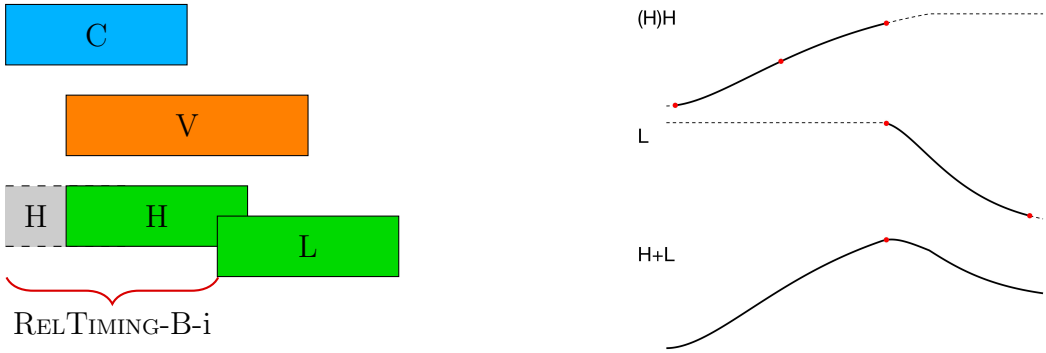
Similar to in Experiment 1A, the differences in the tone-to-segment alignment result from the differences in the degree of overlap between the H and L gestures, at least for the first three TimingStep-B-s. For earlier TimingStep-B-s, the greater the overlap between the H and L gesture overlap, i.e., the closer the H gesture is initiated after the L gesture is initiated, the shorter the duration of the upward movement of f_0 before f_0 starts to fall, the earlier the imitation TP occurs, and the higher the TP F0 (Table 3.28(A)). As TP2 progresses on the timing continuum, the gestural overlap of H and L decreases, and therefore the imitation TP occurs later with higher F0 (Table 3.28(B)). As for the late TimingStep-B-s, the coordinative system of Tone4-bearing syllables is posited by most participants. In this case, the L gesture of the second Tone2 is initiated around the target achievement time of the H gesture, which results in the largest RelTiming-B-i acoustically (Table 3.28(C)).



(A) Early TimingStep-B-s



(B) Mid TimingStep-B-s



(C) Late TimingStep-B-s

Table 3.28: Schematic illustrations of the imitations of Early (A), Mid (B), and Late (C) TimingStep-B-s in Experiment 1B.

Note that in Figure 3.29(a), along the diagonal line from upper left to lower right, the effects of TimingStep-B-s and F0Step-B-s do not cancel out one another. This is due to the relatively large effect size of TimingStep-B-s. However, along this diagonal, the effect of TimingStep-B-s is still slightly compromised by the effect of F0Step-B-s. For example, on the upper left corner, where the stimulus TP occurs early with high F0, the imitation TP occurs earlier than the median RelTiming-B-i with F0 lower than the median F0-B-i, as opposed to higher. However, the F0 deviation in these "conflicting" imitations is smaller than in the "congruent" imitations, which are along the diagonal line from upper right to lower left. The effect size of TimingStep-B-s is larger because there are two distinct modes of alignment in the imitations in Experiment 1B, while only some participants adopt a secondary mode of alignment in Experiment 1A.

As discussed above, the changes in the gestural coordination, depending on the participant, can either be sub-phonemic variation within the same category (Tone2) or categorical changes from Tone2 to Full Tone3 (Experiment 1A) or from Tone2 to Tone4 (Experiment 1B). Regardless of the nature of the changes, the relationship between RelTiming-i and F0-i should always hold from a gestural point of view, which is summarized in Table 3.27 and 3.28.

CHAPTER 4

EXPERIMENT 2

4.1 Introduction

The interaction between lexical tones and intonation has long been a contentious topic, and can be best assessed in the context of a tone language. In a tone language, f_0 is utilized by both lexical tones and intonation. The difference is that lexical tones make use of f_0 contrastively for lexical or grammatical meaning, and intonation uses f_0 for the expression of disclosure meaning and for marking pauses. Two categories of models have been proposed to account for the interaction—the overlay model and the unification model.

Previous research has treated this issue an acoustic one, providing evidence of f_0, intensity, duration, etc (Yuan, 2004; Gibson, 2013). Few studies have made articulatory arguments about the tone-intonation interaction. However, a large body of work within the framework of Articulatory Phonology has been done to pave the way for an alternative, i.e., articulatory, perspective on this issue. Thus, the objective of the current experiment is to examine the interaction between lexical tones and interaction within the framework of Articulatory Phonology.

As discussed before, the control of f_0 can be conceptualized as gestures that are coordinated with oral constriction gestures to form larger phonological units such as syllables, moras, and words. Both lexical tones and intonational tones have been analyzed as gestures, although the articulatory alignment patterns differ depending on the nature of the tone gesture (Gao, 2008; Mücke et al., 2012; Katsika et al., 2014; Yi and Tilsen, 2014). The tone-intonation interaction can thus be

investigated by analyzing the temporal coordinative patterns of f_0 gestures and oral constriction gestures under this framework.

Five (5) female native Mandarin speakers participated in a production experiment using Electromagnetic Articulography (EMA). Participants were asked to read the prompted sentences according to the context on the screen. Each sentence ended with either a question mark or period, eliciting the high and low boundary tones respectively. The target syllable [ma] (Tone2 or Tone4) was embedded in each sentence, and it occurred either at the phrase-final position or the phrase-medial position. Furthermore, each target syllable can either be accented or un-accented depending on the context. The target syllable that occurred at the phrase-final position with or without prosodic focus rendered itself the site where the lexical tones interacts intonation. The articulatory alignment pattern between oral constriction gestures and f_0 gestures (including both lexical and intonational tone gestures) were measured.

4.2 Background

4.2.1 Interaction Between Lexical Tones and Intonation

Previous research has primarily focused on providing phonological accounts for intonation. For example, Autosegmental-Metrical theory provides a detailed account of intonational use of f_0 for pitch accents, phrase accents, and boundary tones in non-tonal languages. Since the overall f_0 is the result of both lexical tones and intonation in lexical tone languages, can f_0 models like the AM theory of intonation be tasked to provide a similar theoretical framework for lexical tones, and to

further explain the interaction between lexical tones and intonation?

In fact, two categories of f_0 models have been proposed to account for the tone-intonation interaction: the overlay model and the AM model (Ladd, 2008). Briefly, the overlay model characterizes the overall f_0 as the result of local perturbations of lexical tones imposed onto the global f_0 contour that carries out intonational functions related to statements, wh-questions, yes-no questions, early focus, late focus, etc. The AM model argues a f_0 contour in any language is simply a sequential structure of of tones, whether their functions are lexical or intonational, therefore there should be no difference in this regard between a lexical tone language and a non-tonal language. Because the AM model treats lexical tones and intonational tones in a unifying way, i.e., they are the same type of phonological entity with different functions, this model will be referred to as the unification model. The two models are described in detail below.

4.2.1.1 Overlay Model

The overlay model is most famously represented in the "ripples-on-waves-on-swells-on-tides" metaphor by Bolinger (1964): "the ripples are the accidental changes in pitch, the irrelevant quavers. The waves are the peaks and valleys that we call accent. The swells are the separations of our discourse into its larger segments. The tides are the tides of emotion." In this view, the "ripples" correspond to the segment-induced micro-prosodic changes in f_0—'the irrelevant quavers'. The 'waves' correspond to word-level accents or lexical tones. The "swells" and "tides" correspond roughly to sentence-level accents or intonation related to the distinction between statements and questions.

Gårding's (1983) grid model is an example of an overlay model. Word-level accents or lexical tones were modeled as local turning points, and intonation is modeled as a grid which the locally determined turning points are superimposed onto. In mathematical implementation, two grid lines are fitted to the local minima and maxima, respectively, to approximate the f_0 range as well as the direction as shown in Figure 4.1.

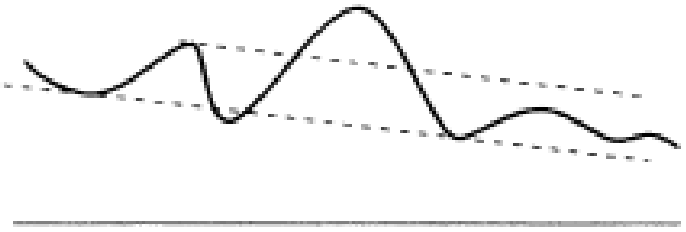


Figure 4.1: Illustration of grid in Gårding's model of intonation. The dashed lines fitted to most of the local minimal and maximal provide an approximation of the f_0 range and the direction of slope. From Gårding (1983) in Ladd (2008).

Fujisaki's (1983) command-response model is another example of an overlay model. Initially proposed to model the lexically specified pitch accents and sentence-level declination in Japanese, Fujisaki (1983)'s model was used to model the interaction between lexical tones and intonation in Cantonese. This model employs the same idea as Gårding (1983), independently implementing locally specified pitch accents/lexical tones and global intonation via "accent command" and "phrase command", respectively. Specifically, the "phrase command" was set up to handle the declination by fitting a exponentially decaying function. The "accent command" is modeled as a sequence of ups and downs that represent the rise and fall in the localized pitch accents. The pitch accent sequence is finally added to the decaying function to compute the f_0 output.

Xu (2005) proposed Parallel Encoding and Target Approximation model, also known as the PENTA model, to tackle the f_0 modeling in Mandarin Chinese. The name of the model gives away the parallel nature in the implementation of sentence-level intonation and lexical tones. Specifically, the PENTA model involves parallel encoding of various communicative functions including lexical, sentential, focal, topical, which can be further expressed as one of the four phonetic primitives: local pitch targets, pitch range, articulatory strength, and duration. Therefore, the communicative functions, which the parameters of the phonetic primitives are associated with, contribute independently to the final f_0 output.

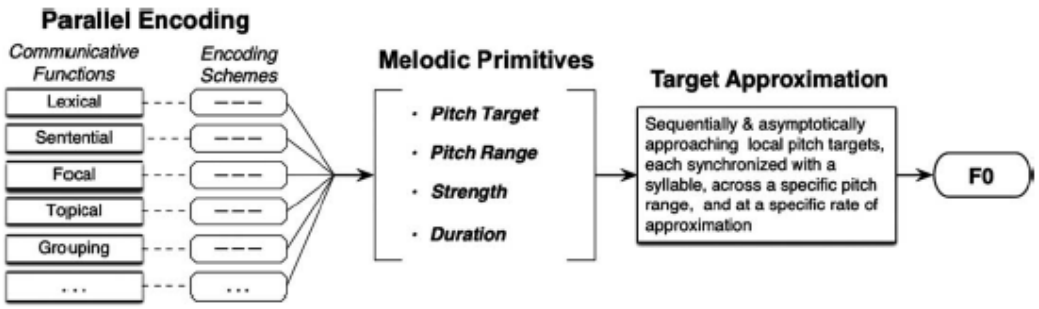


Figure 4.2: A schematic representation of the PETNA model. From Xu (2005).

4.2.1.2 Unification Model

Recall that in Swedish Accent I and II words, the word-level accent and the sentence-level accent form a tone string, and the difference in surface f_0 contour between the two types of words result from the difference in the alignment between the word-level accent and the segmental string. The Swedish example elegantly argues for the same treatment for the word-level accents and sentence-level accents in the f_0 model, which can be further extended to the same treatment for "lexically" specified pitch accents and tone, and "post-lexical" intonation such as phrase

accent and boundary tone. In such an f_0 model, because lexical tones and intonational tones are treated with no difference from the standpoint of phonology, they interact at the same stage, constituting a mixture of tone targets, before a unifying phonetic implementation takes place. Therefore, the AM theory of Intonational Phonology should be placed squarely in the unification model camp.

Pierrehumbert and Beckman (1988) modeled Japanese speech prosody under this framework. In their autosegmental model, lexical pitch accents are assigned to the accented mora, and boundary tones are assigned to higher-level "accentual phrases" and "utterances". The string of pitch accents and boundary tones form the underlying representation of tone targets, which is further mediated by the phonological processes, i.e., the lexical pitch accents interact locally with the boundary tones according to certain phonological rules. The output phonological representation, different from the underlying tonal sequence, is taken up by the phonetic encoding scheme to produce the surface f_0 contour.

4.2.1.3 Comparison of Two Models

The unification model originates from the body of research into segmental anchoring, whereas the overlay model is intuitively more appealing in dealing with the interaction between lexical tones and intonation. One example often put forward as strong evidence for the overlay model is in Mandarin Chinese the overall f_0 contour of a statement is distinctly different from that of a question, provided that the two sentences are identical in composition.

Shen (1990, as cited in Ladd, 2008) showed that three distinct intonational tunes are associated with statements (solid line), wh-questions (dash-dot line), and yes-no questions (dashed line), as illustrated in Figure 4.3. For instance, the

f_0 contour of a statement is almost parallel to that of a yes-no question if the phrase-final part of the contour is left out of consideration. This observation fits right into the interpretation of the overlay model: the difference in the overall f_0 contours between a statement and a question is the global intonation which the local lexical tone targets are superimposed on. In this case, a yes-no question has an overall higher f_0 contour. However, the two f_0 contours are not entirely parallel thanks to the final f_0 rise in yes-no questions. This is not beyond the reach of the overlay model: it can be stipulated that the intonational tune associated with a yes-no question in Mandarin Chinese rises towards the end of an utterance in addition to having an overall higher f_0 contour than a statement.

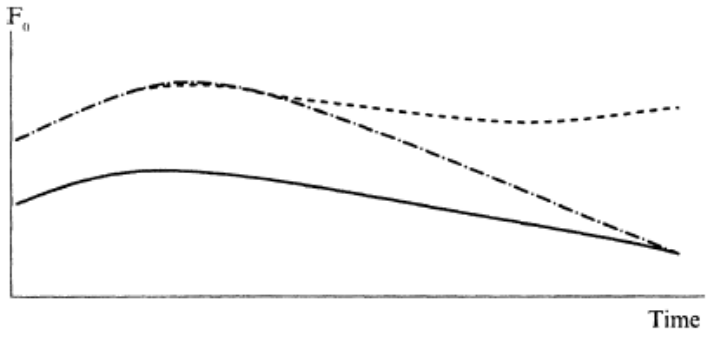


Figure 4.3: Three distinct tunes are associated with statements (solid line), wh-questions (dash-dot line), and yes-no questions (dashed line) in Mandarin Chinese. From Shen (as cited in Ladd, 2008).

In order to accommodate the above observations in the framework of AM theory of Intonation, Ladd (2008) argued that the overall f_0 contour difference between a statement and a yes-no question arises out of successive expansions or contractions of local pitch ranges. That is, for both types of utterances, the pitch range of the lexically specified tone targets is expanded or contracted through the interaction with intonation. In this case, in a yes-no question, the pitch ranges of the local pitch targets are expanded. Moreover, intonational tones, e.g., the high boundary tone, fixated at the end of the utterance, interact with lexical tones, forming a

serialized phonological representation before the phonetic implementation takes place.

Another example of contention comes from Hausa, a tone language (Lindau, as cited in Ladd, 2008). In a Hausa statement, the H's are realized with progressively lower f_0 in a tone sequence like H-L-H-L-H-L. This phenomenon is often referred to as 'downstep', which is conditioned by an alternating tone sequence. Such a downward is not observed in a non-alternating sequence like H-H-H-H-H in a statement. Despite the slight declination in f_0, the top line of f_0 (indicating the upper bound of the f_0 range) is much less steep than than in an alternating sequence. Therefore, Hausa intonation patterns influence the realizations of the lexically specified tones. More importantly, the implementation of the intonation patterns is phonologically conditioned by the tonal make-up of the sequence. Citing this example, Ladd (2008) argued that the overlay model cannot provide a convincing explanation for the observation that the same type of intonation, i.e., statement, produces vastly different f_0 slopes for different tone sequences. Instead, he argued that in keeping with the unification model, the intonation patterns in Hausa are produced locally, operating the downstep rule successively depending on the tonal make-up of the sequence.

It is possible that both the overlay model and the unification model are adopted in implementing intonational processes in a lexical tone language. Yuan (2004) argued that statement intonation in Mandarin Chinese is unmarked, and that question intonation is marked and thus requires certain question mechanisms. Three mechanisms for question intonation were proposed: an overall higher phrase curve, higher strengths of sentence final tones, and a tone-dependent mechanism that flattens the final falling tone and steepens the final rising tone.

The phrase curve mechanism raises the overall f_0 curve for question, compared to the unmarked statement intonation. Therefore, it was argued that the phrase curve mechanism is global and tone-independent. However, as discussed above, the overall f_0 difference can be accounted for by successive expansion of local targets. The strength mechanism explains longer duration of sentence-final syllables and higher strength of sentence-final tones, therefore is not global and but tone-independent. The last mechanism is strictly local and tone-dependent.

To sum up, in the overlay model, intonation and lexical tones are encoded and implemented in a parallel fashion. The phonetics accomplishes most of the implementation without a significant contribution by phonological processes. That is, there is no phonological interaction between intonation and lexical tones. This model, as argued by Gibson (2013), predicts no lexical-tone-specific effects of utterance-level intonation or focus.

On the other hand, the unification model holds that intonational tones are produced locally, and they interact with lexical tones before the phonetic implementation takes effect. The primary difference is that phonological processes take place before the phonetics in the unification model. As a result, the unification model will predict phonologically conditioned, i.e., lexical-tone-specific, effects of utterance-level intonation or focus.

4.3 Hypotheses and Predictions

Whether the overlay model or the unification model best characterizes the tone-intonation interaction in Mandarin is investigated in the current experiment. A carefully designed corpus introduces various types of interaction between lexical

tones and intonational events such as prosodic focus and boundary tones. The target syllables are embedded in the phrase-medial or phrase-final position. The phrase-final contexts are designed to introduce boundary tones elicited by statements or questions. The evidence is drawn from the relative timing patterns between the oral constriction gestures—the C and V gestures—and the f_0 gestures of the target syllables in various stimulus conditions.

It has been established that in Mandarin lexical tone gestures behave like additional onset gestures, which gives rise to the c-center in a tone-bearing CV syllable: the V gesture is approximately initiated halfway between the gestural initiations of the C and the lexical tone gestures (Gao, 2008). On the other hand, intonational tones, such as German and Catalan pitch accents and Greek boundary tones, do not alter the intra-syllabic relative timing of gestural activations of the C and V gestures (Mücke et al., 2012; Katsika et al., 2014). Lexical tones thus contrast with pitch accents and boundary tones in that lexical tones gestures appear to be more tightly integrated to the intra-syllabic coordinative network than the intonational gestures. Even though these studies only investigated the effects of lexical tones and intonational tones on gestural coordination of articulatory gestures in lexical tones (Mandarin) and non-tonal (German, Catalan, and Greek) languages, respectively, they pointed to a new avenue for investigating the tone-intonation interaction: the effects of the intonational tones (such as boundary tones and focus) on the intra-syllabic articulatory alignment between the C, V and lexical tone gestures.

The overlay model predicts that the presence of intonational processes such as prosodic focus and boundary tones will not substantially alter the intra-syllabic relative timing of gestural activations of the C, V and T gestures (C-V-T). However,

under this model, the imposition of the f_0 contour may alter the detection of the T gestural onsets, and slightly alter the C-V-T coordinative patterns. On the other hand, under the unification model, intonational tones and lexical tones are the same type of phonological entities, and arise out of the same type of f_0 control. Therefore, intonational processes can influence the realizations of lexically specified tones, and furthermore the coordination between lexical tone gestures and the oral articulatory gestures. It follows that the C-V-T coordinative pattern at the phrase-final position will differ from that at the phrase-medial position, and that the C-V-T coordinative pattern of un-accented syllables will also differ from accented syllables.

The hypotheses and predictions are detailed as follows:

Hypothesis H1: (Overlay model) Intonational processes operate globally, and intonational tones and lexical tones are implemented in a parallel fashion.

Prediction P1: The relative timing of gestural activations of C-V-T of the target syllable will not be altered substantially in the presence of intonational events such as focus and boundary tones.[1]

Hypothesis H2: (Unification model) Intonational processes operate locally, and intonational tones can influence the realizations of lexical tones, and consequently, the gestural coordination of lexical tones gestures and oral articulatory gestures.

Prediction P2: The relative timing of gestural activations of C-V-T of the target syllable may be altered because of the intonational events such as focus and boundary tones.

[1]Under this model, the C-V-T coordinative patterns may also change, albeit not significantly, because the intonational contour alters the detection of the f_0 gestural onsets. See Section 4.6.

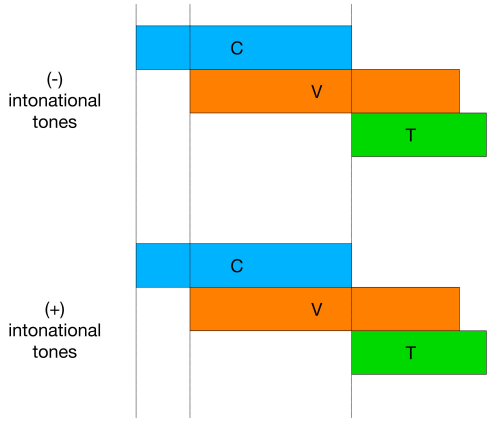


Figure 4.4: Schematic illustration of prediction for Hypothesis H1. The C-V-T coordinative patterns remain unchanged in the presence of intonational tones such as prosodic focus and boundary tones.

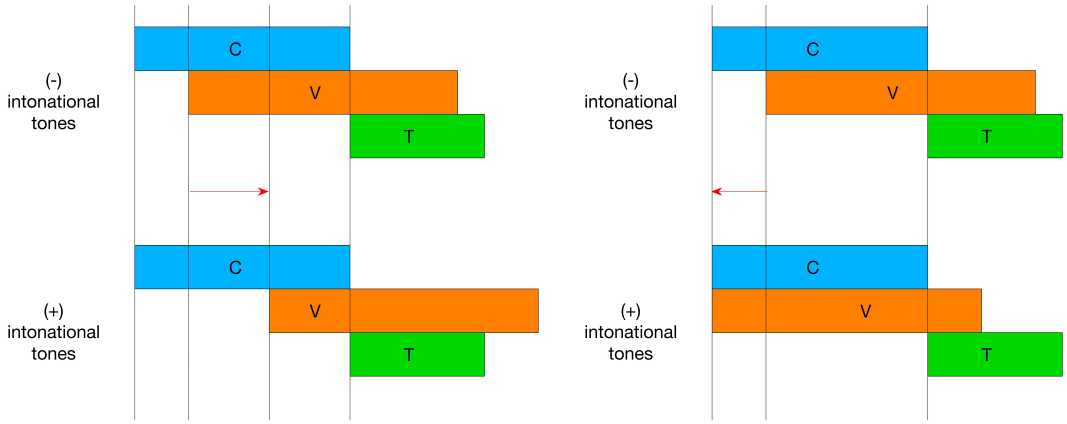


Figure 4.5: Schematic illustration of prediction for Hypothesis H2. The C-V-T coordinative patterns change in the presence of intonational tones such as prosodic focus and boundary tones. Note that the direction of the change of the relative alignment of the V gesture is yet unknown.

The predictions for Hypothesis H1 and H2 are illustrated in Figure 4.4 and Figure 4.5. Note that in the latter prediction, the change in the relative phasing of the V gesture is conditioned by the change in the relative coupling strength between gestures, therefore cannot be predicted here.

4.4 Methodology

4.4.1 Participants

Five female participants who are native speakers of Beijing Mandarin participated in this experiment. They were born and raised in Beijing, and were graduate students at Cornell University at the time of recording. The participants were naïve to the purpose of the study. They gave informed consent and received financial compensation for their participation. The experiment took place in the Cornell Phonetics Lab in the Department of Linguistics at Cornell University.

4.4.2 Stimuli Construction

The stimuli were a series of Mandarin utterances in which the target syllable [ma] was embedded. The target syllable [ma] was composed of a labial consonant [m] and a low central vowel [a], and was preceded by a high front vowel [i]. This segmental sequence, i.e., [ima], allowed for a clear observation of articulatory movements associated with the consonants and vowels. This is because the flesh points corresponding to articulators involved in [i] and [m], and in [m] and [a], are located on different speech apparatuses: tongue for [i] and [a], and lips for [m]. Moreover, the adjacent vowels [i] and [a] have contrastive tongue positions in terms of both height and backness: [i] is a high front vowel and [a] is low back vowel.

The target syllable either bore Tone2 (rising) or Tone4 (falling), and was always immediately preceded by a syllable bearing the same tone. Because Tone2 has a low onset and a high offset, and Tone4 has a high onset and a low offset, the

juxtaposition of two of the same tone rendered a contrasting tonal environment between the preceding and the target syllable, assuring clear observation of f_0 movement.

As noted before, the target syllable was embedded in one of the three types of phrasal contexts: phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc). Each phrasal context further adopted one of two types of sentence-level intonation: Question and Statement, respectively indicated by a question mark and a period. Two elicitations can be exactly the same except for the punctuation at the end of the sentence. High boundary tone (H%) and low boundary tone (L%) were elicited with Question and Statement, respectively. Note that boundary tones were elicited only in the phrase-final phrasal contexts.

For the phrase-final (FinUn and FinAc) elicitations, an prompt preceded the target elicitation to induce the intended prosodic context. Moreover, for these phrasal contexts, an extra intonational phrase immediately followed the target elicitation in order to render it more felicitous. Specifically, a short statement followed a Question elicitation, and a short question followed a Statement elicitation. No prompts or extra intonational phrases were provided for the phrase-medial (MedUn) elicitations.

Stimuli for Question and Statement elicitations of Tone2-bearing target syllables are shown in Table 4.1 and 4.2. Stimuli for Tone4 elicitations are shown in Appendix B.

Phrasal Context	Target	(Prompt) Stimulus
MedUn	Tone2	(–) lu^2 yan^2 yi^2 *ma*^{2} yi^2 de^5 hen^3 kuai^4? 'Lu Yan moves *ma*^{2} very fast?'
FinUn	Tone2 + H%	(bu^2 shi^4 luo^2 yan^2? bu^2 shi^4 li^3 yan^2?) ('Not Luo Yan? Not Li Yan?') **lu**^{2} yan^2 yao^4 yi^2 *ma*^{2}? mei^2 ting^1 shuo^1. '**Lu** Yan will move *ma*^{2}? I have not heard.'
FinAc	**Tone2** + H%	(bu^2 shi^4 ma^3? bu^2 shi^4 ma^4?) ('Not ma^3? Not ma^4?') lu^2 yan^2 yao^4 yi^2 ***ma***^{2}? mei^2 ting^1 shuo^1. 'Lu Yan will move ***ma***^{2}? I have not heard.'

Table 4.1: Question elicitations of Tone2-bearing target syllables in three different phrasal contexts: phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc). Target syllables are italicized. Accented syllables are in bold. Prompts are provided for FinUn and FinAc elicitations.

Phrasal Context	Target	(Prompt) Stimulus
MedUn	Tone2	(−) lu^2 yan^2 yi^2 *ma*^{2} yi^2 de^5 hen^3 kuai^4. 'Lu Yan moves *ma*^{2} very fast.'
FinUn	Tone2 + L%	(bu^2 shi^4 luo^2 yan^2. bu^2 shi^4 li^3 yan^2.) ('Not Luo Yan. Not Li Yan.') **lu**^{2} yan^2 yao^4 yi^2 *ma*^{2}. mei^2 ting^1 shuo^1 ma^5? '**Lu** Yan will move *ma*^{2}. Have you not heard?'
FinAc	**Tone2** + L%	(bu^2 shi^4 ma^3. bu^2 shi^4 ma^4.) ('Not ma^3. Not ma^4.') lu^2 yan^2 yao^4 yi^2 ***ma***^{2}. mei^2 ting^1 shuo^1 ma^5? 'Lu Yan will move ***ma***^{2}. Have you not heard?'

Table 4.2: Statement elicitations of Tone2-bearing target syllables in three different phrasal contexts: phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc). Target syllables are italicized. Accented syllables are in bold. Prompts are provided for FinUn and FinAc elicitations.

4.4.3 Procedure

Participants were seated half a meter away from a monitor, on which the stimuli were presented. Articulatory data were collected with an NDI^{TM} Wave Electro-

magnetic Articulography (EMA). EMA tracks real-time articulatory movements in speech production using small sensors attached to the articulators, such as the lips, jaw, and tongue. For the purpose of the current experiment, eight sensors were used: three sensors attached to the nasion, the left, and right mastoid as reference points; one sensor each attached to the jaw (JAW), the upper lip (UL), the lower lip (LL), the tongue tip (TT), and the tongue body (TB), to tract the movement of the respective articulator. The TT sensor was placed approximately one centimeter behind the tip of the tongue. The TB sensor was placed 4-5 cm posterior of the TT sensor. Acoustic data were simultaneously collected at the sampling rate of 22.5 kHz.

The experiment was organized into blocks. In each block, 16 ([Tone2, Tone4] $\times$ [Question, Statement] $\times$ 4) stimuli (containing target syllables) of the same phrasal context appeared in random order. The stimuli were presented in simplified Chinese characters. The prompts were presented on the same screen as the stimuli. However, the participants were instructed to say only the stimuli as naturally as possible. Subvocalization was also encouraged for these elicitations in order to render the prosodic context more accurately. Hence, more time was allotted for the phrase-final elicitations than the phrase-medial elicitations. During the familiarization phase, if a participant was judged to be speaking unnaturally, including pausing after the target syllable in a MedUn elicitation, the experimenter would demonstrate saying the stimuli without pause in a more colloquial way until the participant can perform the intended task. Each participant completed approximately 16 blocks. Including preparation time, each experiment session took approximately 90 minutes.

4.4.4 Data Processing

Segmentation was done using the HTK forced alignment algorithm, previously described in Section 3.4.4. 16 MFCCs were extracted for each elicitation. The mini dictionary created for this experiment consisted of all the syllables in the stimuli, such as *lu*, *yan*. For each speaker, roughly 2% of the elicitations were manually labeled before being submitted to the forced alignment algorithm. An example of the forced alignment is shown in Figure 4.6.

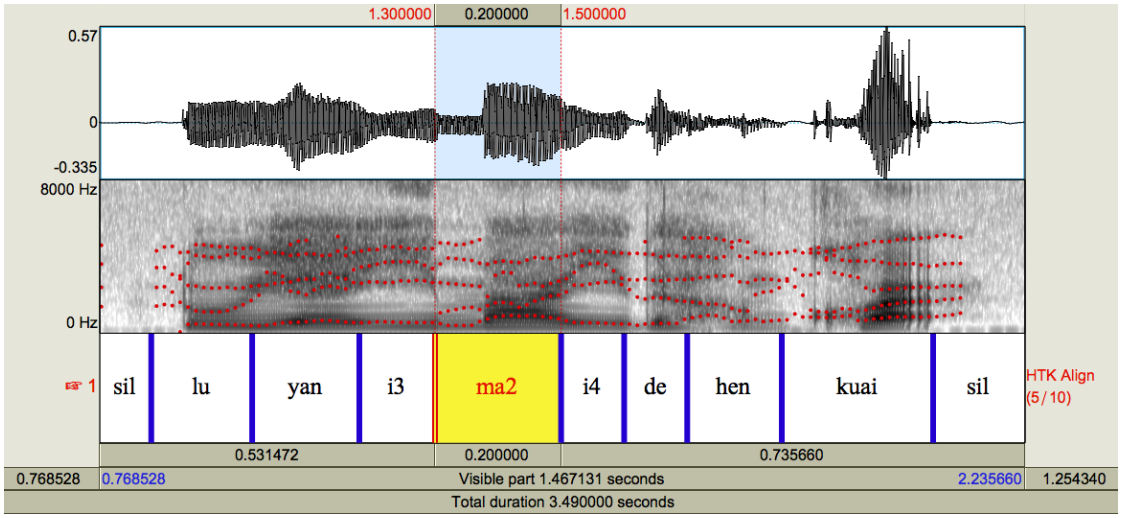
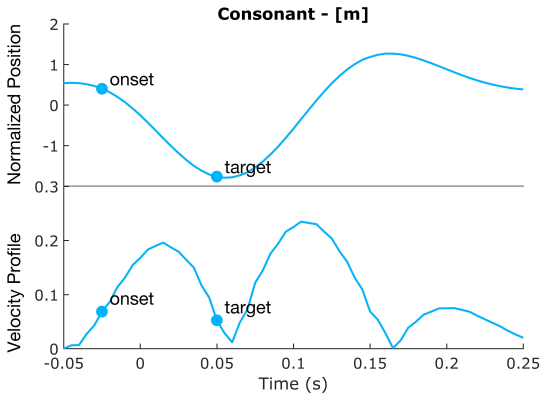


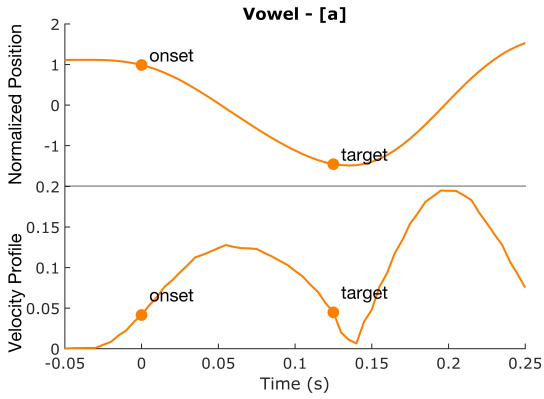
Figure 4.6: Example of using HTK to segment a MedUn elicitation.

Kinematic and f_0 contours were extracted. The articulatory gestures involved in the target syllable [ma] are a bilabial closure (C) gesture for [m] and a tongue root retraction (V) gesture for [a]. The C gesture is associated with the lip aperture (LA), the vertical distance between UL and LL; the V gesture is associated with the tongue body height (TBy), the vertical displacement of TB. Therefore, the trajectories of LA and TBy were extracted from the EMA signals. f_0 contours, with which the f_0 gesture is associated, were extracted using *ProsodyPro* and *B-spline* smoothing, previously described in Section 3.4.4.

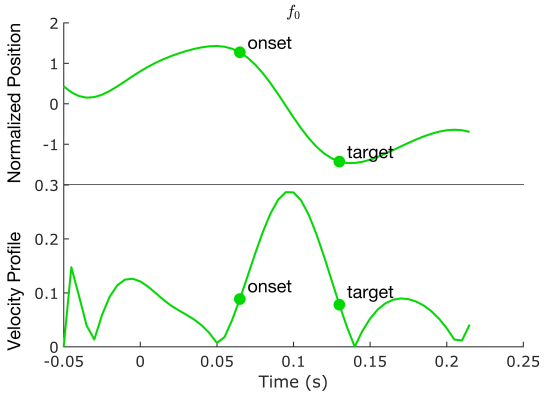
For both articulatory tract variables—LA (Figure 4.7(a)) and TBy (Figure 4.7(b)), their corresponding velocity profiles were computed to determine the articulatory landmarks: minimum and maximum velocity, onset, and target, etc. The onset was defined as the point in time when, starting from the minimum velocity, 30% of the velocity range (between the minimum and maximum velocity) had passed; the target was defined as the point in time when, starting from the maximum velocity, 70% of the velocity range had passed. The f_0 landmarks were defined in the same way as the kinematic landmarks (Figure 4.7(c)). Importantly, the onset of a high f_0 gesture was defined with reference to the f_0 minimum immediately preceding the f_0 peak (high); the onset of a low f_0 gesture was defined with reference to the f_0 maximum immediately preceding the f_0 valley (low).



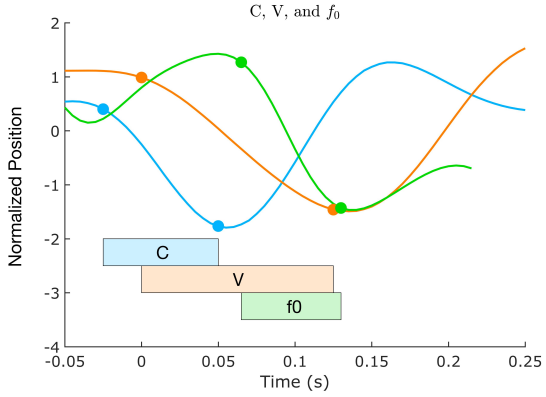
(a) Lip Aperture (LA)



(b) Tongue Body Height (TBy)



(c) f_0



(d) Overlay of C, V, and f_0 tract variables

Figure 4.7: (a-c) Example of the normalized contour and the corresponding velocity profile of the tract variables lip aperture (LA), tongue body height (TBy), and f_0 for a Tone2-bearing [ma]. "Onset" and "target" represent the onset and target of a gesture, determined by the 30% threshold on the corresponding velocity profile. (d) Top: Overlay of the tract variables LA (blue), TBy (orange), and f_0 (green). Bottom: C, V, and f_0 gestures associated with the tract variables.

Temporal lags between onsets were calculated from the landmarks, as shown in Figure 4.8. The C-V lag was defined as the temporal lag between the onset of the C gesture and the onset of the V gesture; the C-T lag was defined as the temporal lag between the onset of the C gesture and the onset of the T gesture. The phase of the V gesture relative to the C-V-T lag was further computed for each trial (CV% = C-V lag / C-V-T lag $\times$ 100%). The smaller the CV%, the closer the activations of the C and V gesture; the larger the CV%, the closer the activations of the V and T gesture.

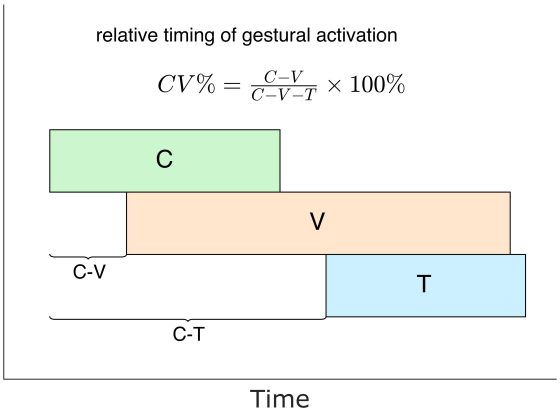


Figure 4.8: Illustration of the measurement: the relative phasing of the V gesture (CV%).

4.4.5 Data Analysis

The CV% was submitted to a mixed effects model, in which the fixed and random terms were as follows:

Context (fixed): three types of phrasal contexts—phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc);

SentInto (fixed): two types of sentence-level intonation—question (Question) and statement (Statement), which corresponds to the high boundary tone (H%) and the low boundary tone (L%), respectively;

Tone (fixed): two types of lexical tones—Tone2 and Tone4;

Participant (random): participants of the current experiment.

4.5 Results

Section 4.5.1 presents the global pattern; Section 4.5.2 break down the results by participant.

4.5.1 Global pattern

The pooled CV% (for all participants) is averaged over each stimulus condition, as shown in Figure 4.9. Therefore, each bar represents the mean CV% for one of the twelve (12) stimulus conditions (= 3 types of Context× 2 types of SentInto× 2 types of Tone).

Visual inspection shows that there is a significant increase in CV% from the phrase-medial position (MedUn) to the two phrase-final positions (FinUn and FinAc). For Tone2-bearing syllables, the increase in CV% from FinUn to FinAc is significant, regardless of SentInto. However, such an increase does not reach statistical significance for Tone4-bearing syllables.

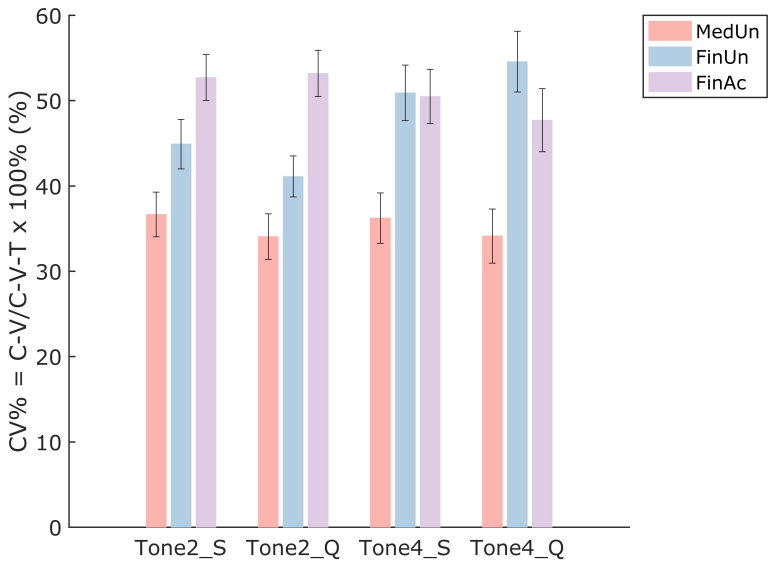


Figure 4.9: Bar plot showing mean and standard error of CV% of target syllables bearing Tone2 and Tone4, elicited in Statement (S) and Question (Q) in MedUn, FinUn, and FinAc. Each bar represents the mean CV% for one stimulus condition. Each error bar represents ±1 standard error of the mean CV%.

		MedUn	FinUn	FinAc
Tone2	Statement	36.7 (2.6)	44.9 (2.9)	52.7 (2.7)
	Question	34.1 (2.7)	41.1 (2.4)	53.2 (2.7)
Tone4	Statement	36.2 (3.0)	50.9 (3.3)	50.5 (3.2)
	Question	34.7 (3.2)	54.6 (3.6)	47.7 (3.7)

Table 4.3: Mean and standard error of the mean of CV% (in %) for all participants.

The CV% is further submitted to ANOVA with Context, SentInto, and Tone as the fixed factors, and participant as the random factor. Interactions between Context and SentInto, between Context and Tone, and between SentInto and Tone are also included.

As can be seen in Table 4.4, there is a significant effect of Context on CV% ($F(2, 1515) = 96.82, p < 0.0001$), suggesting that the C-V-T coordinative patterns are significantly different among three Context. The effect of Toneon CV% is also significant ($F(1, 1515) = 9.87, p = 0.0017$). This has in large part to do

with the significant interaction between Context and Tone (Context×Tone , $F(2, 1515) = 19.54, p < 0.0001$). The main effect of SentInto does not have a significant effect on CV%, nor does the interaction between Context and SentInto (Context×SentInto) or between Tone and SentInto (Tone×SentInto).

Source	F	dF1	dF2	p-value
Context	96.82	2	1515	< 0.0001***
SentInto	1.89	1	1515	0.1696
Tone	9.87	1	1515	0.0017**
Context×SentInto	2.14	2	1515	0.1186
Context×Tone	19.54	2	1515	< 0.0001***
Tone×SentInto	0.01126	1	1515	0.9155

Table 4.4: ANOVA on the CV% for all participants. The main effects of Context and Tone, and the interaction effect between Context and Tone, reach statistical significance ($p < 0.05$).

The CV% is analyzed in a linear mixed effect regression model with Context, SentInto, and Toneas the fixed terms, and Participant as the random term (same as ANOVA). The advantage of a linear mixed effect model over ANOVA is that the former quantifies the effect of each influencing factor. The mixed effect model analyzed here is as follows:

$$\text{CV\%} \sim \text{Context} + \text{SentInto} + \text{Tone} + \text{Context}\times\text{SentInto} + \text{Context}\times\text{Tone} + \text{SentInto}\times\text{Tone} + (1 \mid \text{Participant}).$$

	Coef.	t	d.f.	p-value
Context-FinUn	0.047	2.61	1515	0.009**

Context-FinAc	0.149	8.74	1515	< 0.0001***
SentInto-Question	-0.025	-1.53	1515	0.1263
Tone-Tone4	0.017	1.01	1515	0.3140
Context-FinUn×Tone-Tone4	0.083	4.03	1515	0.0001***
Context-FinAc×Tone-Tone4	-0.046	-2.25	1515	0.0243*
Context-FinUn×SentInto-Question	0.040	1.95	1515	0.0509·
Context-FinAc×SentInto-Question	0.007	0.37	1515	0.7119
Tone-Tone4×SentInto-Question	-0.002	-0.11	1515	0.9155

Table 4.5: Linear mixed effects regression on CV%. Only fixed effects are shown.

The results are shown in Table 4.5. Note that the reference level for Context is MedUn, i.e., the CV% for MedUn is the baseline CV%. Similarly, the reference level for SentInto is Statement, and for Tone is Tone2.

The coefficients for FinUn and FinAc are both significant, suggesting that CV% for FinUn is 4.7% higher than for MedUn, and CV% for FinAc is 14.9% higher than for MedUn. This is in line with the ANOVA result that Context has an significant effect on CV%.

The coefficients for Tone4 is not significant ($t(1515) = 1.01, p > 0.05$), which contrasts with the ANOVA result that the effect of Tone on CV% is significant. This discrepancy can be explained by the interaction between Context and Tone. Specifically, the interaction term FinUn×Tone4 is significant ($t(1515) = 4.03, p = 0.0001$), and has a coefficient of 8.3%, which means that setting aside the main effects of Context and Tone, compared to Tone2-bearing syllables in MedUn, the CV% for Tone4-bearing syllables in FinUn is 8.3% higher. Combining that the coefficient of FinUn—4.7%, the CV% for Tone-4 bearing syllables in FinUn

is about 13% higher than for Tone2-bearing syllables in MedUn. Similarly, the interaction term FinAc×Tone4 is also significant ($t(1515) = -2.25, p = 0.0243$), and has a coefficient of -4.6%, which means that setting aside the main effects of Context and Tone, compared to Tone2-bearing syllables in MedUn, the CV% for Tone4-bearing syllables in FinAc is 4.6% lower. Combining the coefficient of FinAc—14.9%, the CV% for Tone4-bearing syllables in FinAc is still 10.3% higher than Tone2-bearing syllables in MedUn.

Note that the coefficients of Question, and interaction terms involving Tone4, are not significant.[2]

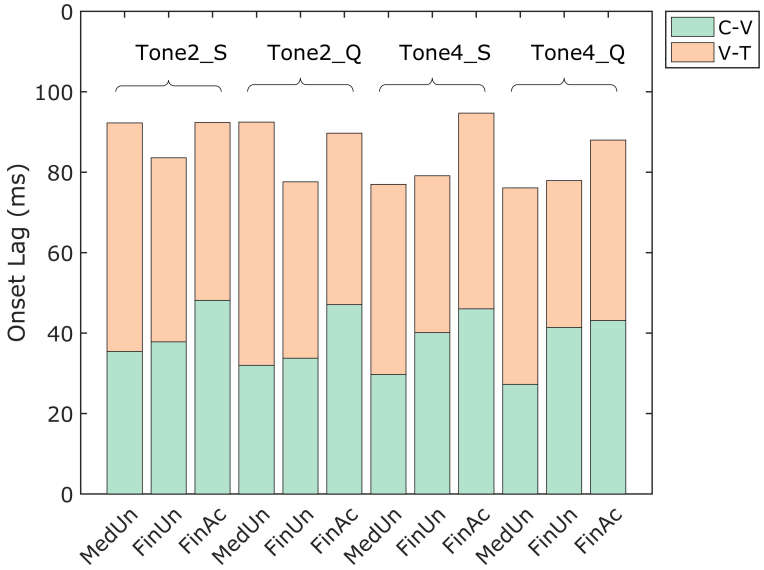


Figure 4.10: Bar plot showing the mean C-V onset lag (green) and the V-T onset lag (orange) for each stimulus condition for all participants.

Figure 4.10 offers a closer look at the two parts of CV%—the C-V onset lag and the V-T onset lag. In the stacked bar plot, the CV% is grouped by Context, in the order of MedUn, FinUn, and FinAc. Within each group, the CV% is further assorted by Tone and SentInto: Tone2×Statement, Tone2×Question,

[2]The interaction term FinUn×Question is marginally significant ($t(1515) = 1.95, p = 0.051$).

Tone4×Statement, and Tone4×Question. Each stacked bar consists of two parts: the C-V onset lag (green) and the V-T onset lag (orange). The mean and the standard error of the mean onset lag between the C and V gestures, between the V and T gestures, and between the C and T gestures, are shown in Table 4.6.

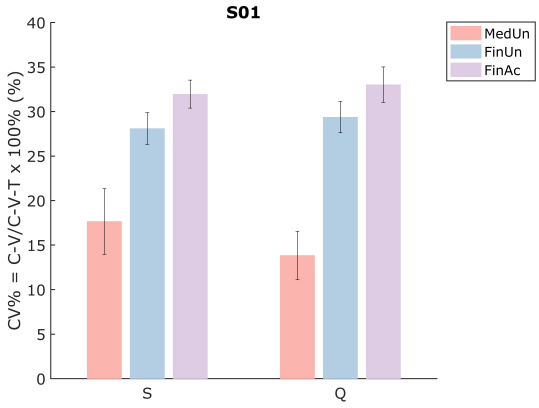
In Tone2-bearing syllables, the C-V onset lag in FinAc is much higher than in FinUn, which approximately equals to in MedUn. However, in Tone4-bearing syllables, the C-V onset lag in FinAc is higher than in FinUn, which is in turn higher than in MedUn. As for SentInto, the C-V onset lag does not differ much between Statement and Question. Turning to the V-T onset lag, in Tone2-bearing syllables, MedUn is higher than both FinUn and FinAc. However, in Tone4-bearing syllables, the V-T onset lag in MedUn is much higher than in FinUn, but not in FinAc. Also, the V-T onset lag does not differ much between Statement and Question in Tone4-bearing syllables.

			MedUn	FinUn	FinAc
C-V	Tone2	Statement	35.4 (2.7)	37.8 (2.5)	48.2 (2.3)
		Question	32.0 (2.6)	33.8 (2.1)	47.1 (2.2)
	Tone4	Statement	29.7 (2.7)	40.1 (2.7)	46.0 (3.1)
		Question	27.2 (2.7)	41.4 (2.9)	43.1 (3.0)
V-T	Tone2	Statement	56.8 (2.5)	45.8 (2.5)	44.2 (2.5)
		Question	60.5 (2.7)	43.9 (2.7)	42.6 (2.4)
	Tone4	Statement	47.3 (2.3)	39.0 (2.7)	48.6 (3.1)
		Question	48.9 (2.5)	36.5 (2.9)	44.9 (3.5)
C-T	Tone2	Statement	92.3 (1.6)	83.6 (1.4)	92.4 (1.6)
		Question	92.5 (1.7)	77.6 (2.2)	89.7 (1.5)
	Tone4	Statement	77.0 (1.3)	79.1 (1.3)	94.7 (2.1)
		Question	76.1 (1.1)	77.9 (2.2)	88.0 (2.2)

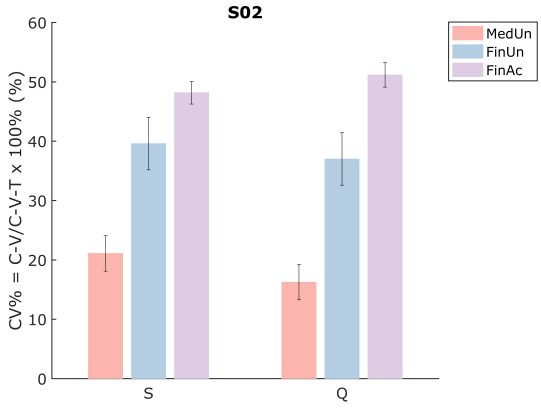
Table 4.6: Mean and standard error of the mean of the C-V, V-T and C-T onset lag (in ms) for each stimulus condition for all participants.

4.5.2 Individual Patterns

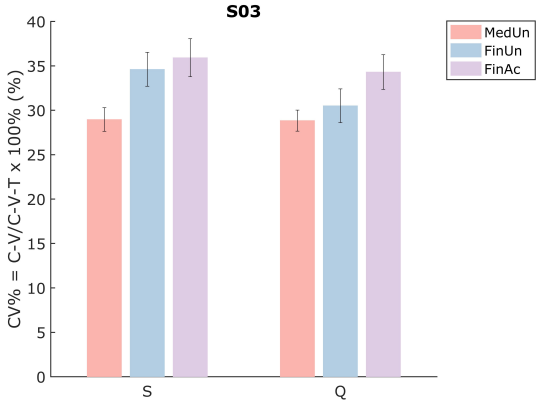
It has been established so far that Tone2- and Tone4-bearing syllables differ in CV%. In this section, individual patterns of CV% are broken down along the line of Tone, i.e., Tone2 and Tone4.



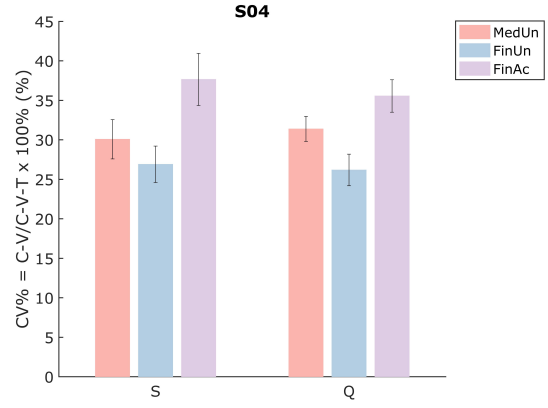
(a) S01



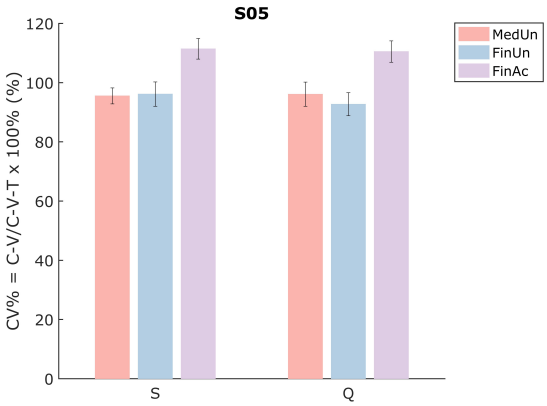
(b) S02



(c) S03



(d) S04

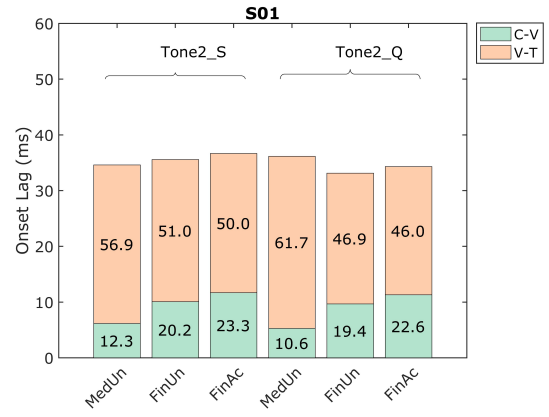


(e) S05

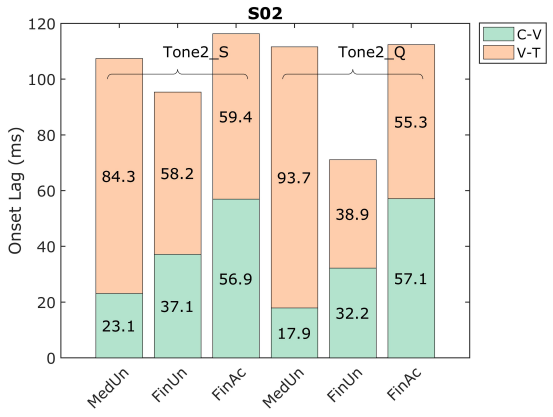
Figure 4.11: Bar plot showing the mean and standard error of the mean CV% for Tone2-bearing syllables for each participant. "S" represents Statement and "Q" represents Question.

Tone2 Figure 4.11 shows the CV% for Tone2-bearing syllables for each participant. For S01 and and S02, the CV% increases from MedUn to FinUn, and further to FinAc, for both Statement and Question. Such a pattern holds for S03, but to a lesser extent: the CV% differences between MedUn and FinUn, and between FinUn and FinAc , do not reach statistical significance, judging from the error bars. Similarly, for S04, the CV% difference between MedUn and FinUn is also negligible. However, for both S03 and S04, there is still a clear increase in CV% from MedUn to FinAc, regardless of SentInto. For S05, a similar pattern emerges despite the fact that the CV% range increases dramatically: the mean CV% comes between 10% and 40% for S01-S04, while it is around 100% for S05.

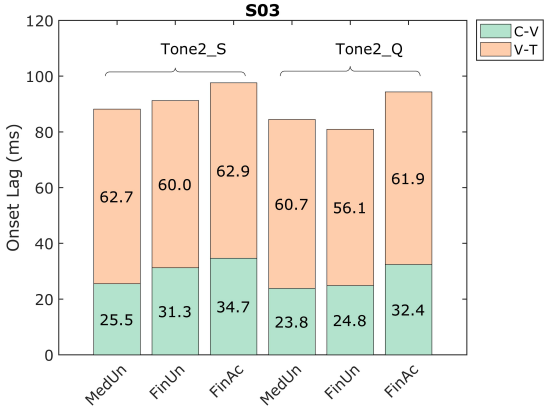
The CV% is further broken down to its two components: the C-V onset lag and the V-T onset lag, as shown in Figure 4.12. The values of the mean C-V and V-T onset lags are also marked in the respective bars.



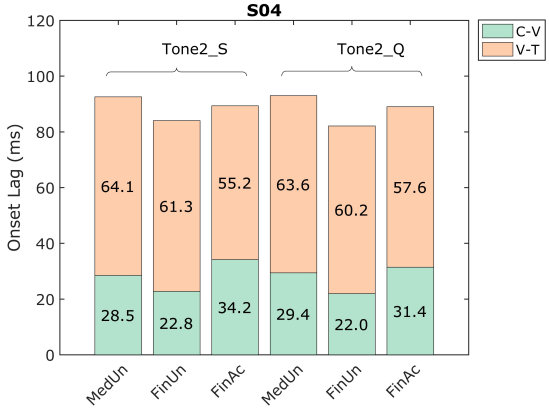
(a) S01



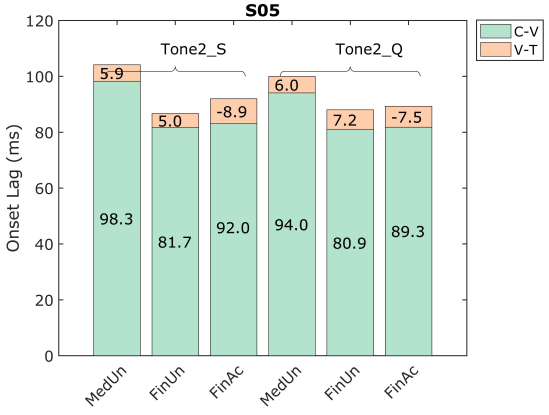
(b) S02



(c) S03



(d) S04



(e) S05

Figure 4.12: Bar plot showing the mean CV onset lag (green) and the V-T onset lag (orange) for Tone2-bearing syllables for each participant.

The global pattern for Tone2-bearing syllables identified in Figure 4.10 generally holds for Participants S01 and S02. For these two participants, the C-V onset lag enjoys a rise from MedUn to FinUn, and from FinUn to FinAc, regardless of SentInto, while the V-T onset lag drops from MedUn to FinUn, and from FinUn to FinAc, regardless of SentInto. For S03, the C-V onset lag increases from MedUn to FinUn, and from FinUn to FinAc, while the V-T onset lag remains unchanged. For S04, the differences in the C-V and V-T onset lags do not differ between MedUn and FinUn . However, the V-T onset lag decreases from MedUn to FinAc, while the C-V onset lag remains unchanged. For S05, the changes in both the C-V and V-T onset lags resemble S04. However, the V-T onset lag becomes negative in FinAc, suggesting that the T gesture is initiated before the V gesture.

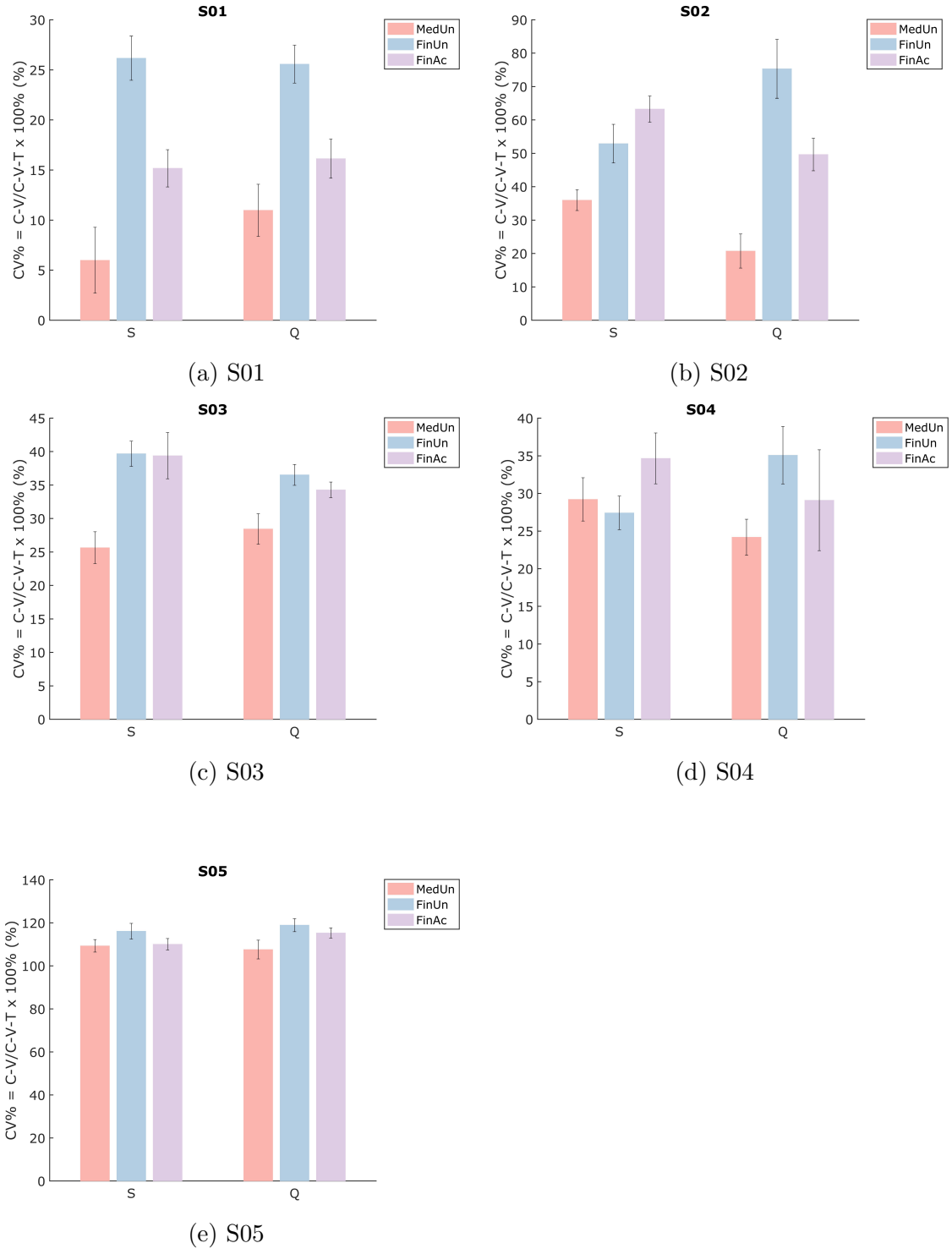
The above observations are further summarized in Table 4.7. The baseline condition is the CV% in Tone2-bearing syllables elicited in Statement in MedUn, as indicated by "0". According to this table, the most consistent pattern in Tone2-bearing syllables is that the CV% increases from MedUn to FinAc, which can be attributed to the increase in the C-V onset lag, or the decrease in the V-T onset lag, or both. The CV% change from MedUn to FinUn does not enjoy a consistent pattern. However, a plurality pattern that consists of S01 and S02 is reminiscent of the CV% change from MedUn to FinAc, i.e., the C-V onset lag increases and the V-T onset lag decreases.

		MedUn		FinUn		FinAc	
		C-V	V-T	C-V	V-T	C-V	V-T
S01	Statement	0	0	↑	↓	↑	↓
	Question	-	-	↑	↓	↑	↓
S02	Statement	0	0	↑	↓	↑	↓
	Question	-	-	↑	↓	↑	↓

S03	Statement	0	0	↑	-	↑	-
	Question	-	-	↑	-	↑	-
S04	Statement	0	0	-	-	-	↓
	Question	-	-	-	-	-	↓
S05	Statement	0	0	↓	-	-	↓
	Question	-	-	↓	-	-	↓

Table 4.7: Direction of change in C-V onset lag and V-T onset lag for Tone2-bearing syllables for each participant. Each cell represents one stimulus condition. Upward arrows "↑" indicate an increase from the baseline condition (indicated by "0"), hyphens "-" indicate no change, and downward arrows "↓" indicate a decrease.

Tone4 Figure 4.13 shows the CV% for Tone4-bearing syllables for each participant. There is a great deal of variation across participants. For S01, the CV% increases from MedUn to FinUn, and decreases from FinUn to FinAc, regardless of SentInto. However, the CV% in FinAc is still higher than in MedUn. For S02, the CV% increases from MedUn to FinUn, but only decreases from FinUn to FinAc for Question. For S03, the CV% increases from MedUn to FinUn, and remains unchanged from FinUn to FinAc, for both Statement and Question. For S04, the CV% in Statement remains unchanged from MedUn to FinUn, and increases from FinUn to FinAc, while the CV% in Question only increases from MedUn to FinUn. For S05, the CV% does not differ across the three levels of Context.



(a) S01

(b) S02

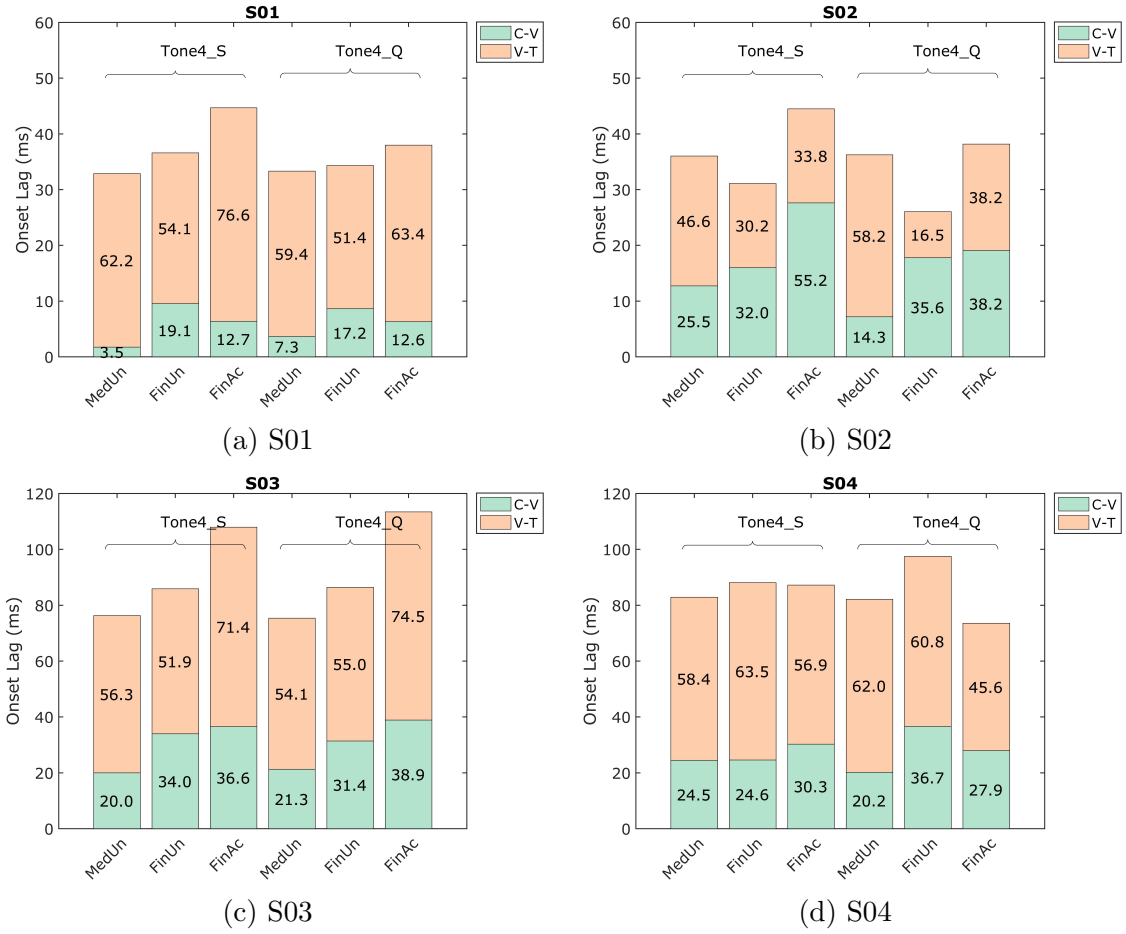
(c) S03

(d) S04

(e) S05

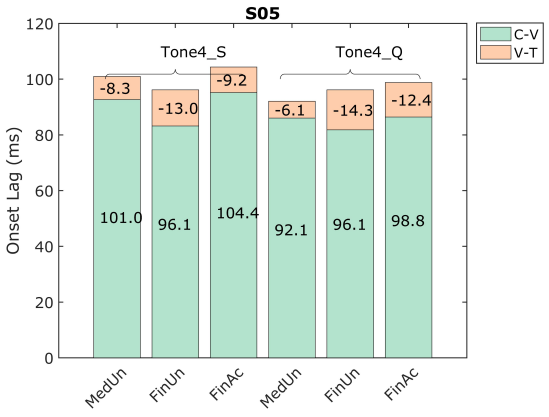
Figure 4.13: Bar plot showing the mean and standard error of the mean CV% for Tone4-bearing syllables for each participant. "S" represents Statement and "Q" represents Question.

Similar to before, the CV% is further broken down to its two components: the C-V onset lag and the V-T onset lag, as shown in Figure 4.14. The values of the mean C-V and V-T onset lags are marked in the respective bars. Despite the speaker-specific variation, the increase in CV% can mostly be attributed to the increase in the C-V onset lag.



(a) S01 (b) S02

(c) S03 (d) S04



(e) S05

Figure 4.14: Bar plot showing the mean CV onset lag (green) and the V-T onset lag (orange) for Tone4-bearing syllables for each participant.

The above observations are summarized in Table 4.8. The baseline condition is the CV% in Tone4-bearing syllables elicited in Statement in MedUn, as indicated by "0". The most consistent pattern is the increase in the C-V onset lag from the baseline condition MedUn, which is seen in all participants except for S05. For FinUn, the V-T onset lag remains unchanged or decreases from the baseline. Consequently, the CV% increases significantly from MedUn to FinUn for these participants (barring the Statement elicitations for S04). For FinAc, the V-T onset lag increases for S01 and S03, which mitigates the increase in the C-V onset lag, contributing the decrease in the CV% from FinUn to FinAcfor these two participants. However, the CV% increase from MedUn to FinAc is still significant. The sporadic changes in the onset lags from Statement to Question can be seen as anomalies.

		MedUn		FinUn		FinAc	
		C-V	V-T	C-V	V-T	C-V	V-T
S01	Statement	0	0	↑	-	↑	↑
	Question	-	-	↑	-	↑	↑

S02	Statement	0	0	↑	↓	↑	↓
	Question	↓	-	↑	↓	↑	-
S03	Statement	0	0	↑	-	↑	↑
	Question	-	-	↑	-	↑	↑
S04	Statement	0	0	-	-	-	-
	Question	-	-	↑	-	-	↓
S05	Statement	0	0	-	-	-	-
	Question	↓	-	↓	-	↓	-

Table 4.8: Direction of change in C-V onset lag and V-T onset lag for Tone4-bearing syllables for each participant. Each cell represents one stimulus condition. Upward arrows '↑' indicate an increase from the baseline condition (indicated by 0), hyphens "-" indicate no change, and downward arrows "↓" indicate a decrease.

To sum up, 1) there is a significant effect of Context on CV%: the CV% in MedUn is lower than that in FinUn and FinAc; 2) the interaction between Context and Tone is significant: globally, for Tone2-bearing syllables, the CV% increases from FinUn to FinAc, whereas for Tone4-bearing syllables, the CV% difference between FinUn and FinAc does not reach statistical significance; 3) the increase in CV% can be attributed to the increase in the C-V onset lag, or the decrease in the V-T onset lag, or both; 4) for two participants, there is a significant CV% decrease in Tone4-bearing syllables from FinUn to FinAc (which is still higher than in MedUn), which can be attributed to the increase in the V-T onset lag; 5) there is no significant main effect of SentInto on CV%.

4.6 Discussion

The finding that Context has a significant effect on CV% supports Hypothesis H2, i.e., the unification model. Intonational tone (such as focus-introduced pitch accent and boundary tone) operate locally, and can influence the gestural coordination of lexical tone gestures and oral articulatory gestures. The CV% varies with Context because the C-V onset lag and the V-T onset lag are differentially affected by Context. It is argued that the changes in the relative timing of gestural activations can be attributed to the changes in the relative phasing, which in turn arise out of the changes in the relative coupling strength between oral articulatory gestures and T gestures. Moreover, the finding that the interaction between Context and Tone also affects the CV% suggests that there is a lexical-tone-specific effect of focus on the gestural coordination of C-V-T. This further supports the notion that intonational tone can influence the intra-syllabic coordination of the C, V, and the lexical tone gestures by altering the gestural coupling relations depending on the tonal make-up of the lexical tones.

4.6.1 Boundary Tone (BT) Gestures

The effect of boundary tones can be assessed by comparing the CV% in FinUn with that in MedUn: the target syllable occurs in the phrase-final position in FinUn, while it occurs in the phrase-medial position in MedUn; in both Context, the target syllables are not associated with prosodic focus.

The global pattern shows that the CV% increases from MedUn to FinUn, regardless of Tone. As shown in Figure 4.10, for both Tone2- and Tone4-bearing

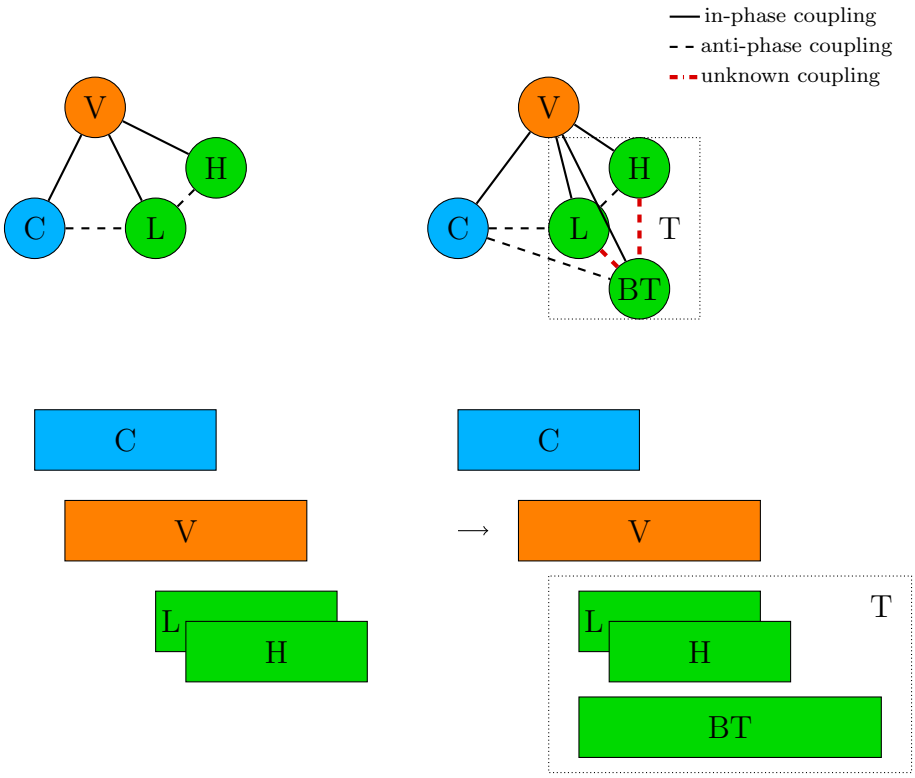
syllables, the global CV% increase can be attributed to the increase in the C-V onset lag, the decrease of the V-T onset lag, or both.[3] The difference in CV% between MedUn and FinUn can be accounted for a boundary tone gesture (BT). It is proposed that the presence of the BT gesture interferes with the intra-syllabic gestural coordination between the C, V and the lexical tone gestures, consequently altering the relative timing patterns of C-V-T from MedUn to FinUn. Specifically, the BT gesture either increases the C-V onset lag, or decreases the V-T onset lag, or both, effectively altering the CV% in the target syllable.

This is not the first time that BT gestures have been proposed. Katsika et al. (2014) proposed that BT gestures in Greek are anti-phase coordinated with V gestures to account for the result that BT gestures are initiated concurrently with the V gesture's target. However, they found that the C-V onset lag does not differ between the final syllable of a de-accented phrase-final word and a phrase-medial word. Therefore, they argued that BT gestures do not alter the intra-syllabic C-V coordination because they are associated with phrase-level tones. The behavior of BT gestures is similar to that of pitch accent gestures in Catalan and German, which are phrase tone gestures, in that pitch accent gestures also do not interfere with the intra-syllabic C-V coordination, and therefore do not cause the c-center effect Mücke et al. (2012).

In the current study, the presence of BT gestures obviously alters the intra-

[3]However, the individual behaviors illustrated in Figure 4.12 and Table 4.7 suggest that the S05 is different from the other participants. For instance, while the CV% is lower than 100% across the stimulus conditions for S01-S04, it is higher than 100% for S05. This suggests that the V gesture is not initiated between the C and T gestures, but after the T gesture. This leads to some unusual patterns in S05's data. For instance, the C-V onset lag increases or remains unchanged from MedUn to FinUn for S01-S04,the C-V onset lag decreases significantly from MedUn to FinUn for S05. Because this decrease for S05 is of much greater magnitude than that of the increases for the other participants, the overall C-V onset lag does not differ between MedUn and FinUn for Tone2-bearing syllables globally. S05's data will be excluded in the subsequent analyses involving global patterns.

syllabic gestural coordination of the C, V and lexical tone gestures, therefore requiring a different proposal than that by Katsika et al. (2014) and Mücke et al. (2012). Two gestural accounts are laid out here.



a) Tone2-bearing syllable in MedUn

b) Tone2-bearing syllable in FinUn

Figure 4.15: Coupling graphs (top) and gestural scores (bottom) of a Tone2-bearing syllable in MedUn (left) and FinUn (right). The black solid lines represent in-phase coupling relations; the black dashed lines represent anti-phase coupling relations; the red dash-dotted lines represent unknown gestural coordinations.

In the first account, it is proposed that BT gestures in Mandarin behave like lexical tone gestures. Recall that Gao (2008) claimed that lexical tone gestures behave like onset C gestures, and form consonant clusters that give rise to the c-center effect. More importantly, lexical tone gestures in Mandarin alter the intra-syllabic

C-V gestural coordination, which would otherwise result in the synchronization of the gestural initiations of the C and V gestures due to their in-phase coupling relation. Because the coordination of BT gestures resemble that of lexical tone gestures, the BT gestures are in-phase coupled to the V gesture, and anti-phase coupled to the C gesture (Figure 4.15). This renders a stronger coordination between the V gesture and the T gestures, which encompass both the lexical tone gestures and the BT gesture. The stronger V-T coordination attracts the V gesture towards the T gestures, and away from the C gesture, accounting for the higher CV% in FinUn. More importantly, as argued earlier, this corroborates the claim that boundary tones interact with lexical tones locally, and interfere with the intra-syllabic coordination between the C, V and the lexical tone gestures.[4]

It also can be argued that the BT gesture is not coupled to the articulatory gestures or the lexical tone gestures, but affects the intra-syllabic gestural coordination through some extra mechanism. The interference of the BT gesture increases the coupling strength between the V gesture and the lexical tone gestures, effectively leading to the increase in the C-V onset lag, the decrease in the V-T onset lag, and eventually the increase in the CV%, which is consistent with the first account.

Another important finding is that SentInto does not affect the CV%: the C-V-T coordinative pattern does not differ between Statement and Question, regardless of Context. The non-significant effect of SentInto indicates in FinUn, where the target syllable occurs phrase-finally, either a boundary tone does not occur or occurs but exerts the same effect on the gestural coordination. Because the CV% differs significantly between MedUn and FinUn, and phrase boundary is associated with the latter but not the former, it is argued that boundary tones

[4]Note that the coordinations between the BT gesture and the lexical tone gestures are not stipulated due to the lack of knowledge in the current field. The unknown gestural coordinations are indicated by red dash-dotted lines in Figure 4.15.

are present in FinUn, and H% and L% are respectively triggered by Question and Statement. Moreover, there is no significant Tone-specific SentInto effect, i.e., no significant interaction between Tone and SentInto. Taking both into consideration, a unified BT gesture is thus proposed. This is in line with the observation that there is no systematic effect of the boundary tone type (L%, H%, and !H%) on the gestural coordination of BT gestures in Greek (Katsika et al., 2014).

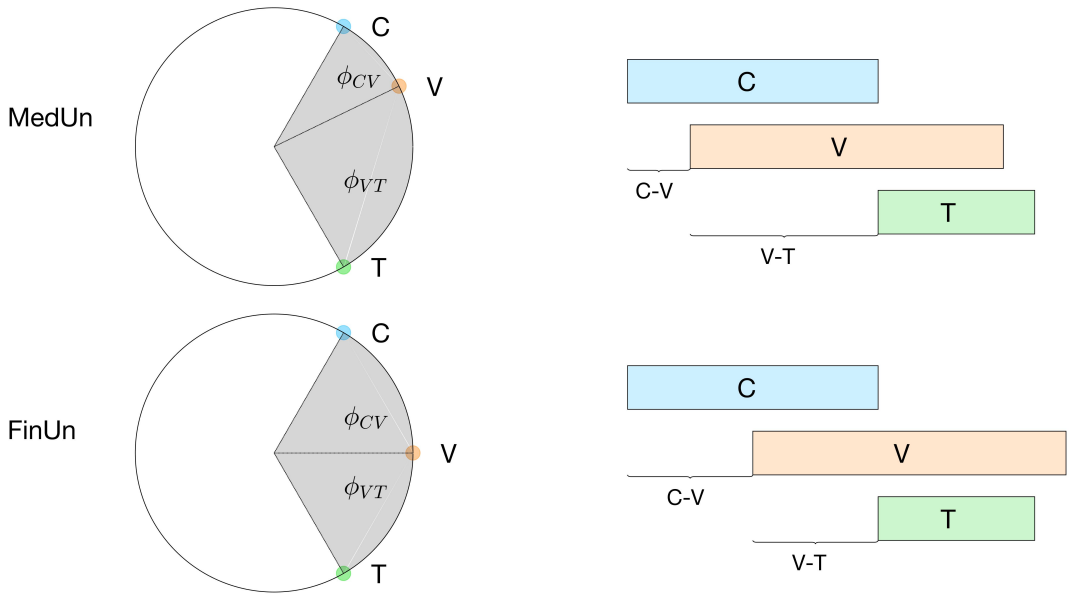


Figure 4.16: Coupled oscillators (left) and gestural scores (right) in MedUn (top) and FinUn (bottom). In FinUn, the presence of the BT gesture increases the coupling strength between the V gesture and the T gestures, drawing the V gesture closer to the T gestures, thereby increasing the CV% in FinUn.

Figure 4.16 summarizes the difference in gestural coordination between MedUn and FinUn. The unit circles on the left illustrate the relative phasing of the coupled oscillators, i.e., the C, V, and T gestures, in both MedUn and FinUn. The gestural scores on the right further demonstrate the same coordinative patterns on the temporal scale. Specifically, in MedUn, the C-V coupling strength is stronger than the V-T coupling strength ($C_{CV} > C_{VT}$). Therefore, the relative phasing of the C

and V gestures is smaller than that of the V and T gestures ($\phi_{CV} < \phi_{VT}$), which in turn causes the V gesture to be initiated closer to the C onset than the T onset. Note in MedUn, the T gesture includes only the lexical tone gestures. In FinUn, the presence of boundary tones triggers the BT gestures. In both aforementioned accounts, the BT gesture renders a stronger coupling strength between the V and T gestures ($C_{CV} \approx C_{VT}$), an increased C-V relative phase ($\phi_{CV} \approx \phi_{VT}$), which eventually leads to an increased CV% in FinUn.[5]

4.6.2 Accent (μ) Gesture

The effect of prosodic focus can be assessed by comparing the CV% in FinAc with that in FinUn: the target syllable is accented in FinAc, whereas it is un-accented in FinUn(the contrastive focus is placed on the phrase-initial syllable); in both Context, the target syllables occur in the phrase-final position.

The global pattern shows that the CV% increases from FinUn to FinAc for Tone2-bearing syllables, while remains unchanged for Tone4-bearing syllables. For Tone2-bearing syllables, the CV% increases from FinUn to FinAc can mostly attributed to the increase in the C-V onset lag for S01-S04. At the same time, the V-T onset lag remains stable. This is case for both the global pattern and most of the individual participants.[6] For Tone4-bearing syllables, both the C-V onset lag and the V-T onset lag increase from FinUn to FinAc, resulting in no significant

[5]The T gesture includes the lexical tone gestures and the BT gesture in the first account, while it only includes the lexical tone gestures in the second account. Nonetheless, both accounts predict the same changes in the CV%, in line with the empirical finding.

[6]For S02, the V-T onset increases from FINUN to FINAC for Tone2-bearing syllables in Question. However, the increase is compensated by an increase of a larger magnitude in the C-V onset lag, resulting an increase in the CV% from FinUn to FinAc. Furthermore, judging from the global pattern, this particular data point, i.e., the V-T onset lag for Tone2-bearing syllables in Question in FinUn might be a outlier measure.

change in CV% globally. For some individual participants, the relatively large increase in the V-T onset lag leads to the decrease in the CV%, as shown in Figure 4.13.

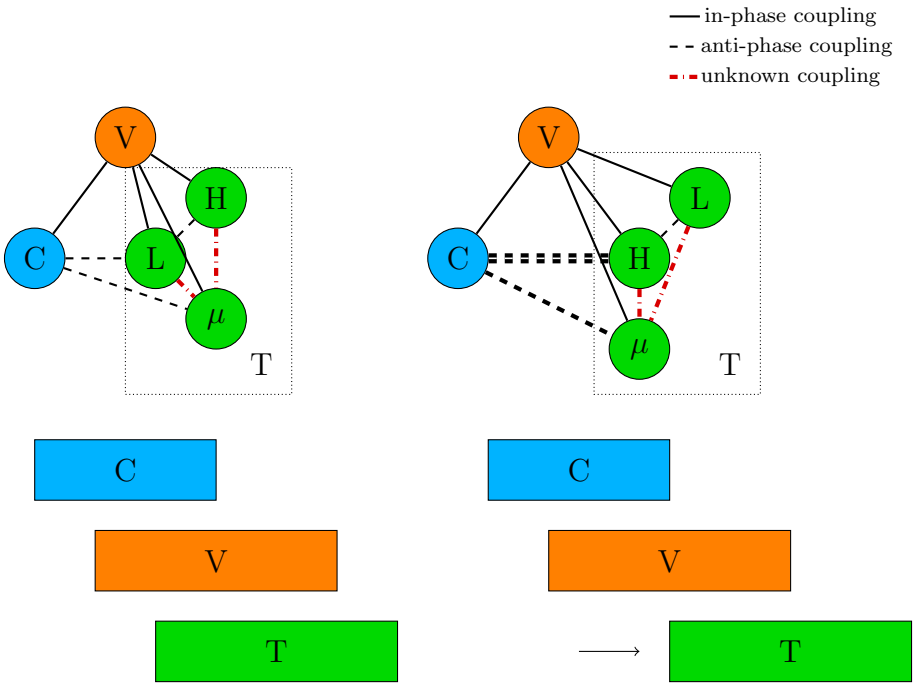
The differences in CV% between FinUn and FinAc and the Tone-specific effects of Context suggest that prosodic focus, like boundary tones, also influences the gestural coordination between the oral articulatory gestures and lexical tone gestures. Therefore, an accent (μ) gesture, triggered by the focus-introduced pitch accent in FinAc, is proposed to account for the changes in relative timing of gestural activations of C-V-T. Similar to the BT gesture, two gestural accounts can be laid out.

In the first account, the μ gesture is coordinated with the V gesture via in-phase coupling. This further strengthens the V-T coupling, therefore attracts the V gesture closer to the T gesture, which now includes the lexical tone gesture, the BT gesture, and the μ gesture.

In this account, the μ gesture and the BT gesture are treated as two closely related intonational gestures with different intonational functions. Moreover, both intonational gestures behave in a similar fashion to the lexical tone gestures. The latter point is line with the premise of the unification model: lexical tones and intonational tones are the same type of phonological entity, and thus should be modeled as the same type of tone gesture. From the standpoint of gestural control, both the lexical tone gestures and intonational tone gestures involve the same type of f_0 control that utilizes a certain group of laryngeal muscles. The control of these muscles is temporally coordinated with the control of oral articulators primarily associated with segments in an integrated fashion.

A unifying account of the μ gesture and the BT gesture can capture the CV% increases for Tone2-bearing syllables from MedUn to FinUn, and from FinUn to FinAc. However, additional explanation is needed to account for the changes for Tone4-bearing syllables: the CV% increases from MedUn to FinUn, whereas it remains approximately unchanged globally—and even decreases for some participants—from FinUn to FinAc, which can be attributed to the increase in the V-T onset lag.

It is speculated that the μ gesture, triggered by focus-introduced pitch accent in FinAc, affects the gestural coordination differently in Tone2-bearing syllables and in Tone4-bearing syllables. As illustrated in Figure 4.17, in addition to being in-phase coupled to the V gesture, the μ gesture is also anti-phase coupled to the C gesture. The presence of the μ gesture disproportionally affects the H gesture in its coordination with the C gesture, leading to a stronger anti-phase coupling relation between the C gesture and the H gestures in Tone4-bearing syllables than in Tone2-bearing syllables. It is possible that the anti-phase coupling between the C gesture and the μ gesture also increases. The stronger coupling strength between the C gesture and the T gesture in Tone4-bearing syllables is indicated by the double dashed lines in Figure 4.17(b). This results in an increase in the C-T onset lag, which can capture the observation that the V-T onset lag is larger and the CV% is lower in Tone4-bearing syllables than in Tone2-bearing syllables.



a) Tone2-bearing syllable in FinAc

b) Tone4-bearing syllable in FinAc

Figure 4.17: Coupling graphs (top) and gestural scores (bottom) of a Tone2-bearing syllable in FinAc (left) and a Tone4-bearing syllable in FinAc (right). The black solid lines represent in-phase coupling relations; the black dashed lines represent anti-phase coupling relations; the red dash-dotted lines represent unknown gestural coordinations.

In the second account, the μ gesture is not coupled to the articulatory gestures or the lexical tone gestures. It influences the intra-syllabic gestural coordination through some unknown mechanism. Specifically, the presence of the μ gesture further increases the in-phase coupling strength between the V gesture and the lexical tone gestures, effectively resulting in the increase in the CV% from FinUn to FinAc for Tone2-bearing syllables. The presence of the μ gesture disproportionally affects the H gesture in its coordination with the C gesture for Tone4-bearing syllables, leading to an increase of the anti-phase coupling strength. As a result,

the C-T onset lag increases, and the CV% remains unchanged or even decreases.

The Tone-specific effect of the μ gesture may arise out of the differences between the H and L lexical tone gestures. As Gao (2008) discussed in her dissertation, the differences between the two types of tone gestures are associated with various aspects of increasing and decrease pitch, such as the muscles involved in production, and the length of time it takes to achieve the same displacement in f_0. In both accounts, the disproportional effect of the μ gesture on the coordination of the H gesture in Tone4-bearing syllables might have to do with the fact that prosodic focus affects the realization of a high tone and a low tone differently. Chen and Gussenhoven (2008) found that while there was a significant increase in maximum f_0 from the NoEmphasis condition to the Emphasis condition, there was no significant difference in minimum f_0 between the two discourse conditions, which further led to a larger f_0 range expansion of the falling tone (Tone4) than the rising tone (Tone2). Thus, Tone4—the H tone in particular—reacts to prosodic focus in a more sensitive manner, and the anti-phase coupling between the C and T gestures becomes stronger in the presence of the μ gesture. The interaction between Tone and Context is in keeping with the unification model in that the intra-syllabic gestural coordination between the C, V, and the lexical tone gestures reacts differently towards the intonational tone gestures depending on the tonal make-up of the lexical tones. This only can arise if intonation tones are locally produced and interact with lexical tones, rather than operating on a global level.

To sum up, the articulatory alignment of C-V-T is altered by the presence of both boundary tones and prosodic focus. It is proposed that like the lexical tone gestures, the intonational tone gestures associated with boundary tones and focus-introduced pitch accent—the BT gestures and the μ gesture, respectively, can be

modeled as f_0 gestures. The rationale is that both intonational and lexical tone gestures use the same group of laryngeal muscles to exert f_0 control. The intonational tone gestures can influence the intra-syllabic gestural coordination by operating at a local level and interacting with the C, V, and lexical tone gestures. Furthermore, they may also exhibit lexical-tone-specific effects. Therefore, the unification model is supported.

4.6.3 Inadequacy of Overlay Model

Under Hypothesis H1, the overlay model predicts the gestural coordination of C-V-T will not be substantially altered in the presence of intonational tones such as focus and boundary tones. In this section, the changes that can arise under this model will be discussed.

First of all, the C-V onset lag will not vary with intonational tones under the overlay model. Intonation is often described as "post-lexical" in previous studies (Ladd, 2008). This indicates that the implementation of intonational tones does not affect the gestural coordination of the articulatory gestures and the lexical tone gestures, because articulatory timing is learned through coupling strengths among the gestures associated with segments and lexical tones, which are stored in the long-term memory, i.e., lexicon. This contrasts with the finding in the current experiment that the C-V onset lag increases from phrase-medial positions to phrase-final positions. To further rule out the possibility that the lengthening of the C-V onset lag arises out of the final lengthening, two more ANOVAs are conducted to assess the effects of Context, SentInto, and their interaction on the C-V onset lag as a proportion of the C gesture for Tone2-

and Tone4-bearing syllables, respectively.[7] Both ANOVAs show that CONTEXT has a significant effect on the increase in the proportional C-V onset lag: for Tone2-bearing syllables, $F(2, 784) = 38.06, p < 0.001$; for Tone4-bearing syllables, $F(2, 599) = 5.99, p < 0.01$. Therefore, the conclusion that the C-V onset lag increases still holds.

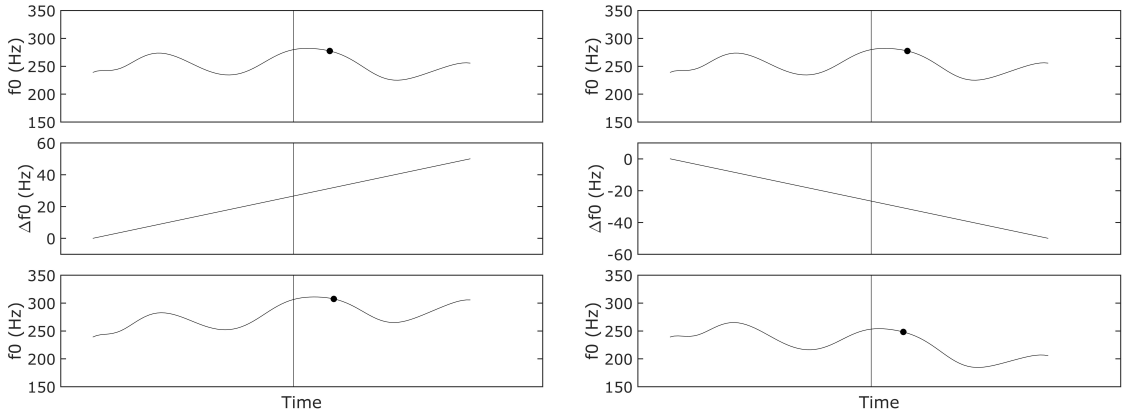
Secondly, the V-T onset lag may be subject to changes due to the way the gestural onset is measured in the current study. Under the assumptions of the overlay model, the overall f_0 contour is the result of imposing local f_0 perturbations (lexical tones) onto global f_0 "grid" (intonation, c.f. Gårding, 1983). When the global grid rises or falls, it shifts the location of the turning points associated with the T gestures.

Simulations are conducted to assess the impact of superimposing a global intonation grid on the measurement of gestural onsets of the T gestures. As illustrated in Figure 4.18, two shapes of intonation—rising and falling—are superimposed for MEDUN utterances with Tone2- or Tone4-bearing syllables. In each of the four subplots, the top panel shows the original contour—a MEDUN elicitation in STATEMENT. The middle panel shows the intonation grid: a rising grid represents a rising intonation, and a falling grid represents a falling intonation. Note that the rising intonation can be associated with FINUN, and the falling intonation can be associated with FINAC, according to the empirical f_0 contours. The bottom panel shows the aggregate f_0 contour after imposing the intonation grid. Gestural onsets for these syllables are then measured the same way as before, as indicated by the filled circles on the f_0 contours.

The magnitude of increase for the rising intonation is set at 10 Hz, 30Hz, and 50

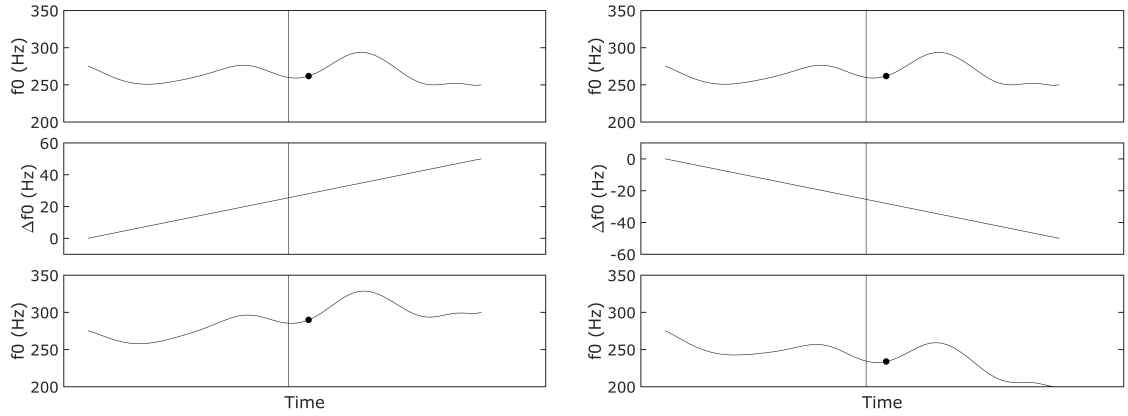
[7] Participant S05 is excluded.

Hz, based on empirical measurements. Similarly, the magnitude of decrease for the falling intonation is set at -50 Hz, -30 Hz, and -10 Hz. The aggregate f_0 contour is the result of the addition of the original contour and the global intonation. For instance, in the falling intonation with the decrease magnitude of -50 Hz, the starting f_0 in the aggregate contour is the same as that in the original contour. The ending f_0 in the aggregate contour is 50 Hz lower than that in the original contour. The f_0 in between is the sum of the f_0 in the original contour and the corresponding Δf_0 that falls on the linear decreasing line from 0 to -50 Hz. Therefore, a falling intonation with the decrease magnitude of -50 Hz is steeper than -30 Hz, and a rising intonation with the increase magnitude of 50 Hz is steeper than 30 Hz.



(a) Tone2 - Rising intonation

(b) Tone2 - Falling intonation



(c) Tone4 - Rising intonation

(d) Tone4 - Falling intonation

Figure 4.18: Illustration of the f_0 contours with a rising (left) and falling intonation (bottom) under the overlay model for Tone2- and Tone4-bearing syllables. Top panels show the original f_0 contours; middle panels show the intonation "grids" (c.f. Gårding, 1983); bottom panels show the aggregate f_0 contours. Vertical lines indicate the acoustic onsets of the target syllables. Filled circles indicate the gestural onsets of the L and H gestures for Tone2- and Tone4-bearing syllables, respectively.

The differences between the onsets of the T gestures in the aggregate f_0 contour and in the original f_0 contour are measured, as shown in Table 4.9. "↑" indicates an increase in the V-T onset lag, "↓" indicates an decrease in the V-T onset lag.

	Falling			Rising		
Tone	-50	-30	-10	10	30	50
Tone2	↓	↓	↓	↑	↑	↑
Tone4	↑	↑	↑	↓	↓	↓

Table 4.9: Simulations showing differences between the onsets of the T gestures in the aggregate f_0 contour and in the original f_0 contour. "↑" indicates increases in the V-T onset lag, whereas "↓" indicates decreases.

Under the overlay model, the V-T onset lag varies with intonation: for Tone2-bearing syllables, the lag increases with the rising intonation and decreases with the falling intonation; for Tone4-bearing syllables, the lag decreases with the rising intonation and increases with the falling intonation.

The simulation results contrast with the empirical findings. Firstly, the simulations cannot capture the finding that Tone has no effect on the intra-syllabic gestural timing in FinUn. In the simulations with falling intonation, which can be associated with FinUn, the V-T onset lag decreases for Tone2-bearing syllables and increases for Tone4-bearing syllables, while the empirical results show that the V-T onset lag decreases regardless of Tone.

Secondly, the simulations fail to account for the finding that in FinAc, the V-T onset lag can be extended for Tone4-bearing syllables but not for Tone2-bearing syllables. However, the V-T onset lag for Tone4-bearing syllables decreases in the simulations with rising intonation, which is most likely to be associated FinAc, where the f_0 range is often expanded and the f_0 maximum increases. Moreover, the V-T onset lag for Tone2-bearing syllables increases in the simulations with rising intonation, which again contrasts with empirical results.

To sum up, the overlay model cannot account for the differences in gestural timing across different stimulus conditions. Admittedly, under the assumptions of the overlay model, the V-T onset lag can be altered. However, the changes in the V-T onset lag come in the opposite direction from the empirical results. Moreover, the overlay model cannot capture the increases in the C-V onset lag.

CHAPTER 5

CONCLUSION

This dissertation asks two questions regarding Mandarin lexical tones: 1) How do native speakers control the alignment between lexical tones and segments? 2) How does sentence-level intonation interact with lexical tones in a tones language? These two questions are addressed respectively by Experiment 1 and 2, of which the results are summarized in Section 5.1 and Section 5.2. On the basis of the empirical findings, gestural accounts are proposed within the framework of AP.

5.1 Experiment 1

In Experiment 1, the participants, who are native speakers of Mandarin, were asked to imitate a series of bi-tonal sequences Tone2+Tone2, in which the relative timing and fundamental frequency of the turning points (TP) varied parametrically. There were steps on the stimulus TP relative timing continuum and four on the stimulus TP F_0 continuum. In Experiment 1A, the parametric variation occurred at the Low-to-High transition with the first tone-bearing syllable, while in Experiment 1B, it occurred at the High-to-Low transition across syllable boundaries.

It is found that speakers encode stimulus variation in a non-linear fashion, and that the parameter space is structured by perceptual categories. Moreover, there is a strong correlation between the discrimination performance and the magnitude of difference between imitations. Specifically, most imitations in Experiment 1A can be grouped into one large distributions while the imitations in Experiment 1B belong to two distinct distributions. Specifically, in Experiment 1B, the imitations for first three TimingStep-B-s can be separated from those for the last two

TimingStep-B-s. The imitation results are further backed up by the discrimination results: in Experiment 1A, there is no discrimination boundary on the stimulus TP relative timing continuum; in Experiment 1B, the discrimination boundary occurs between TimingStep-B-s 3 and 4.

Therefore, it is argued that there are categorical modes of gestural coordination between lexical tone gestures and oral articulatory gestures. Moreover, the relative timing patterns are governed via different modes of categorical coordination. In Experiment 1A, most tone-to-segment alignment patterns are governed by one mode of gestural coordination, corresponding to that of Tone2-bearing syllables. In Experiment 1B, the tone-to-segment alignment patterns are indicative of two distinct modes of gestural coordination, corresponding to that of Tone2- and Tone4-bearing syllables. Specifically, the imitations for the first three TimingStep-B-s resemble Tone2, while the imitations for the last two TimingStep-B-s resemble Tone4.

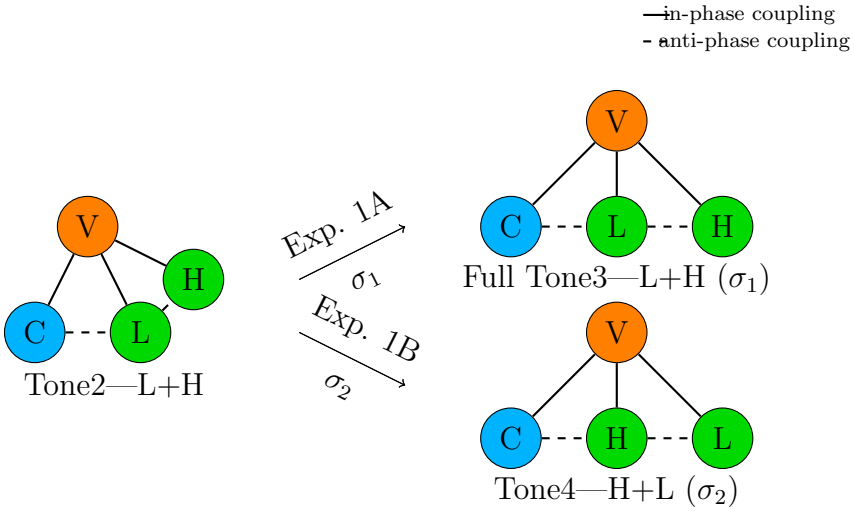


Figure 5.1: Schematic illustration of the changes in coupling modes in Experiment 1A (top right) and in Experiment 1B (bottom right). Note that only the tone-bearing syllables are shown, i.e., σ_1 in Experiment 1A and σ_2 in Experiment 1B.

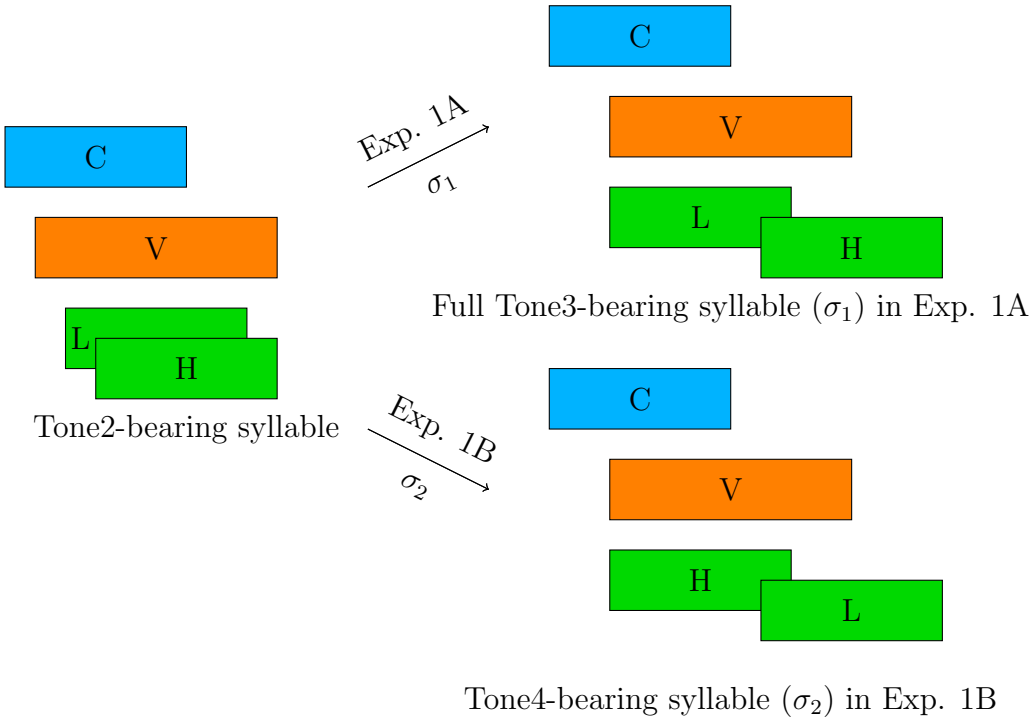


Figure 5.2: Schematic illustration of the changes in gestural scores in Experiment 1A (top right) and in Experiment 1B (bottom right). Note that only the tone-bearing syllables are shown, i.e., σ_1 in Experiment 1A and σ_2 in Experiment 1B.

There is one caveat about claiming that there is only one categorical mode of gestural coordination in Experiment 1A: when the TP occurs late in the tone-bearing syllable with low F0, some imitations appear to exhibit a secondary mode of tone-to-segment alignment resembling that of a Full Tone3-bearing syllable. This is also echoed by the increase in the discrimination performance when this subset of stimuli is involved. Therefore, the gestural accounts for the two experiments can be unified, as summarized in Figure 5.1 and Figure 5.2.

In Experiment 1A, the primary mode of gestural coordination is that of Tone2-bearing syllables. As the TPs occurs late with low F0 (e.g., TimingStep-A-s 5 and F0Step-A-s 1), some participants can adopt a secondary mode of gestural

coordination reminiscent of that of Full Tone3-bearing syllables. This is achieved by increasing the coupling strength between the C gesture and the H gesture. As a result, the H gesture is attracted away from the C gesture, rendering a late TP with low F_0. However, as stated earlier, this categorical shift in alignment patterns only occurs for a small subset of stimuli for some participants.

In Experiment 1B, the two categorical modes of gestural coordination are easier to discern: the participants adopt the gestural coordination of Tone2-bearing syllables for the first three TimingStep-B-s, and Tone4-bearing syllables for the last two TimingStep-B-s. The shift in the categorical modes is accomplished by interpreting the H gesture as spanning over both tone-bearing syllables, and adopting the gestural coordinative system of Tone4-bearing syllables.

Moreover, the categorical patterns of tone-to-segment alignment are subject to stimulus-induced changes in both experiments. Within each category of tone-to-segment alignment patterns, the TP relative timing in imitation is positively correlated with that in stimulus. It is proposed that the gradient variation in the stimulus-induced adjustments can be accounted for by the continuous changes in the coupling strength between lexical tone gestures and oral articulatory gestures.

The manipulation of coupling strength is essential for Articulatory Phonology to function as both a phonological model and a phonetic model. Changes in coupling strength lead to changes in relative phasing, which in turn lead to changes in relative timing of gestural activations (c.f. Figure 4.16). The manipulation of coupling strength thus gives rise to different patterns of phasing relationship between gestures. The differences can be either phonological (categorical) or phonetic (continuous).

For example, the changes in the tone-to-segment alignment in imitation for TimingStep-A-s 1-4 can be largely regarded as continuous. Specifically, for later TimingStep-A-s, increases in the L-H anti-phase coupling force lead to larger degree of overlap between the L and H gestures, attracting the H gesture away from the C gesture, resulting in greater RelTiming-A-i.[1] Despite the incremental changes in RelTiming-A-i, the differences in tone-to-segment alignment are largely continuous within one category—Tone2. When the gestural overlap is large enough, as in some imitations for TimingStep-A-s 5, the alignment patterns can be seen as forming a different category, resembling a different lexical tone category—Full Tone3.

Similarly, the changes in the tone-to-segment alignment in imitation for TimingStep-B-s 1-3 can be attributed to the increases in the C-L anti-phase coupling force, which attracts the L gesture away from the C gesture, resulting in greater RelTiming-B-i. The changes in RelTiming-B-i are deemed gradient because the magnitude of changes is much smaller than the magnitude of stimulus variation. For TimingStep-B-s 4-5, a different tone-to-segment alignment, resembling that of Tone4-bearing syllables, is adopted by most speakers to imitate the late stimulus TP timing. The increase in RelTiming-B-i from TimingStep-B-s 4 to 5 can be accounted for by the increase in L-H anti-phase coupling force.

This is not the first time that differences in alignment pattern are accounted for by manipulation of coupling force. Recall that Mandarin tone-bearing syllables (such as Tone1 and Half Tone3) exhibit c-center effect under the assumption that the C-V and V-T in-phase coupling relations are of the same coupling strength. The competition between the C-T anti-phase coupling relation and the C-V and V-T in-phase coupling relations renders the c-center effect (Gao, 2008). In addition,

[1]Note that the V-H in-phase coupling force also increases in the imitations for later TimingStep-A-s, resulting in shorter latency between the V and L gestures. However, this cannot be tested in the current experiment due to the lack of articulatory data.

it was observed that one alternative pattern to the c-center paradigm: the onset of the V gesture is aligned closer to the onset of the C gesture than to that of the T gesture in a Half Tone3-bearing syllable. It was argued that the C-V in-phase coupling force is stronger than the V-T in-phase coupling force, and thus the V-T coupling relation is suppressed. Therefore, the T gesture is attracted away from the V gesture, giving rise to the alternative pattern.

Proposing coupling relations with uneven strength brings in more degree of freedom to the coupling model (Gao, 2008). Specifically, it accounts for the sub-phonemic variation in addition to providing a pure phonological model. In this sense, Articulatory Phonology functions as a unified theory of phonetics and phonology, whereas Autosegmental Metrical theory is more of a phonological theory that requires additional mechanisms for phonetic implementation. This is in line with the view expressed in Browman and Goldstein (1989) that phonological and phonetic representations are essentially congruent (c.f. Clements, 1992).

In the current experiment, the phonological presentations correspond to the categorical modes of tone-to-segment alignment, functioning as attractors in the cognitive organization in which abstract linguistics categories are formed. The phonetic variation corresponds to the stimulus-induced gradient changes within in each lexical tone category. The manipulation of coupling force allows for an elegant explanation for both phonological categories and phonetic variation, which renders Articulatory Phonology a better integrated model of phonology and phonetics.

However, there is one caveat about manipulating coupling strength in Articulatory Phonology: it can lead to an infinite number of phasing relationships between gestures. Recall that in Chapter 2 the in-phase coupling relation is associated with a relative phase of 0° between two planning oscillators, while the anti-phase

coupling relation is associated with a relative phase of 180°. These two coupling relations are the most stable ones out of an infinite number of coupling relations. That is, a pair of planning oscillators (thus gestures) can be coupled with a relative phase of 10°, 20°, 100°, etc. Therefore, the in-phase coupling relation with a relative phase of 0° is, more accurately, one mode of attractive coupling relations, whereas the anti-phase coupling relation with a relative phase of 180° is one mode of repulsive coupling relations (Tilsen, 2016). There could potentially be an infinite number of attractive and repulsive coupling relations, which can be accounted for manipulating coupling strength between gestures.

The proliferation of gestural coupling relations could render Articulatory Phonology too rich a model as a theory of phonology. It is thus an important task to differentiate the phonological aspect from the phonetic aspect by manipulating coupling force. It is often assumed that different languages consistently select a limited number of coordinative structures qualitatively while there can exist some variation in coupling strength.

Steriade's (1990) discussion of Dorsey's law can shed light on this issue. Dorsey's Law turns CCV(C) syllables into CvCV(C), where v stands for a copy of V. For example, /par/ becomes /pa.ra/, where the inserted vowel matches the nucleus with the complex onset. It is also possible that the second consonant in a onset cluster is moved to syllable coda position. In this case, /par/ becomes /par/, where the rightward displacement of /r/ carries it all the way to a peripheral position.

Both instances of Dorsey's Law can find an explanation in Articulatory Phonology: both can be viewed as a change in the relative phasing of gestures, which is attributable to a change in the coupling strength. Specifically, when the two consonant gestures nearly overlap with each other at the left peripheral of the vowel

gesture, the output sequence is /pra/. A significant delay in the onset of the second consonant gesture ceases the overlap of the two consonant gestures, yielding /pa.ra/. The displacement of the second consonant gesture to the right peripheral of the vowel gesture gives rise to /par/.

However, in the case of the vowel insertion /pa.ra/, there can be free variation in the actual size of non-peripheral displacement which turns /pra/ into /pe.ra/. The inserted vowel /e/ can be construed as a vowel of indeterminate quality, which arises out of a displacement that is large enough to leave behind a vowel quality /e/ but too small to form a full vowel /a/ (Steriade, 1990). It was thus argued that languages or dialects consistently select either displacement to the peripheral position or non-peripheral position, while allowing for gradient variation in the actual size of timing adjustment.

A pertinent issue brought up by Steriade (1990) is the internal duration of articulatory gestures. Recall that the duration of a gesture is determined by the stiffness parameter of the planning oscillator associated with the gesture. The notion of internal duration allows for the possibility of distinguishing multiple phasing relationships between gestures. In the current experiment, an account in which both the articulatory and lexical tone gestures have internal duration is essential to address both the categorical and gradient changes in imitation relative timing, which an alternative autosegmental analysis falls short of tackling.

In Autosegmental Metrical theory, autosegmental units are usually viewed as points in time thus do not have a beginning or ending point. The punctuality of autosegments dictates that the timing relationship between two autosegmental units can only be sequential, i.e., precedence or subsequence, leaving no explanation for overlapping. Even when auxiliary tiers like the the CV timing tier (c.f.

Steriade, 1990, and references therein) are introduced, an autosegmental analysis is still unable to distinguish simultaneity from partial overlapping, which is both a phonological reality and a phonetic reality. Hence, the proposal of internal duration in Articulatory Phonology is the right step towards a theory with greater integration of phonological and phonetics.

Additional findings of Experiment 1 support the notion that TP relative timing and F0 vary with one another depending on the degree of gestural overlap between the two tone gestures in a bi-tonal sequence. Specifically, for an L+H sequence (Experiment 1A), as the gestural overlap decreases, i.e., the gestural onset of the H gesture occurs later, the TP occurs later with lower F0. For an H+L sequence (Experiment 1B), as the gestural overlap decreases, i.e., the gestural onset of the L gesture occurs later, the TP occurs later with higher F0. Note that the changes in the gestural coordination that further lead to the changes in the gestural overlap, depending on the participant, can either be sub-phonemic variation within the same category (Tone2), or categorical changes from Tone2 to Full Tone3 (Experiment 1A) or from Tone2 to Tone4 (Experiment 1B). From a gestural point of view, the relationship between TP relative timing and F0 should nonetheless hold regardless of the nature of the changes in the gestural coordination.

The co-variation of TP relative timing and F0 can be conveniently accounted for by the internal duration proposal, which allows for multiple phasing relationships between the two lexical tone gestures. On the other hand, under an autosegmental analysis, one has to resort to external mechanisms like f_0 range expansion (c.f. Arvaniti et al., 2006) to account for such gradient changes.

In sum, the results of Experiment 1 show that the tone-to-segment alignment pattern is governed by the categorical modes of gestural coordination that corre-

spond to lexical tones in Mandarin. The categorical alignment patterns, however, are subject to gradient stimulus-induced adjustments, which can be interpreted as the result of the gradient changes in the coupling strength between lexical tone gestures and oral articulatory gestures. Therefore, both the categorical shifts and the gradient changes in the alignment patterns can be accounted for by adjusting the coupling strength between certain gestures in a gestural constellation, which is made possible by the notion that gestures have internal duration, as proposed by Articulatory Phonology. The idea that f_0 control can be conceptualized as gestures is further supported by the finding regarding the co-variation between TP relative timing and F0.

5.2 Experiment 2

In Experiment 2, the interaction between lexical tones and intonation was investigated within the framework of Articulatory Phonology, under the premise that f_0 control can be conceptualized as f_0 gestures that are coordinated with oral articulatory gestures. Two models of the interaction between lexical tones and intonation, i.e., the overlay model and the unification model, were tested with articulatory and acoustic data in an EMA experiment. The temporal lags between the C and V gestures, and between the V and T gestures, were collected for the target syllables [ma] occurring in various prosodic contexts.

It is found that the relative timing pattern of gestural activations of the C, V, and T gestures (CV%) is altered by intonational events such as boundary tones and prosodic focus. The changes in the CV% can be broken down into the changes in both the C-V onset lag and the V-T onset lag. Comparing the CV% in MedUn

to that in FinUn, the presence of boundary tones increases the C-V onset lag and decreases the V-T onset lag, therefore increasing CV%. Comparing the CV% in FinUn to that in FinAc, prosodic focus functions in a similar way to boundary tones for Tone2-bearing syllables, but not for Tone4-bearing syllables. In the latter case, both the C-V onset lag and the V-T onset lag increase, resulting in decreases in the CV% for some participants, and consequently, no significant change in the CV% globally. The unification model is thus supported by the empirical results, suggesting that the intonational tones can influence the intra-syllabic gestural coordination of lexical tone gestures and oral articulatory gestures.

It is proposed that BT gestures are triggered in FinUn. The presence of the BT gesture interferes with the intra-syllabic gestural coordination between the C, V, and the lexical tone gestures, altering the relative timing patterns of C-V-T from MedUn to FinUn. Specifically, the changes in the relative timing patterns are attributed to the changes in the relative phasing, which in turn arise out of the changes in the relative coupling strength between the oral articulatory gestures and the T gestures. The latter changes occur due to the influence of the BT gesture on the intra-syllabic gestural coordination of the target syllable, which is illustrated in Figure 5.3. In FinUn, the presence of the BT gesture increases the coupling strength between the V gesture and the T gestures, drawing the V gesture closer to the T gestures, thereby increasing the CV%.

The influence of the BT gesture on the intra-syllabic gestural coordination can be brought into play via its coordination with the oral articulatory gestures and the lexical tone gestures. An alternative account could be that the BT gesture is not directly coupled to the articulatory gestures or the lexical tone gestures, but affects the intra-syllabic gestural coordination through some external mechanism.

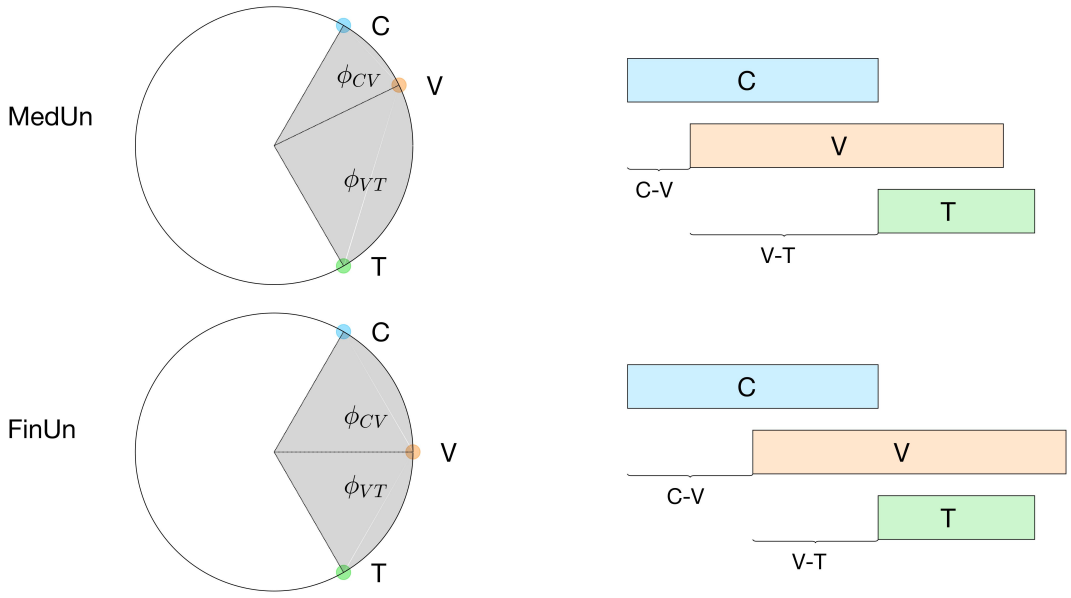


Figure 5.3: Coupled oscillators (left) and gestural scores (right) in MedUn (top) and FinUn (bottom).

It is speculated that the mechanism is related to the exertive force (attention, effort, etc.) account proposed by Tilsen (2017). It was argued that each planning oscillator is associated with an ensemble of neurons. When an exertive force applies, the neurons in the ensemble exhibit a collective oscillation that can increase the coupling strength by way of increasing the the ensemble size. In the current experiment, it is possible that the phrase boundary introduces an extertive force, increasing the ensemble size associated with the lexical tone gestures, and the V-T in-phase coupling strength. In both accounts, it is undeniable that the intra-syllabic gestural coordination is altered by the BT gesture.

Note that it is also possible that the presence of the BT gesture decreases the coupling strength between the C and V gestures, which also leads to a relative increase in the V-T lag and CV%. This explanation is compatible with the exertive force account as the influence of the BT gesture on the intra-syllabic gestural co-

ordination is indirect. However, it is more likely that the BT gesture indirectly influences the intra-syllabic gestural coordination by affecting the coupling relations of the lexical tone gestures.

Moreover, it is proposed that the μ gesture is triggered by focus-introduced pitch accent. The μ gesture functions in a similar way to the BT gesture. However, for Tone4-bearing syllables, the anti-phase coupling between the C and T gestures becomes stronger in the presence of the μ gesture, resulting in a larger C-T onset lag, which captures the increases in both the C-V onset lag and the V-T onset lag. The μ gesture can affect the intra-syllabic gestural coordination by directing coupling or via an exertive mechanism. Importantly, the lexical-tone-specific influence of the μ gesture supports the notion that intonational tones are locally produced and are interacting with the lexical tones, rather than operating on a global level.

Note that throughout the discussion, the term "sequential" has been avoided on purpose when describing the relationship between intonation and lexical tones, although such usage can be seen elsewhere (c.f. Gibson, 2013). Sequentiality is the by-product of AM theory of intonation, because phonological units are points in time and have to be arranged in a sequential fashion. But sequentiality does not necessary hold true within a gestural framework because gestures have internal duration that allows for multiple modes of relative phasing. Therefore, in the proposed unification model, intonational tone gestures are locally produced, and can be coordinated with lexical tone gestures with certain degree of gestural overlap (c.f. Figure 4.15). Importantly, the locally produced intonational tone gestures affect intra-syllabic gestural coordination via coupling relations associated with lexical tone gestures.

Another finding of note is that SentInto does not affect the CV%: regardless

of Context, the C-V-T coordinative pattern does not differ between Statement and Question. A unified BT gesture is thus proposed to account for the coordinative differences between different Context. This seems in direct contrast with acoustic evidence that question intonation is implemented differently from statement intonation. Yuan (2004) found that question intonation results in higher overall f_0 curve, higher strength of sentence-final tones, and steeper (Tone2) or flatter (Tone4) final tones than statement intonation. It was argued that these acoustic differences can be attributed to a combination of "question" mechanisms: a global phrase curve mechanism, a partially global, tone-independent strength mechanism, and a local tone-dependent mechanism. The discrepancy here is that the proposed BT gesture is intonation-independent, while the question mechanisms are intonation-dependent. One possible explanation is that the unified BT gesture is associated with phrase boundary, which is tone-insensitive, and that different boundary tones, e.g., H% and L%, interact differently with lexical tones.

In sum, the results of Experiment 2 show that intonation affects the alignment pattern of lexical tones and segments, supporting the unification model. A gesture account is put forward within the AP model: both the BT gesture and the μ gesture, triggered by boundary tones and focus-introduced pitch accents, respectively, interfere with the coupling relations between the lexical tone gestures and the oral articulatory gestures, altering the intra-syllabic gestural coordination, i.e., CV%. Moreover, the μ gesture affects the intra-syllabic gestural coordination of Tone2-bearing syllables differently from that of Tone4-bearing syllables, further driving home the point that intonational tones are locally produced.

5.3 Future Research

Conceptualizing f_0 control as gestures is not a novel idea. Both intonational and lexical use of f_0 have been modeled as gestures (Gao, 2008; Mücke et al., 2012; Katsika et al., 2014; Yi and Tilsen, 2014). The tone-to-segment alignment patterns can be associated with the interaction between f_0 gestures and oral articulatory gestures associated segments within the framework of AP.

In Experiment 1, rather than stipulating arbitrary acoustic anchors, AP accounts for the categorical differences in alignment pattern with different modes of gestural coordination between oral articulatory gestures and f_0 gestures. By further manipulating the coupling strength between certain pairs of gestures, gradient changes in alignment pattern can be achieved.

The gestural model is perceptually helpful in understanding the relative timing patterns in Experiment 1. Nevertheless, these patterns are acoustic in nature. Admittedly, acoustic tone-to-segment alignment is an indirect approximation of articulatory gestural coordination, but it cannot shed light on the coupling relationship between the C and V gestures. Without proper articulatory data, significant parts of the proposed gestural account like the C-V coupling would remain speculative. Therefore, in order to justify this model, future research can seek to collect articulatory data in an imitation study like the current experiment, and investigate the interplays between oral articulatory gestures and lexical tone gestures in various stimulus conditions.

In Experiment 2, intonational tone gestures (such as the BT gesture and the μ gesture) are proposed to capture the differences in relative timing pattern of

gestural activations in various prosodic contexts. The presence of the BT gesture and the μ gesture interfere with the intra-syllabic gestural coordination of the C, V and the lexical tone gestures, therefore altering the relative timing patterns.

Within the gestural framework, it is proposed in one account that the intonational tone gestures are coupled to the oral articulatory gestures like the lexical tone gestures. This account falls in line with the premise that both intonation and lexical tones use the same type of f_0 control, i.e., the same group of laryngeal muscles. However, whether this is true remains unclear since intonational processes also involve other aspects of control (such as intensity and duration) than only f_0. Moreover, kinematic data of laryngeal activities in f_0 control are still lacking. In this case, future studies can contribute to the current understanding of the field by shedding light on the similarities and differences in the physiological aspects of the f_0-related gestures, i.e., lexical tone gestures and intonational tone gestures.

APPENDIX A

EXPERIMENT I

Onset		TP1		TP2		TP3		Offset	
Time (ms)	F0 (Hz)	Time	F0	Time	F0	Time	F0	Time	F0
0	220	80	165	375	230	550	190	737	240
0	220	80	170	375	230	550	190	737	240
0	220	80	175	375	230	550	190	737	240
0	220	80	180	375	230	550	190	737	240
0	220	120	165	375	230	550	190	737	240
0	220	120	170	375	230	550	190	737	240
0	220	120	175	375	230	550	190	737	240
0	220	120	180	375	230	550	190	737	240
0	220	160	165	375	230	550	190	737	240
0	220	160	170	375	230	550	190	737	240
0	220	160	175	375	230	550	190	737	240
0	220	160	180	375	230	550	190	737	240
0	220	200	165	375	230	550	190	737	240
0	220	200	170	375	230	550	190	737	240
0	220	200	175	375	230	550	190	737	240
0	220	200	180	375	230	550	190	737	240
0	220	240	165	375	230	550	190	737	240
0	220	240	170	375	230	550	190	737	240
0	220	240	175	375	230	550	190	737	240
0	220	240	180	375	230	550	190	737	240

Table A.1: f_0 parameters of synthesized stimuli in Experiment 1A for female participants. Only f_0 turning point 1 (TP1) vary in two dimensions: its relative timing to the acoustic onset of the first [ma2] (from 80 ms to 240 ms) and its fundamental frequency (from 165 Hz to 180 Hz).

Onset		TP1		TP2		TP3		Offset	
Time (ms)	F0 (Hz)	Time	F0	Time	F0	Time	F0	Time	F0
0	130	80	102	375	135	550	115	737	140
0	130	80	105	375	135	550	115	737	140
0	130	80	108	375	135	550	115	737	140
0	130	80	110	375	135	550	115	737	140
0	130	120	102	375	135	550	115	737	140
0	130	120	105	375	135	550	115	737	140
0	130	120	108	375	135	550	115	737	140
0	130	120	110	375	135	550	115	737	140
0	130	160	102	375	135	550	115	737	140
0	130	160	105	375	135	550	115	737	140
0	130	160	108	375	135	550	115	737	140
0	130	160	110	375	135	550	115	737	140
0	130	200	102	375	135	550	115	737	140
0	130	200	105	375	135	550	115	737	140
0	130	200	108	375	135	550	115	737	140
0	130	200	110	375	135	550	115	737	140
0	130	240	102	375	135	550	115	737	140
0	130	240	105	375	135	550	115	737	140
0	130	240	108	375	135	550	115	737	140

0	130	240	110	375	135	550	115	737	140

Table A.2: f_0 parameters of synthesized stimuli in Experiment 1A for male participants. Only f_0 turning point 1 (TP1) vary in two dimensions: its relative timing to the acoustic onset of the first [ma2] (from 80 ms to 240 ms) and its fundamental frequency (from 102 Hz to 110 Hz).

Onset		TP1		TP2		TP3		Offset	
Time (ms)	F0 (Hz)	Time	F0	Time	F0	Time	F0	Time	F0
0	220	160	180	275	225	550	190	737	240
0	220	160	180	275	230	550	190	737	240
0	220	160	180	275	235	550	190	737	240
0	220	160	180	275	240	550	190	737	240
0	220	160	180	325	225	550	190	737	240
0	220	160	180	325	230	550	190	737	240
0	220	160	180	325	235	550	190	737	240
0	220	160	180	325	240	550	190	737	240
0	220	160	180	375	225	550	190	737	240
0	220	160	180	375	230	550	190	737	240
0	220	160	180	375	235	550	190	737	240
0	220	160	180	375	240	550	190	737	240
0	220	160	180	425	225	550	190	737	240
0	220	160	180	425	230	550	190	737	240
0	220	160	180	425	235	550	190	737	240
0	220	160	180	425	240	550	190	737	240
0	220	160	180	475	225	550	190	737	240
0	220	160	180	475	230	550	190	737	240

0	220	160	180	475	235	550	190	737	240
0	220	160	180	475	240	550	190	737	240

Table A.3: f_0 parameters of synthesized stimuli in Experiment 1B for female participants. Only f_0 turning point 2 (TP2) vary in two dimensions: relative timing to the acoustic onset of the first [ma2] (from 275 ms to 475 ms) and fundamental frequency (from 225 Hz to 240 Hz).

Onset		TP1		TP2		TP3		Offset	
Time (ms)	F0 (Hz)	Time	F0	Time	F0	Time	F0	Time	F0
0	130	160	102	275	132	550	115	737	140
0	130	160	105	275	135	550	115	737	140
0	130	160	108	275	138	550	115	737	140
0	130	160	110	275	140	550	115	737	140
0	130	160	102	325	132	550	115	737	140
0	130	160	105	325	135	550	115	737	140
0	130	160	108	325	138	550	115	737	140
0	130	160	110	325	140	550	115	737	140
0	130	160	102	375	132	550	115	737	140
0	130	160	105	375	135	550	115	737	140
0	130	160	108	375	138	550	115	737	140
0	130	160	110	375	140	550	115	737	140
0	130	160	102	425	132	550	115	737	140
0	130	160	105	425	135	550	115	737	140
0	130	160	108	425	138	550	115	737	140
0	130	160	110	425	140	550	115	737	140
0	130	160	102	475	132	550	115	737	140

0	130	160	105	475	135	550	115	737	140
0	130	160	108	475	138	550	115	737	140
0	130	160	110	475	140	550	115	737	140

Table A.4: f_0 parameters of synthesized stimuli in Experiment 1B for male participants. Only f_0 turning point 1 (TP1) vary in two dimensions: its relative timing to the acoustic onset of the first [ma2] (from 275 ms to 475 ms) and its fundamental frequency (from 132 Hz to 140 Hz).

APPENDIX B

EXPERIMENT 2

Phrasal Context	Target	(Prompt) Stimulus
MedUn	Tone4	(−) lu^4 yan^4 yi^4 *ma*^{4} yi^4 de^5 hen^3 kuai^4? 'Lu Yan translates *ma*^{4} very fast?'
FinUn	Tone4 + H%	(bu^2 shi^4 luo^2 yan^4? bu^2 shi^4 li^3 yan^4?) ('Not Luo Yan? Not Li Yan?') **lu**^{4} yan^4 lai^2 yi^4 *ma*^{4}? mei^2 ting^1 shuo^1. '**Lu** Yan (will) come and translate *ma*^{4}? I have not heard.'
FinAc	**Tone4** + H%	(bu^2 shi^4 ma^2? bu^2 shi^4 ma^3?) ('Not ma^2? Not ma^3?') lu^4 yan^4 lai^2 yi^4 ***ma***^{4}? mei^2 ting^1 shuo^1. 'Lu Yan (will) come and translate ***ma***^{4}? I have not heard.'

Table B.1: Question elicitations of Tone4-bearing target syllables in three different carriers: phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc). Target syllables are italicized. Accented syllables are in bold. Prompts are provided for FinUn and FinAc elicitations.

Phrasal Context	Target	(Prompt) Stimulus
MedUn	Tone4	(−) lu^4 yan^4 yi^4 *ma*^{4} yi^4 de^5 hen^3 kuai^4. 'Lu Yan translates *ma*^{4} very fast.'
FinUn	Tone4 + L%	(bu^2 shi^4 luo^2 yan^4? bu^2 shi^4 li^3 yan^4.) ('Not Luo Yan. Not Li Yan.') **lu**^{4} yan^4 lai^2 yi^4 *ma*^{4}. mei^2 ting^1 shuo^1 ma^5? '**Lu** Yan (will) come and translate *ma*^{4}. Have you not heard?'
FinAc	**Tone4** + L%	(bu^2 shi^4 ma^2. bu^2 shi^4 ma^3.) ('Not ma^2. Not ma^3.') lu^4 yan^4 lai^2 yi^4 ***ma***^{4}. mei^2 ting^1 shuo^1 ma^5? 'Lu Yan (will) come and translate ***ma***^{4}. Have you not heard?'

Table B.2: Statement elicitations of Tone4-bearing target syllables in three different carriers: phrase-medial un-accented (MedUn), phrase-final un-accented (FinUn), and phrase-final accented (FinAc). Target syllables are italicized. Accented syllables are in bold. Prompts are provided for FinUn and FinAc elicitations.

BIBLIOGRAPHY

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

Arvaniti, A., Ladd, D. R., and Mennen, I. (2006). Phonetic effects of focus and "tonal crowding" in intonation: Evidence from greek polar questions. *Speech Communication*, 48(6):667–696.

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

Beckman, M. and Ayers Elam, G. (1997). *Guidelines for ToBI labelling, version 3*. The Ohio State University Research Foundation, Ohio State University.

Bolinger, D. (1964). Intonation: Around the edge of language. *Harvard Educational Review*, 34:282–296.

Browman, C. P. and Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252.

Browman, C. P. and Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45:140–155.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6:201–251.

Browman, C. P. and Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320.

Browman, C. P. and Goldstein, L. (1990b). Representation and reality: Physical systems and phonological structure. *Journal of Phonetics*, 18:411–424.

Browman, C. P. and Goldstein, L. (1990c). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 341–376. Cambridge University Press.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49:155–180.

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Chen, Y. and Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36:724–746.

Clements, G. (1992). Phonological primes: Features or gestures? *Phonetica*, 49:181–193.

de Boor, C. (2001). *A practical guide to splines (revised edition)*. New York: Springer.

D'Imperio, M., Espesser, R., Lœvenbruck, H., Menezes, C., Nguyen, N., and Welby, P. (2004). Are tones aligned to articulatory events? Evidence from Italian and French. In Cole, J. and Hualde, J. I., editors, *Laboratory Phonology 9*. Berlin: Mouton de Gruyter.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In MacNeilage, P. F., editor, *The Production of Speech*, pages 39–55. New York: Spring-Verlag.

Gao, M. (2008). *Mandarin tones: An articulatory phonology account*. PhD thesis, Yale University.

Gårding, E. (1983). A generative model of intonation. In Cutler, A. and Ladd, D. R., editors, *Prosody, Models and Measurements*, pages 11–21. Springer Verlag.

Gibson, M. K. (2013). *Lexical tone, intonation and their interaction: a scopal theory of tune association*. PhD thesis, Cornell University.

Goldstein, L., Nam, H., Saltzman, E., and Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. *Frontiers in phonetics and speech science*, pages 239–250.

Grice, M. (1995). Leading tones and downstep in english. *Phonology*, 12(02):183–

233.

Gubian, M., Torreira, F., and Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49:16–40.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: from global to local prosody. In *Speech Prosody 2004, International Conference*.

Katsika, A., Krivopapić, J., Moonshammer, C., Tiede, M., and Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics*, 44:62–82.

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

Ladd, D. R., Faulkner, D., Faulkner, H., and Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rage. *Journal of Acoustical Society of America*, 106(3):1543–1554.

Ladd, D. R., Mennen, I., and Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America*, 107(5):2685–2696.

McGowan, R. and Saltzman, E. (1995). Incorporating aerodynamics and laryngeal components into task dynamics. *Journal of Phonetics*.

Moisik, S. R., Lin, H., and Esling, J. H. (2014). A study of laryngeal gestures in mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus). *International Phonetic Association. Journal of the International Phonetic Association*, 44(1):21.

Mücke, D., Nam, H., Hermes, A., and Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. In Hoole, P., Bombien, L., Pouplier, M., Moonshammer, C., and Kühnert, B., editors, *Consonat clusters and structural*

complexity, pages 205–230. Berlin: Mouton de Gruyter.

Nam, H. and Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th International Congress of Phonetic Sciences*.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(132-142).

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

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics*, 18:375–394.

Pierrehumbert, J. and Beckman, M. (1988). *Japanese tone structure*. Cambridge, MA: MIT Press.

Prieto, P., D'Imperio, M., and Fivela, B. G. (2005). Pitch accent alignment in Romance: Primary and secondary associations with metrical structure. *Language and Speech*, 48(4):359–396.

Prieto, P., van Santen, J., and Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23:429–451.

Saltzman, E. and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382.

Scheider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime User's Guide*. Psychology Software Tools, Inc., Pittsburgh, PA.

Steriade, D. (1990). Gestures and autosegments. In Beckman, M. and Kingston, J., editors, *Papers in Laboratory Phonology*, pages 382–397. Cambridge University Press.

The MathWorks, Inc. (2016). MATLAB and Statistics Toolbox Release 2016a.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emer-

gence of phonological structure. *Journal of Phonetics*.

Tilsen, S. (2017). Exertive modulation of speech and articulatory planning. *Journal of Phonetics*.

Tilsen, S. and Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of Acoustical Society of America*, 134(1):628–639.

Tilsen, S., Burgess, D., and Lants, E. (2013). Imitation of intonational gestures: A preliminary report. *Cornell Working Papers in Phonetics and Phonology*, 1:1–17.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25:61–83.

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. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46:220–251.

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, pages 7–10, Aix-en-Provence, France.

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33:319–337.

Yi, H. and Tilsen, S. (2014). A gestural account of mandarin tone sandhi. In *Proceedings of Meetings on Acoustics*, volume 136.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006).

The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.

Yuan, J. (2004). *Intonation in Mandarin Chinese: acoustic, perception, and computational modeling*. PhD thesis, Cornell University.