A Gestural Account of Mandarin Tone Sandhi

Hao Yi *

1 Introduction

Recently tones have been analyzed as articulatory gestures, which may be coordinated with segmental gestures (Gao 2008). Using Electromagnetic Articulometry (EMA), this paper investigates the timing of tonal gestures and articulatory gestures in Mandarin tone sandhi under the framework of Articulatory Phonology (AP).

Analyses of the timing patterns show that purported neutralized phonological contrast can nonetheless exhibit coordinative differences in Mandarin tone sandhi. Furthermore, it is likely that a bias towards the underlying tone (lexical Tone3) is responsible for the incomplete neutralization between Tone2 and the output of third tone sandhi (henceforth T3S) .

The rest of the paper is organized as follows: Section 2 offers background on Mandarin tone sandhi and the motivation for the current study; Section 3 lays out the theoretical framework, i.e. AP and two major AP-based accounts that inspired this paper; Section 4 formulates the hypotheses; Section 5 covers the methodology, including the experiment design and statistical methods; Section 6 presents the results, and Section 7 provides closing discussion and some final thoughts.

2 Background

2.1 Mandarin Tone sandhi

Mandarin has a four-tone system: high-level Tone1 (55), rising Tone2 (35), falling-rising Tone3 (213), and falling Tone4 (51).[1] Tone3 participates in two oft-cited tone sandhis: half third sandhi and third tone sandhi. Half third sandhi applies to the first syllable of a disyllabic sequence Tone3 + ToneX, where ToneX is not Tone3. As a result, Tone3 changes from 213 to 21, giving rise to a half third tone (henceforth T3H). Third tone sandhi, on the other hand, applies to the first syllable of a disyllabic sequence Tone3 + Tone3. The output of third tone sandhi (henceforth T3S) surfaces as having a rising pitch contour that resembles Tone2. Figure 1 shows the F0 contours of Tone2, T3S, and T3H.

*I thank Sam Tilsen for his guidance and feedback. I also thank Abby Cohn, Draga Zec, and Robin Karlin for their feedback.

[1]Numerical values in parenthesis represent pitch height on a five-point scale. 5 represents the maximal value, while 1 represents the minimal value (Chao 1968).

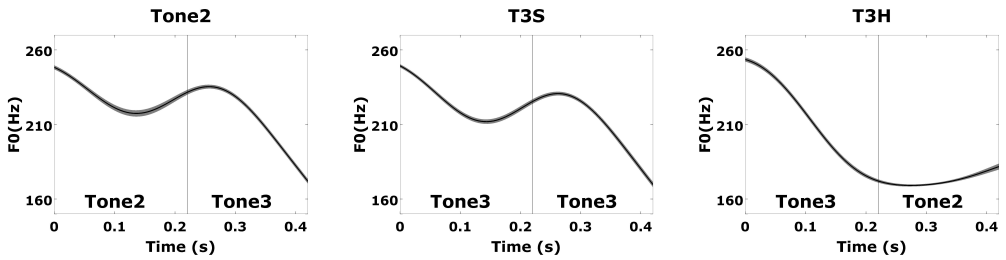


Figure 1: Mean F0 + / - 1.0 stand error of Tone2 (*left*), T3S (*middle*), and T3H (*right*) following Tone2. The base tones are illustrated at the bottom of each figure. T3S is acoustically similar to Tone2 in that both have a rising F0 contour; this contrasts to T3H, which is low low tone.

Despite the similarity between Tone2 and T3S, subtle acoustic differences between the two rising tones have been reported in several studies. Chen and Yuan (2007) showed that the F0 rise of T3S was both shorter in duration and smaller in range than that of Tone2. Therefore, it was suggested by Chen and Yuan (2007) that third tone sandhi was not the change of one toneme (Tone3) to anther toneme (Tone2). Instead, T3S and Tone2 were not completely neutralized, i.e. they were not merged into one toneme.

Moreover, following studies showed more evidence that demonstrated that there were acoustic differences between T3S and Tone2: the F0 turning point that marks the transition between the dipping and rising was both later and lower in T3S than that in Tone2; the rime duration was longer in T3S-bearing syllables than Tone2-bearing syllables (Peng 2000, Chen and Yuan 2007, Zhang and Lai 2010). As suggested by Zhang and Lai (2010), the shape of T3S is reminiscent of lexical Tone3, which also has a later and lower F0 turning point; similarly, the rime duration of Tone3-bearing syllables is longer than that of Tone2-bearing syllables. In other words, T3S and Tone3 are still acoustically similar to some extent.

It will be of great help to understand the difference in underlying representation between T3S and Tone2. However, previous work on the incomplete neutralization of T3S has predominantly done so from an acoustic perspective. Little work has been done from an articulatory perspective. The timing of tones and segments can lend valuable insight into this long-standing yet still perplexing phenomenon. This study investigates both the third tone sandhi and the half third sandhi, from an articulatory perspective. We conducted a production study using EMA (electromagnetic articulometry), with a special focus on the timing of tone gestures and articulatory gestures in both sandhi phenomena. The aim of the current study is to formulate a model that takes into account all three variants of Tone3.

3 Framework

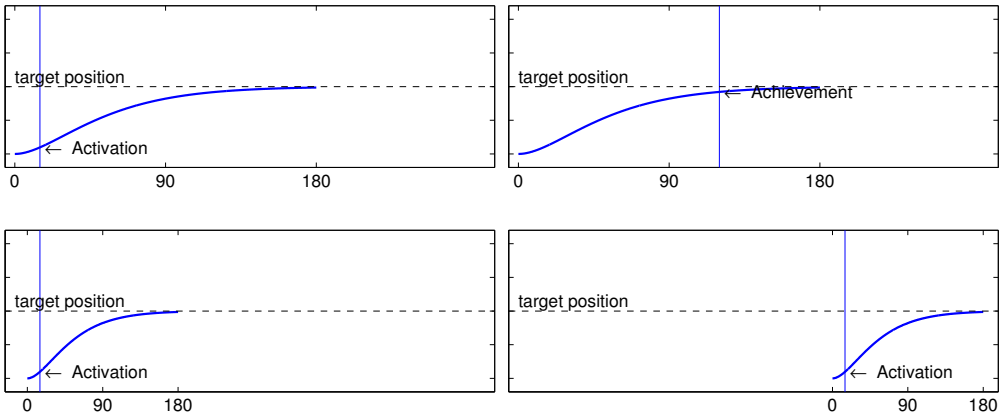
3.1 Articulatory Phonology

Under the framework of AP, articulatory gestures are proposed to be the basic units of phonological structure (Browman and Goldstein 1989, 1990, 1992). Gestures are one-dimensional point-attractor dynamical systems that are associated with target values of vocal tract geometry that are achieved by the coordinated movements of articulators (such as the lower lip, upper lip, jaw, tongue, and velum). For example, a bilabial closure gesture is associated with three articulators, namely the upper lip, the lower lip and the jaw; move-

ments of these articulators are coordinated so as to achieve a negative value of the tract variable, in this case lip aperture.

Two or more gestures can be organized with each other in a specific way to form larger structures. As a one-dimensional pointer system, each gesture is modeled as an oscillator with its own virtual cycle, and is activated by a defined phase (e.g. 0° point of the virtual cycle). When two gestures are coordinated, two points (or phases), one on each virtual cycle of the gesture, must coincide temporally. That is, two gestures are coordinated by coupling their corresponding virtual cycle with a relative phase.

There are two preferred ways in which a pair of gestures can be coupled: in-phase or anti-phase; the former is the more stable coupling relation. Besides these two coupling modes, gestures can also couple in modes with other phasings. Also note that not every gesture in an utterance is coordinated with every other gesture.



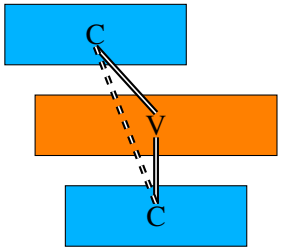
(a) In-phase coupling mode (b) Anti-phase coupling mode

Figure 2: Illustrations of two coupling modes – in-phase coupling (2a) and anti-phase coupling (2b). In the in-phase coupling mode, two gestures are coupled to each other with a relative phase of 0°; in the anti-phase coupling mode, two gestures are coupled to each other with a relative phase of 180°.

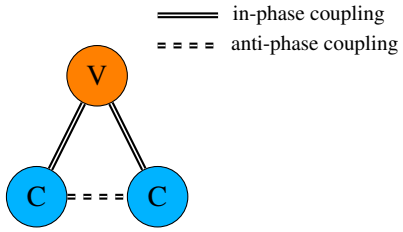
According to the original C-V coupling hypothesis, an onset consonant (henceforth C) gesture is in-phase coupled to the vowel (henceforth V). In a English word such as *me*, which forms a CV syllable, the C gesture is in-phase coupled to the V gesture. What is meant by in-phase coupling is that the C gesture and the V gesture, in this example the bilabial gesture of [m] and the tongue body gesture of [i], are initiated at the same time. This is in contrast to an acoustic point of view, which indicates that the two sounds are articulated sequentially.

Moreover, the C gestures in an onset cluster are anti-phase coupled to each other. In an English word such as *plea*, which forms a CCV syllable, the onset C gestures (a bilabial gesture of [p^h] and a tongue tip gesture of [l]) are anti-phase coupled to each other, while both C gestures are in-phase coupled to the V gesture (a tongue body gesture of [i]). The collective force of these coupling relations results in a pattern in which the onsets of the C gestures are displaced equally in opposite directions in time from the onset of the V gesture (Nam and Saltzman 2003); this is known as the C-center effect (Browman and Goldstein 1989). Therefore, in *plea*, the midpoint between the onsets of the C gestures (the bilabial gesture and the tongue tip gesture) corresponds to the onset of the V gesture (the tongue

body gesture). Figure 3a shows the gestural score and Figure 3b shows the abstract coupling relations in a CCV syllable in English.



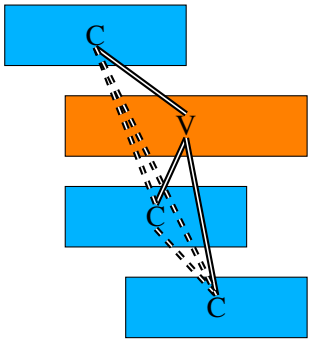
(a) Gestural score



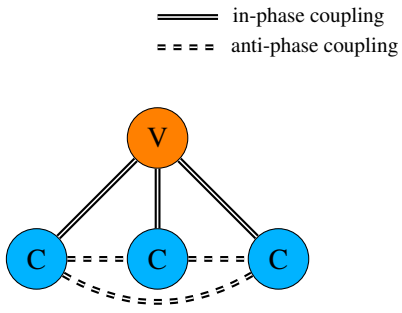
(b) Coupling relations

Figure 3: Illustrations of the C-center effect in a CCV syllable in English. The onsets of the C gestures are displaced equally in opposite directions in time from the onset of the V gesture. The C gestures are in-phase coupled to the V gesture and anti-phase coupled to each other. Solid lines indicate in-phase coupling and dashed lines indicate anti-phase coupling.

Similarly, in a English word such as *split*, which forms a CCCV syllable, the collective force of the coupling interactions renders a near synchronization of the V gesture and the second C gesture. Therefore, in *split*, the tongue body gesture of [ɪ] is initiated at approximately the same time as the bilabial gesture of [p]. Figure 4a shows the gestural score and Figure 4b shows the abstract coupling relations in a CCCV syllable in English.



(a) Gestural score



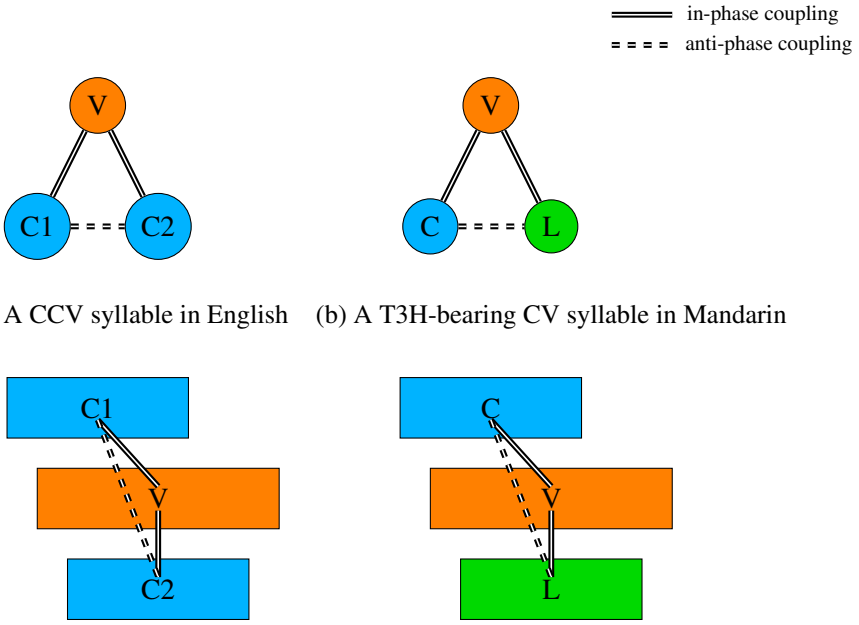
(b) Coupling relations

Figure 4: Illustrations of the C-center effect in a CCCV syllable. The overall timing of the center of the C onsets with respect to the V onset is preserved. The C gestures are in-phase coupled to the V gesture, and anti-phase coupled to each other. Solid lines indicate in-phase coupling and dashed lines indicate anti-phase coupling.

3.2 An AP Account of Mandarin Tones

With tone coming into play, Mandarin shows somewhat different coupling interactions from non-tonal languages like English. Gao (2008) proposed that lexical tones in Mandarin Chinese can be analyzed as a single tone (henceforth T) gesture or combinations of T gestures. T gestures, namely High and Low (henceforth H and L), can be coordinated with segmental

gestures like onset C gestures. C and T gestures, like C gestures in an onset cluster, are anti-phase coupled to each other, and both C and T gestures are in-phase coupled to V gestures. Therefore, a C-center effect emerges in a CV syllable that bears a level tone (Tone1 or T3H) in Mandarin: the onset of the V gesture is initiated halfway between the onsets of the C gesture and the T gesture, as illustrated in Figure 5.



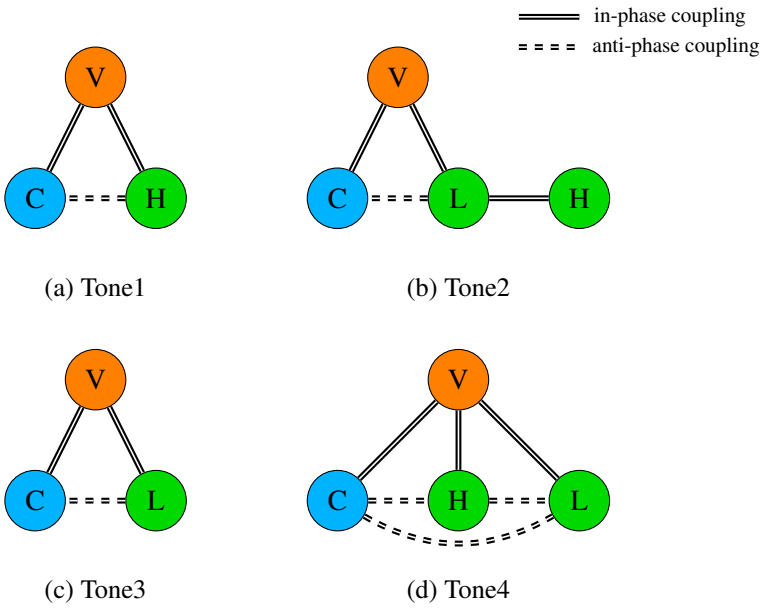
(a) A CCV syllable in English (b) A T3H-bearing CV syllable in Mandarin

(c) A CCV syllable in English (d) A T3H-bearing CV syllable in Mandarin

Figure 5: Analogy in coupling relations (*top*) and gesture scores (*bottom*) between an English CCV syllable and a Mandarin T3H-bearing CV syllable. The T (L) gesture behaves like an onset C gesture. Both syllables display a C-center effect.

Figure 6 further illustrates the proposed coupling relations of four Mandarin tones in Gao (2008).[2] The four tones of Mandarin fall into two categories: Tone1, Tone2 and T3H fall into one, and Tone4 stands on its own. In the first group, the T gesture is initiated halfway between the C gesture and the T gesture. Specifically, the high-level Tone1 (55) has one single H gesture, and low T3H (21) has one single L gesture. These two tones are good fits for the aforementioned C-center model due to their relatively simple tonal compositions. In the case of the rising Tone2 (35), two T gestures (L and H) are involved. Based on empirical speech data, Gao (2008) argued that the two T gestures function as only one additional onset-like T gesture. Moreover, instead of being arranged in a sequential fashion, the L gesture is initiated in synchrony with the H gesture and has intrinsically shorter duration than the H gesture. The overlap between the two tone gestures is responsible for the early dip in the pitch contour of Tone2, which is often characterized as an 'undershoot'. Therefore, Tone2 is no different from Tone1 and Tone3 in terms of the relative timing of the V gesture with respect to the C-center. As shown in Figure 7 and 8, an analogy can be drawn between a Tone2-bearing syllable and, for example, the CCV syllable *smee* in English, where [m] has two C gestures, namely the bilabial gesture and the velum gesture,

[2]Note that Tone3 (21) in this analysis refers to T3H.

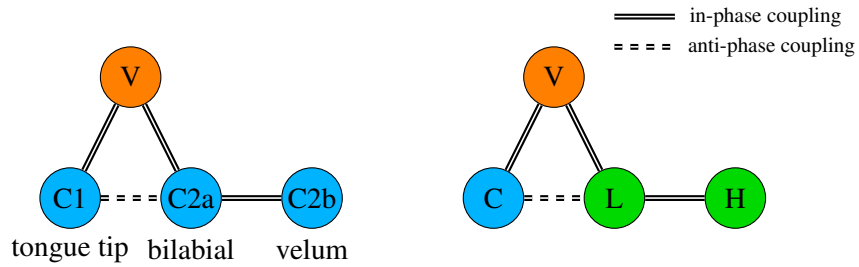


(a) Tone1 (b) Tone2

(c) Tone3 (d) Tone4

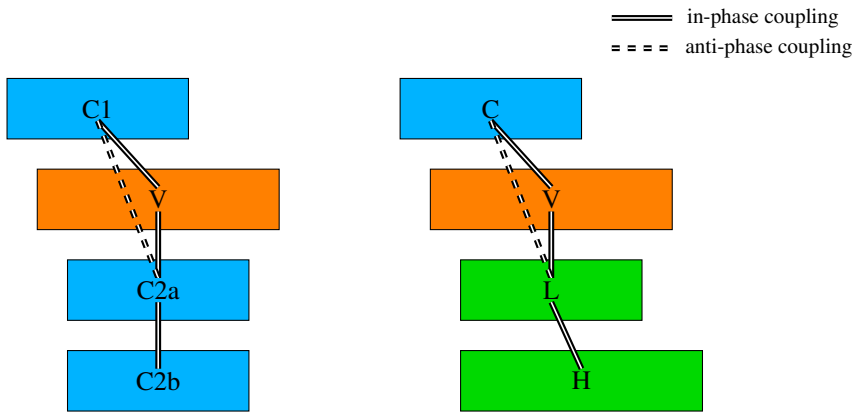
Figure 6: Coupling relations of Mandarin four tones in Gao (2008). T gestures behave like C gestures: they are in-phase coupled to V gesture and anti-phase coupled to onset C gestures. For Tone1, Tone2 and Tone3 (T3H), the V gesture is initiated halfway between the C gesture and the T gesture; for Tone4, the V gesture is initiated in synchrony with the T1 (H) gesture.

[s] has a tongue tip gesture, and [i] has a tongue body gesture. The V gesture is initiated halfway between the C1 (tongue tip) gesture of [s] and the C2 (bilabial and velum) gestures of [m], ruling out the possibility that the two C2 gestures are anti-phase coupled. Instead, the two C2 gestures are initiated in synchrony with each other. That is, the two C2 gestures act like one onset C gesture.



(a) A CCV syllable in English (b) A Tone-2 bearing CV syllable in Mandarin

Figure 7: Analogy in coupling relations between a Mandarin Tone-2 bearing CV syllable and an English CCV syllable (where C2 consists of two gestures). In 7a the C2b gesture is in-phase coupled to the C2a gesture only. Similarly in 7b the T2 (H) gesture is in-phase coupled to the T1 (L) gesture .

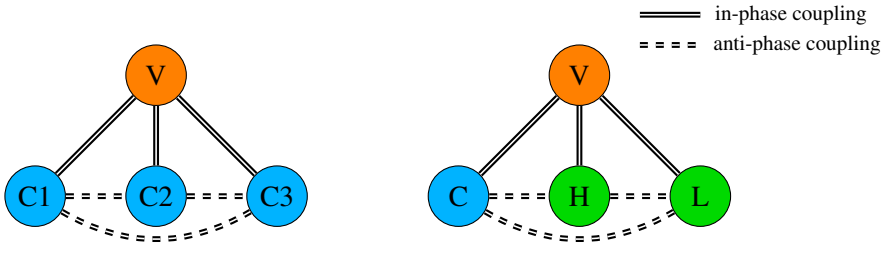


(a) A CCV syllable in English (b) A Tone2-bearing CV syllable in Mandarin

Figure 8: Analogy in gestural score between a Mandarin Tone2-bearing CV syllable and an English CCV syllable (C2 consists of two gestures). The V gestures are initiated halfway through in both 8a and 8b, because in 8a the two C gestures act as one C gesture and in 8b two T gestures act as one T gesture.

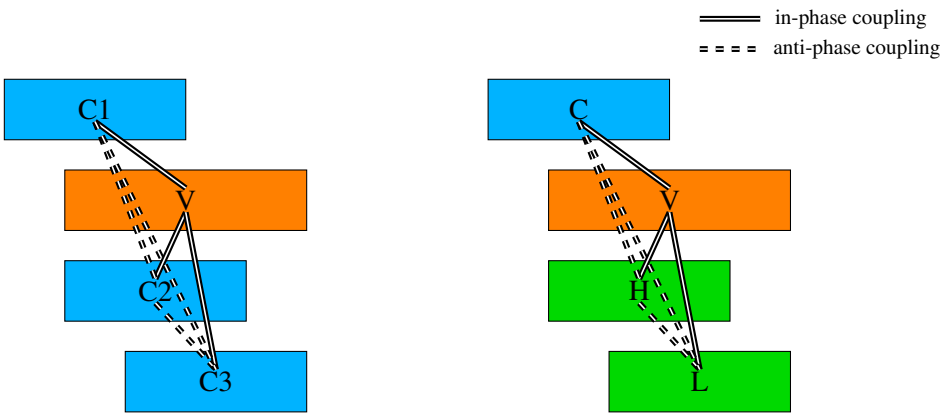
In the second group stands only Tone4 (51), the falling tone. Two T gestures, H and L, are anti-phase coupled to each other. Moreover, unlike in Tone2, both of the T gestures are coupled to the C and V gestures. This distinguishes Tone4 from the other three tones. The additional coupling introduced by the T2 (L) gesture of Tone4 changes the timing of the V gesture with respect to the C gestures in contrast to Tone2. The overall timing relationship between the center of the C/C-like gestures and the V gesture is still preserved. That is, the V gesture is initiated approximately at the same time as the T1 (H) gesture.

An analogy can be drawn between a Tone4-bearing syllable and a CCCV syllable in English, as shown in Figures 9 and 10. The onset cluster contains three consonants that are in an anti-phase coupling mode. Moreover, all of the three C gestures are in-phase coupled with the V gesture.



(a) A CCV syllable in English (b) A Tone4-bearing CV syllable in Mandarin

Figure 9: Analogy in coupling relations between a Mandarin Tone4-bearing CV syllable and an English CCCV syllable. In 9a all three C gestures are in-phase coupled to the V gesture and they are anti-phase coupled to the each other. Similarly in 9b the two T gestures and the C gesture are in-phase coupled to the V gesture, and the three C/C-like gestures are anti-phase coupled to each other.

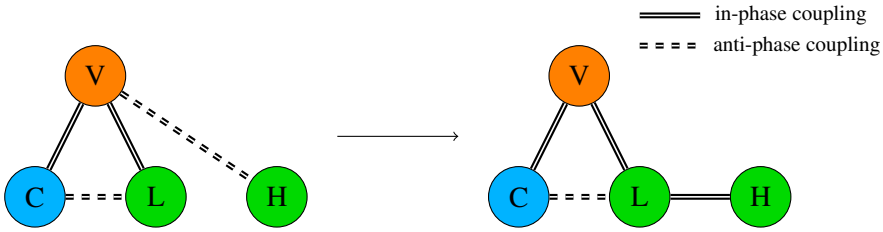


(a) A CCCV syllable in English (b) A Tone4-bearing CV syllable in Mandarin

Figure 10: Analogy in gestural score between a Mandarin Tone4-bearing CV syllable and an English CCCV syllable. The V gesture is initiated in synchrony with the C2 gesture in 10a; the V gesture is initiated in synchrony with the T1 (H) gesture in 10b.

3.3 An AP Account of Mandarin Tone Sandhi

Following Gao's (2008) proposal, Hsieh (2011) has proposed that T3S arises from the reorganization of the T gestures in Tone3. The H gesture of lexical Tone3 (falling-rising) acts as a coda consonant, which is anti-phase coupled to the V gesture only. During the application of third tone sandhi, the T2 (H) gesture of Tone3 undergoes a qualitative shift, 'advancing' to be in-phase coupled to the L gesture, resulting in T3S, as shown in Figure 11.



(a) Lexical Tone3 (b) T3S

Figure 11: Third tone sandhi based on Hsieh (2011). The H gesture is anti-phase coupled to the V gesture in its underlying form whereas it is in-phase coupled to the L gesture in T3S.

Despite being in the direction from anti-phase coupling to more stable in-phase coupling, the change is counterintuitive because it is unlikely that a coda C gesture would 'advance' to be in-phase coupled to the onset C gesture. Another question that arises out of Hsieh (2011) is that no distinction in the gestural score between Tone2 and T3S was proposed to account for acoustic differences between the two rising tones, as reported in previous work. However, Hsieh (2011) did not detail the difference between T3S and Tone2 in terms of gestural phasing (especially not jointly with segmental gestures). Moreover, she

only conducted acoustic experiments and therefore did not replicate the findings of Gao (2008) before accepting her proposal.

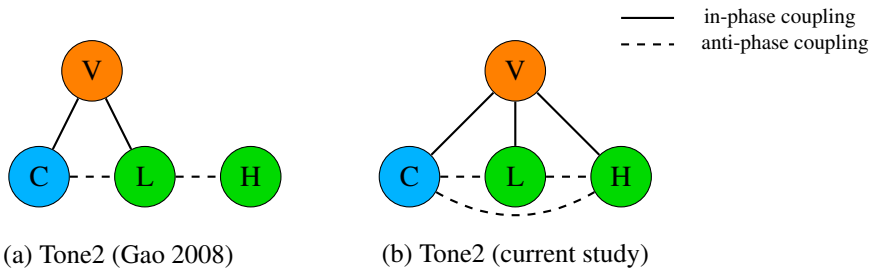
The current study focuses on the timing of the T gestures and segmental gestures in Tone3 variants and aims to offer explanations for the difference between T3S and Tone2 from an articulatory perspective.

4 Hypotheses and Predictions

First, unlike Gao (2008), we hypothesize that the second T gesture of Tone2 is coupled to the C, V, and L gestures, as illustrated in Figure 12.[3] Provided that T3H (a single L gesture) is the baseline of a C-center effect, this hypothesis predicts differences between Tone2 and T3H in terms of the relative timing of the C, V, and T gestures.

H_1: The T2 (H) gesture is coupled to the segmental gestures; the two T gestures (L and H) in Tone 2 act as two additional onset C gestures.

P_1: The V gesture is initiated after the midpoint between the C gesture and the T gesture in Tone2. The gestural organization of Tone2 will differ significantly from T3H in terms of relative timing of the C, V and T gestures.



(a) Tone2 (Gao 2008) (b) Tone2 (current study)

Figure 12: Comparison of coupling relations of a Tone2-bearing syllable between Gao's (2008) proposal (*left*) and the current study (*right*).

Second, we hypothesize that the underlying coupling structure of Tone3 influences that of the derived T3S. This predicts a difference between T3S and Tone2 in terms of the timing of the V gesture with respect to the C-center.

H_2: The underlying coupling structure of Tone3 influences that of the derived T3S; the T2 (H) in both tones acts as an additional C gesture.

P_2: T3S is different from Tone2 in terms of the relative timing of the V gesture with respect to the C-center.

5 Methodology

Participants

Four native speakers of Mandarin Chinese participated in the current study. They were born in Beijing, and were graduate students at Cornell University at the time of recording. From their own report, none of the participants suffered from any speech or hearing problems. Analyses of 2 female speakers (S1 and S2, hereafter) of Beijing Mandarin are

[3]Note that Figure 12 only represents the generic coupling relations, i.e. no specific coupling parameters are stipulated.

presented; analyses of the other two are not presented here because frequent use of creaky voice precludes reliable analysis of tone gesture timing.

Task

Four tone sequences were chosen as targets: Tone3 + Tone3, Tone2 + Tone3, Tone3 + Tone2, Tone3 + Tone4. The first tone of each disyllabic sequence was the target tone, while the second tone provided the conditioning environment for tone sandhi. In particular, the first two tone sequences resulted in a rising tone (i.e. T3S and Tone2), while the latter two resulted in a low tone (i.e. T3H) that preceded an L gesture (of Tone2) and an H gesture (of Tone4). Target sequences were embedded into a carrier sentence where the preceding tone was Tone2. This offers a contrasting tonal environment in that Tone2 ends with a high offset and all the target tones start with a low onset.

The chosen target syllables were restricted to syllables containing a labial consonant [m] and a low vowel [a]. The surrounding syllables, i.e. the syllable that precedes the target syllable and the syllable that follows the target syllable, contain coronal consonants [l] and [n], respectively, and a high front vowel [i]. The stimuli were chosen in consideration of the movement of articulators. In order to ensure a clear observation of articulatory movements that are associated with consonants and vowels, the flesh points corresponding to the articulators should be either located on different speech organs or be as far apart as possible if the same (bilabial stop [m] vs. vowel [a]). Moreover, the vowels in adjacent syllables must have contrastive tongue positions both vertically and horizontally (low back vowel [a] vs. high front vowel [i]).

Table 1 demonstrates the stimuli. The numerical values following the syllable correspond to Mandarin tones on a five-point scale (Chao 1968).

Preceding	*Target (Base)* + Conditioning	Example
Tone2	*T3S (Tone3)* + Tone3	li35] *ma213* ni213
	Tone2 (Tone2) + Tone3	li35] *ma35* ni213
	T3H (Tone3) + Tone2	li35] *ma213* ni35
	T3H (Tone3) + Tone4	li35] *ma213* ni51

Table 1: Stimuli used in the experiment: T3S + Tone3, Tone2 + Tone3, T3H + Tone2, and T3H + Tone4.

Following Gao (2008), focus on the target sequences (target tone + conditioning tone) was avoided by topicalizing the subject of the sentence with the following syntactically well-formed carrier sentences:[4]

ʂz̩51	wɔ213	jau51	ljəu35 li35	_____	i51 tɕia55
be	*I*	*want*	*Liu Li (Proper Name)*	_____	*one family*

'It is I that wants the whole family of Liu Li and _____.'

All the speech material was presented with most frequently used Chinese characters. Infrequent characters and homographic characters corresponding to more than one readings were avoided.

[4]Target tone sequences (in red) in the carrier are made-up names.

The experiment was coded using a MATLAB™ script. The screen was approximately 1.5 m away from the participant. For each trial, the visual word stimuli appeared alongside a red box moving from the bottom to the top of the screen. The box moved at one of the two programmed speeds: 1 second for fast speech and 2.5 seconds for slow speech. That is, it took 1 second for the red box to move from the bottom to the top of the screen in fast speech, and 2.5 seconds in slow speech. The participants were instructed to read out the sentence displayed on the screen at the indicated speech rate after the red box disappeared. The stimulus, as well as the speech rate of the stimuli, was presented in a random sequence.

Data processing and analysis

Articulatory data were collected using an NDI™ WAVE Electromagnetic Articulograph (EMA) at the Cornell Phonetics Lab in the Department of Linguistics at Cornell University. The system tracks real-time articulatory orofacial movements in speech production using small sensors attached to articulators, such as the lips, jaw, and tongue. For the purpose of this study, eight sensors were used in each experiment: three as reference points attached to the nasion, left and right mastoid process, one each to the jaw under the lower teeth (JAW), upper lip (UL) and lower lip (LL), the tongue tip (TT) and tongue body (TB). The TT sensor was placed about 1 centimeter posterior to the tip of the tongue, and the TB sensor was placed approximately 4-5 cm posterior to the TT sensor. Acoustic data were simultaneously collected at a sampling rate of 22.5 kHz.

Kinematic and F0 trajectories were extracted using MATLAB™. The gestures that were involved in the target syllable [ma] were a bilabial closure for [m] and tongue root retraction for [a]. The time course of these gestures was measured with lip aperture (henceforth LA), the vertical distance between the UL and LL, and tongue body height (henceforth TBy), the vertical displacement of the TB. For each trajectory of interest (i.e. LA and TBy), a corresponding velocity profile was computed to determine the articulatory landmarks: minimum velocity, onset, and peak velocity. Specifically, the onset was defined as the point when 20% of the velocity range between the minimum velocity and the peak velocity had passed. Similarly, the target was defined as the point when 80% of the velocity range between the peak velocity and the following (not the first one) minimum velocity had passed. F0 contours were extracted in MATLAB™, using a script developed by the Cornell Phonetics Lab. The script incorporates VOICEBOX, a third-party speech processing toolbox (Brookes 2005). F0 landmarks were consistently defined in the same way as kinematic landmarks.

Figure 13 illustrates the velocity-based landmarking in LA channel. The upper panel shows the LA trajectory and the lower panel shows the corresponding velocity profile (absolute value). The 20% threshold was used to determine the onset (blue) between the minimum velocity point (green) and the maximum velocity point (red).

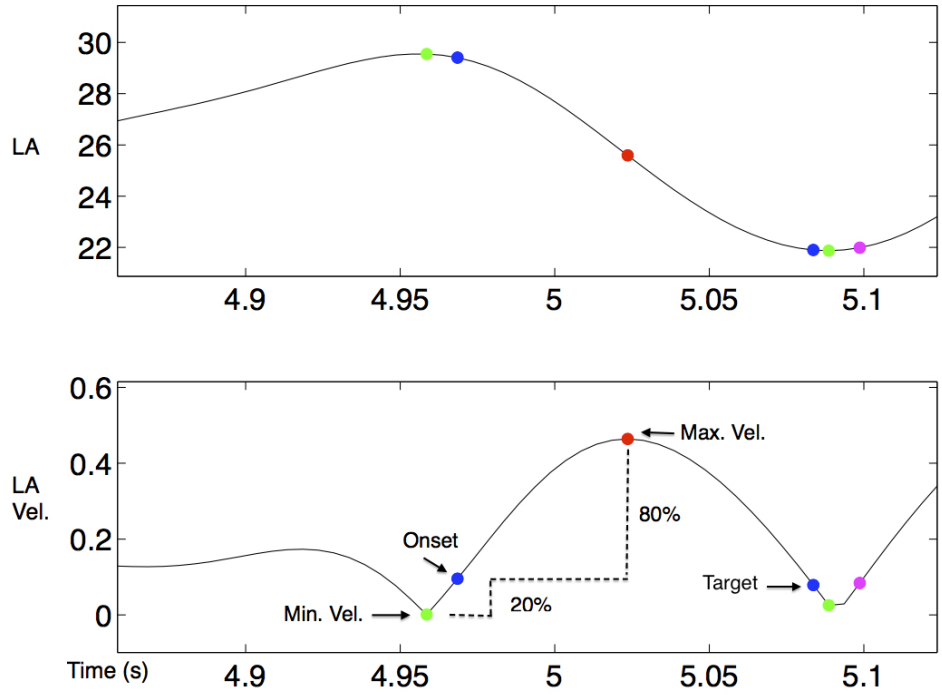


Figure 13: Illustration of velocity-based landmarking. The upper panel demonstrates the trajectory of the LA trajectory; the lower panel demonstrates the velocity (absolute value) of the trajectory at corresponding times with illustrations of the landmarks. Onset (blue) is defined as the point when 20% of the velocity range between the Min. Vel. (green) and the Max. Vel. (red) has passed.

After the landmarks on the kinematic and F0 trajectories were recorded, temporal lags between onsets of different channels were computed for further analysis, as shown in Figure 14. Specifically, CV lag is the temporal lag between the onset of the C gesture and the onset of the V gesture; VT1 lag is the temporal lag between the onset of the V gesture and the onset of the T1 gesture (L in T3H). Positive values of lags indicate that the first landmark precedes the second one, whereas negative values indicate that the first landmark follows the second one. Generally speaking, we would expect both the CV lag and the VT1 lag to be positive because the V gesture is initiated halfway between the C and T gestures (Gao 2008).

The phase of the V gesture relative to the CT1 lag (CV%) was further computed for each trial (CV% = CV onset lag / CT1 onset lag). The CV% is predicted to be 50% in the ideal C-center circumstance; a value larger than 50% indicates that the V gesture is initiated after the C-center, a value smaller than 50% indicates that the V gesture is initiated before the C-center.

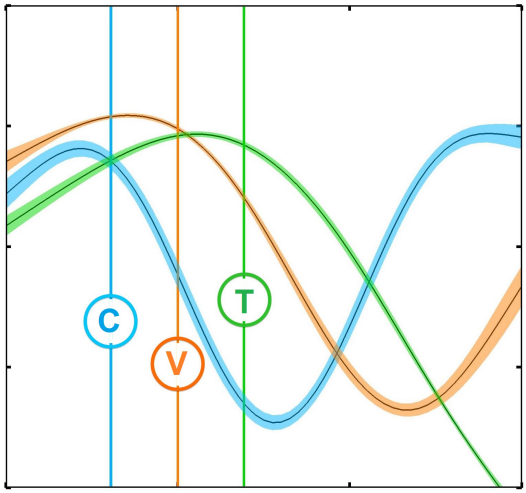


Figure 14: Normalized trajectory overlays of a T3H-bearing syllable. The vertical lines mark the onsets of the gestures. CV lag is the temporal lag between the onset of the C gesture and the onset of the V gesture; VT1 lag is the temporal lag between the onset of the V gesture and the onset of the T1 gesture. The C, V, and L gestures are coded in blue, orange, and green, respectively.

6 Results

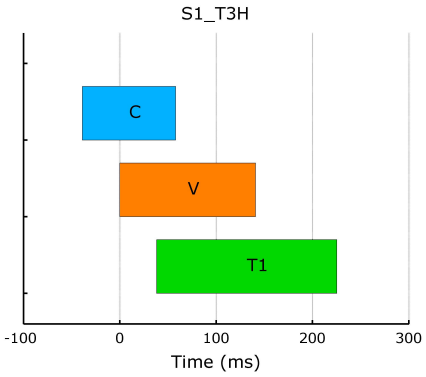
Patterns of articulatory timing between tones and oral segmental gestures show a significant difference between Tone2 and T3H; this supports the hypothesis that the additional H gesture in Tone2 (compared to the T3H) influences articulatory timing. Regarding the second hypothesis, the results were split: one speaker (S1) showed incomplete neutralization, consistent with the idea that the coupling structure of T3S is influenced by its relation to T3; the other speaker (S2) showed no significant difference.

6.1 Comparison between Tone2 and T3H

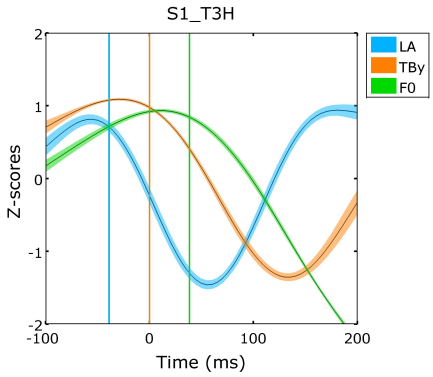
The simplicity of T3H renders it the baseline of a C-center effect in that T3H consists of a single L gesture and no additional competition is involved. Analyses of T3H show that the V gesture is initiated halfway between the C gesture and the L gesture for both speakers.

The gestural score and the overlay of normalized trajectories of a T3H-bearing syllable for both speakes are illustrated in Figure 15.[5] Gestural scores and normalized trajectories with onsets highlighted are displayed in a parallel fashion. Both the gestural score and the trajectories are aligned to the onset of the V gesture. Moreover, one standard error is plotted alongside the average contour for each channel in the overlay of normalized trajectories.

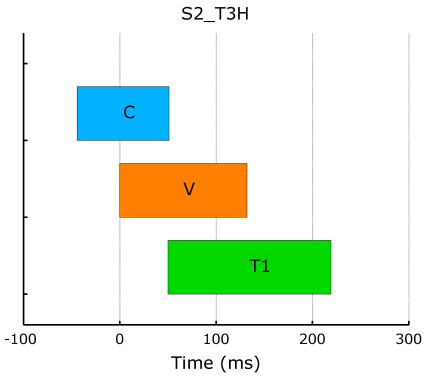
[5]The trajectory was normalized to a [0,1] scale for illustrative clarity, but the actual values were kept for statistical analysis.



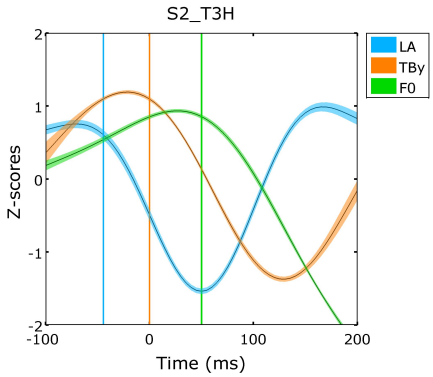
(a) Gestural score of T3H (S1)



(b) Normalized trajectories of T3H (S1)



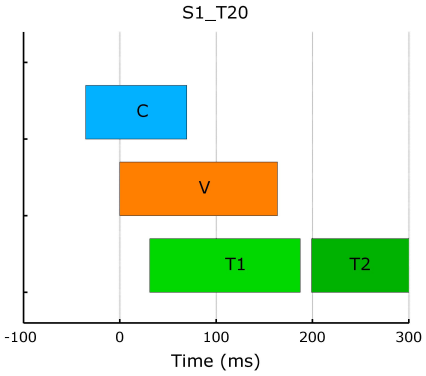
(c) Gestural score of T3H (S2)



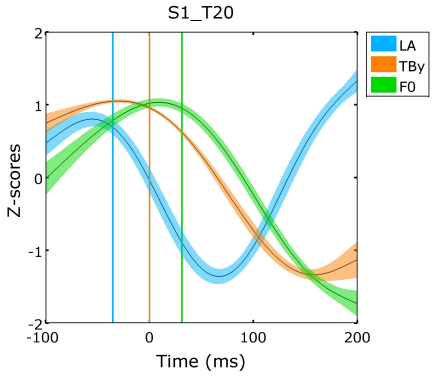
(d) Normalized trajectories of T3H (S2)

Figure 15: Timing of articulatory gestures in a T3H-bearing syllable

In contrast, Tone2 did not exhibit this pattern. The gestural score and the overlay normalized trajectories of a Tone2-bearing syllable for both speakers are illustrated in Figure 16.[6]

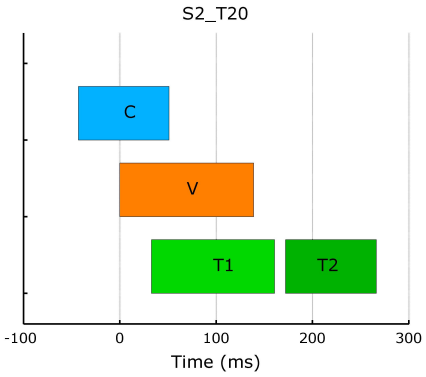


(a) Gestural score of Tone2 (S1)

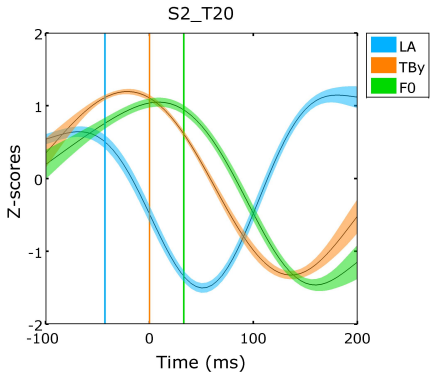


(b) Normalized trajectories of Tone2 (S1)

[6]Tone2 is referred to as 'T20' in the following plots.



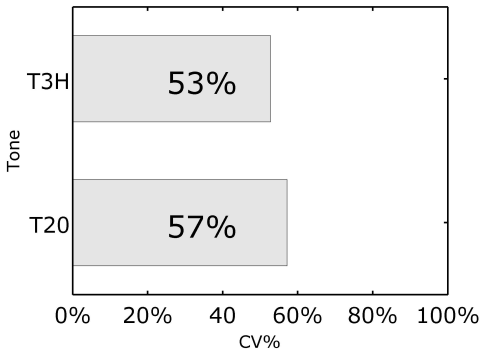
(c) Gestural score of Tone2 (S2)



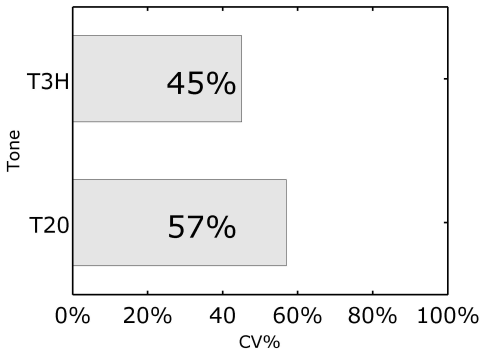
(d) Normalized trajectories of Tone2 (S2).

Figure 16: Timing of articulatory gestures in a Tone2-bearing syllable.

Figure 17 compares CV% between T3H and Tone2 for both speakers. For S1, the CV lag takes up 53% of the CT1 lag in T3H, whereas it takes up 57% in Tone2; for S2, the CV lag takes up 45% of the CT1 lag in T3H and it takes up 57% in Tone2. The difference in CV% between Tone2 and T3S is significant for S2 ($p < 0.0001$), whereas the difference is non-significant in S1 ($p = 0.2216$).



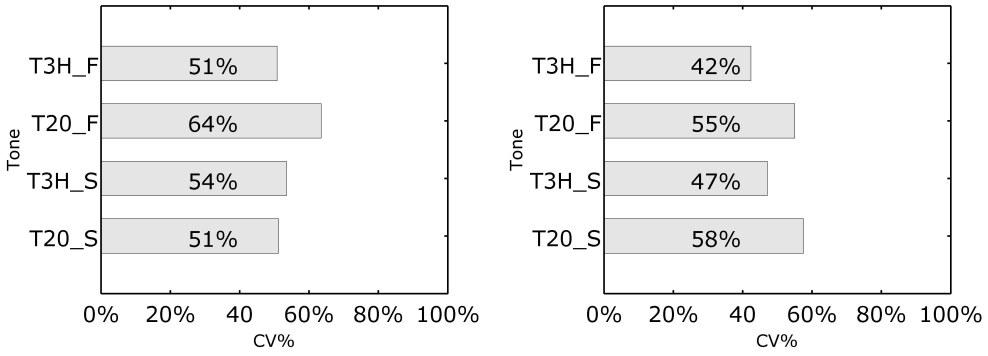
(a) S1



(b) S2

Figure 17: Comparison of CV% between T3H and T20 (by target) for S1 (*left*) and S2 (*right*). The difference in CV% between T3H and Tone2 is significant for S2, and marginally significant for S1.

However, a closer examination of S1's data shows that the CV% difference is marginally significant in fast speech (51% in T3H, 64% in Tone2, $p = 0.0613$). However, in slow speech, no significant difference in the CV% is observed between Tone2 and T3H (54% in T3H, 51% in Tone2, $p = 0.9667$). For S2, the CV% difference between Tone2 and T3H is significant in both fast and slow speech: $p = 0.0003$ in fast speech and $p = 0.0150$ in slow speech (Figure 18).



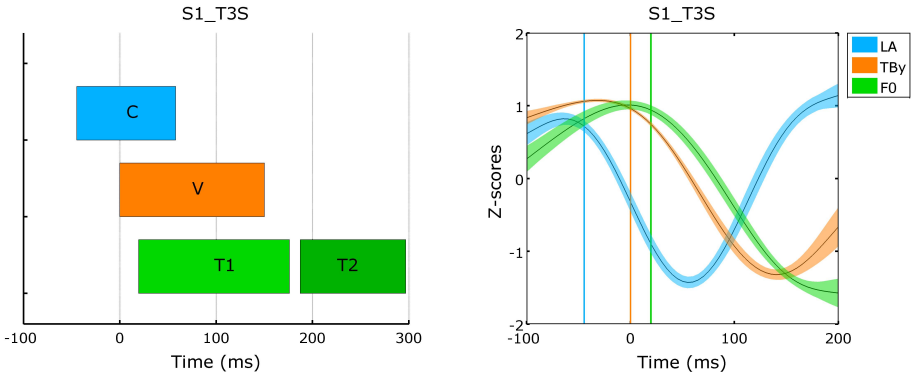
(a) S1 (b) S2

Figure 18: Comparison of CV% between CV% in T3H and Tone2 (by target and speed) for S1 (*left*) and S2 (*right*). The difference in CV% between T3H and Tone2 is significant in both fast and slow speech for S2 and in fast speech for S1, but not significant in slow speech for S1.

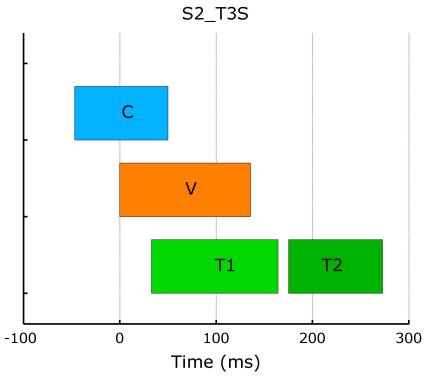
The above observation of Tone2 and T3H generally agrees with H_1. That is, both T gestures of the Tone2 should be treated as additional onset C gestures. The fact that the V gesture is closer to the T1 (L) gesture in Tone2 suggests that the T2 (H) gesture is coupled to segmental gestures, acting as a third consonant-like gesture.

6.2 Comparison between T3S and Tone2

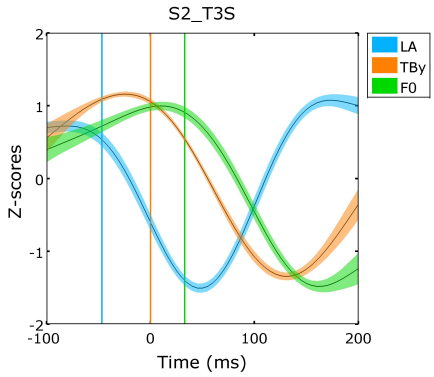
The gestural score and the overlay of the normalized trajectories of a T3S-bearing syllable for both speakers are illustrated in Figure 19. For S1, the V gesture is initiated closer to the onset of T1 (L) gesture in T3S, compared to Tone2. For S2, the VT1 lag of T3S approximately equals that of Tone2. Therefore, there exist two patterns of gestural organization in T3S: for S1, T3S is different from Tone2; for S2, T3S and Tone2 share the same gestural organization.



(a) Gestural score of T3S (S1) (b) Normalized trajectories of T3S (S1)



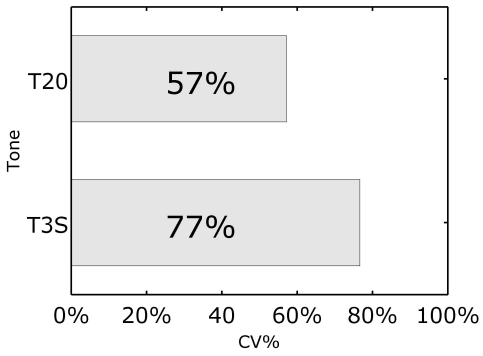
(c) Gestural score of T3S (S2)



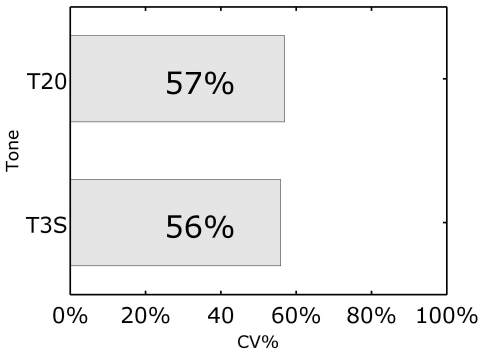
(d) Normalized trajectories of T3S (S2).

Figure 19: Timing of articulatory gestures in a T3S-bearing syllable

Figure 20 compares CV% between T3S and Tone2 for both speakers. For S1, the CV lag takes up 57% of the CT1 lag in Tone2, whereas it takes up 77% in T3S; for S2, the CV lag takes up 57% of the CT1 lag in Tone2 and it takes up 56% in T3S. Moreover, the difference in CV% between Tone2 and T3S is significant in S1 ($p = 0.0020$).



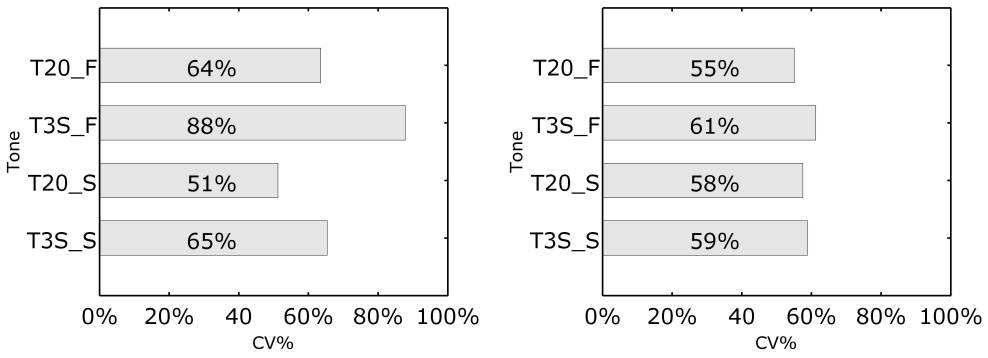
(a) S1



(b) S2

Figure 20: Comparison of CV% between Tone2 and T3S (by target) for S1 (*left*) and S2 (*right*). The difference in CV% between Tone2 and T3S is significant for S1 only.

A closer examination shows the CV% is larger in T3S than in Tone2 in both fast and slow speech. However, for S2, the difference in CV% between Tone2 and T3S is much smaller at both speech rates (Figure 21). The results of a two-way ANOVA confirm this: the target has a main effect on CV% for S1 but not for S2 ($p = 0.0012$ in S1, $p = 0.2860$ in S2). For both speakers there is no interaction effect.



(a) S1 (b) S2

Figure 21: Comparison of CV% between Tone2 and T3S (by target and speed) for S1 (*left*) and S2 (*right*). For S1, the CV% is significantly larger in T3S than in Tone2 at both speech rates; for S2, the difference is non significant at both speech rates.

The above result shows that for S1, the onset of the V gesture is initiated closer to the onset of the T1 (L) gesture in T3S, compared to in Tone2, whereas for S2 the VT1 lag of T3S is approximately the same as that of Tone2. Therefore, there exist two patterns of gestural organizations in T3S: for S1, T3S is different from Tone2; for S2, T3S and Tone2 share the same gestural organization.

7 Discussion and Conclusion

Patterns of the articulatory timing between tone gestures and segmental gestures indicate that the additional T gesture in Tone2, namely H, introduces more coupling interactions, therefore resulting in timing differences that distinguish it from T3H . Furthermore, the bias towards the underlying tone (lexical Tone3), which is further evidenced by the involvement of H, is responsible for an incomplete neutralization between Tone2 and T3S for one of the speakers.

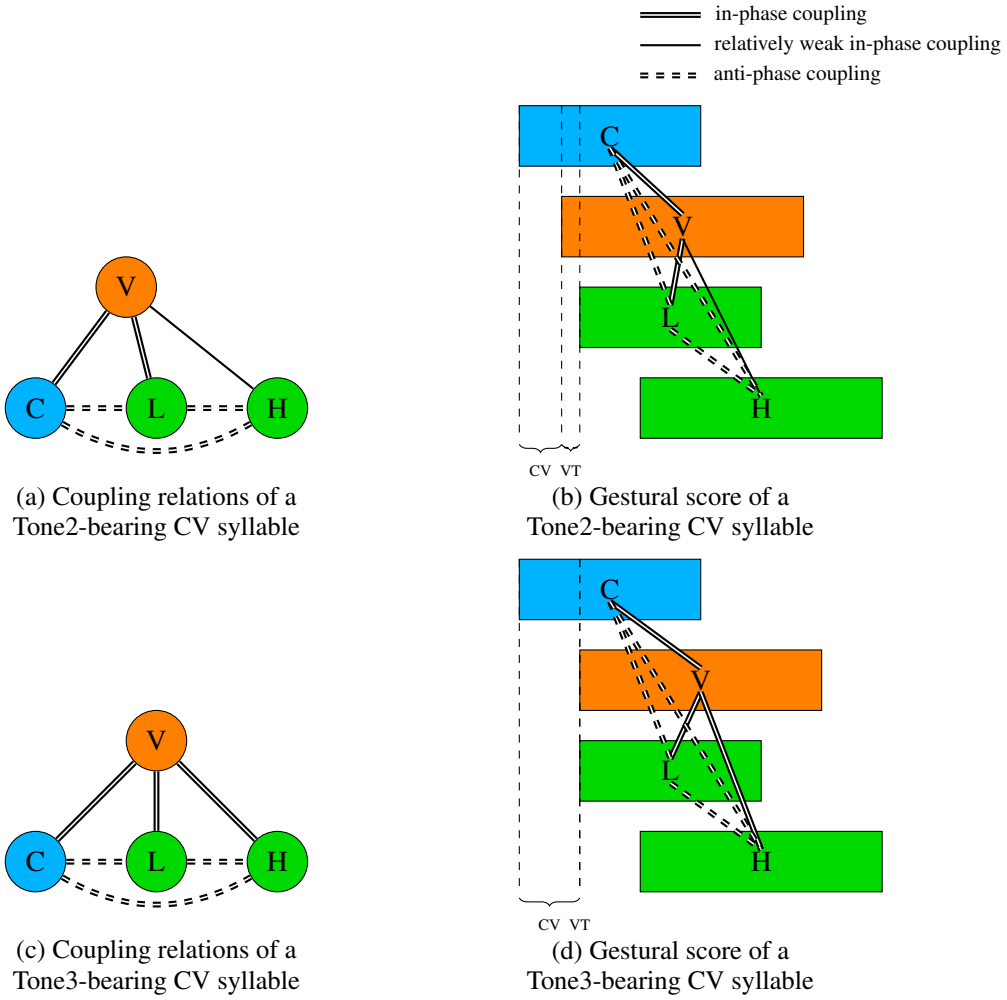
7.1 Modeling Tone2

The CV% of Tone2 is significantly larger than that of T3H for both speakers. That is, the onset of the V gesture is placed closer to the onset of the T1 (L) gesture in Tone2 than in T3H. The difference in gesture composition between Tone2 and T3H is that Tone2 has the T2 (H) gesture besides the shared T1 (L) gesture. This additional T gesture, i.e. T2 (L) introduces more couplings in syllables bearing Tone2.

An analogy can be drawn from Tone3. It is proposed that in Tone3, two T gestures, namely L and H, both act as onset C gestures.[7] Given that T gestures behave like C gestures, both T gestures together with the C gesture, act as onset C gestures that are anti-phase coupled to each other and in-phase coupled with V gesture. The collective force in a Tone3-bearing syllable renders a near synchronization between the V gesture and the T1 gesture. Put differently, the V gesture is initiated after the gestural midpoint between the C gesture and the T1 gesture due to the additional coupling interactions introduced by the T2 (H)

[7] Gao (2008) proposed a similar gestural composition for Tone4, where two T gestures, namely H and L, are both involved in the coupling interactions.

gesture in Tone3. Therefore, the onset of the V gesture is closer to the onset of the T1 (L) gesture in Tone3. Similarly, due to the additional coupling interactions introduced by the T2 (H) gesture in Tone2, the V gesture is initiated after the gestural midpoint between the C and T1 (L) gestures in Tone2.



(a) Coupling relations of a Tone2-bearing CV syllable

(b) Gestural score of a Tone2-bearing CV syllable

(c) Coupling relations of a Tone3-bearing CV syllable

(d) Gestural score of a Tone3-bearing CV syllable

Figure 22: Comparison of the coupling relations and gestural score between Tone2 and Tone3. The V gestures are initiated after the gestural midpoint in both tones. The difference is that the VT1 lag in Tone2 is larger than that in Tone4 to the extent that the V gesture is in synchrony with the T1 (H) gesture in Tone3. Double solid lines indicate stronger coupling strength (in-phase) than single solid lines within the coordinative unit.

However, as shown in Figure 22, the extent to which the V gesture is close to the T1 (L) gesture is different between Tone3 and Tone2: in Tone3, there is a near-synchronization between the V gesture and the T1 (L) gesture, whereas the onset of the V gesture is closer to the gestural midpoint than to the onset of T1 (L) gesture, meaning there is still a temporal lag between the V gesture and the T1 (L) gesture. Therefore, the T2 (H) gesture should be regarded as active in coupling but not having as strong of a coupling with the V gesture as the T1 (L) gesture in Tone2. Put differently, the V gesture in Tone2 maintains an in-phase coupling relation with the T2 (H) gesture, and the T2 (H)-V coupling is weaker in strength

compared to the T1 (L)-V coupling and C-V coupling, thus the difference in CV% between Tone2 and Tone3.

Alternatively, such differences could also arise out of the relatively strong anti-phase coupling relation wherein the T2 (H) gesture is involved. That is, the coupling of C-T2 (H) and T1 (L)-T2 (H) is stronger than C-T1 (L). However, this hypothesis seems somewhat counterintuitive in that the T2 (H) gesture bears a stronger coupling relation with the C gesture and the T1 (L) gesture given that it is a more peripheral gesture in the sense of sequence. Treating T2 (H)-V as having the same coupling strength as T1 (L)-V also seems counterintuitive in that regard.

Another possible alternative is that the differences arise out of the systematic differences in landmarking between different tones. Thus, the significant difference in CV% across tones is a result of an artifact of the landmarking. As for the comparison between T3H and Tone2, which have different pitch contours, the landmarking for the low tone (i.e. T3H) might produce certain systematic errors, given that low tones in Mandarin are notorious for their 'creak'. Thus it might be difficult to track the pitch contours.

7.2 Modeling T3S

The CV% of T3S is larger than that of Tone2 in both fast and slow speech for S1. The V gesture is initiated closer to the T1 (L) gesture in T3S, inducing incomplete neutralization between T3S and Tone2. However, for S2, T3S and Tone2 are in complete neutralization in the sense that the difference in CV% is not significant.

Given that both T gestures are active in coupling in T3S and Tone2, the evidence that the V gesture is initiated closer to the T1 (L) gesture in T3S than in Tone2 is suggestive of a stronger coupling of T2 (H)-V in T3S than in Tone2 (The stronger the T2 (H)-V coupling, the smaller the VT1 lag, the larger the CV%). Also note that the V gesture of Tone2 is initiated closer to the T1 (L) gesture in Tone2 than in T3H for both speakers, because there is no coupling, i.e. zero coupling force, between the V gesture and the T2 (H) gesture in T3H. Hence, a gradual decrease in CV% across the three tones is observed in S1's data.

A few questions remain to be answered. Why does T3S behaves quite like Tone2, but at the same time distance itself from Tone2? What is responsible for the incomplete neutralization? How does third tone sandhi come into play?

Figure 23 shows the derivations of T3H and T3S in half third sandhi and third tone sandhi, respectively. In producing T3H, speakers dissociate the coupling between the T2 (H) gesture and the V gesture, from in-phase coupled to uncoupled, giving rise to T3H. The T2 (H) is co-selected with the T1 (L) gesture, to which T2 (H) is anti-phase coupled to. The T2 (H) gesture does not bear any coupling relations with the segmental gestures. The T2 (H) gesture not surfacing might result from the lack of a second timing unit that it might anchor to. In Mandarin, it is often argued that there is only one tone bearing unit — the syllable itself. The T2 (H) gesture does surface as H when the Tone3 bearing syllable is followed by a toneless syllable (syllable with Tone0). Thus, the dissociation of T2 (H) during the half third sandhi renders the Mid-to-V lag of T3H to fluctuate around 0 ms.

In producing T3S, speakers cannot fully decohere the coupling when Tone3 is followed by another Tone3 due to a higher phonological constraint that disallows two L gestures from surfacing next to each other (*LL)(Yip 2002).[8] To observe the higher-level constraint, speakers will adjust the T2 (H)-V coupling to assimilate to Tone2 to preserve the structure

[8]Despite the fact that a constraint like *LL cannot be incorporated into the framework of AP in a satisfactory manner, we suggest that it is functioning at a higher level before the gestures enter into the coordinative systems.

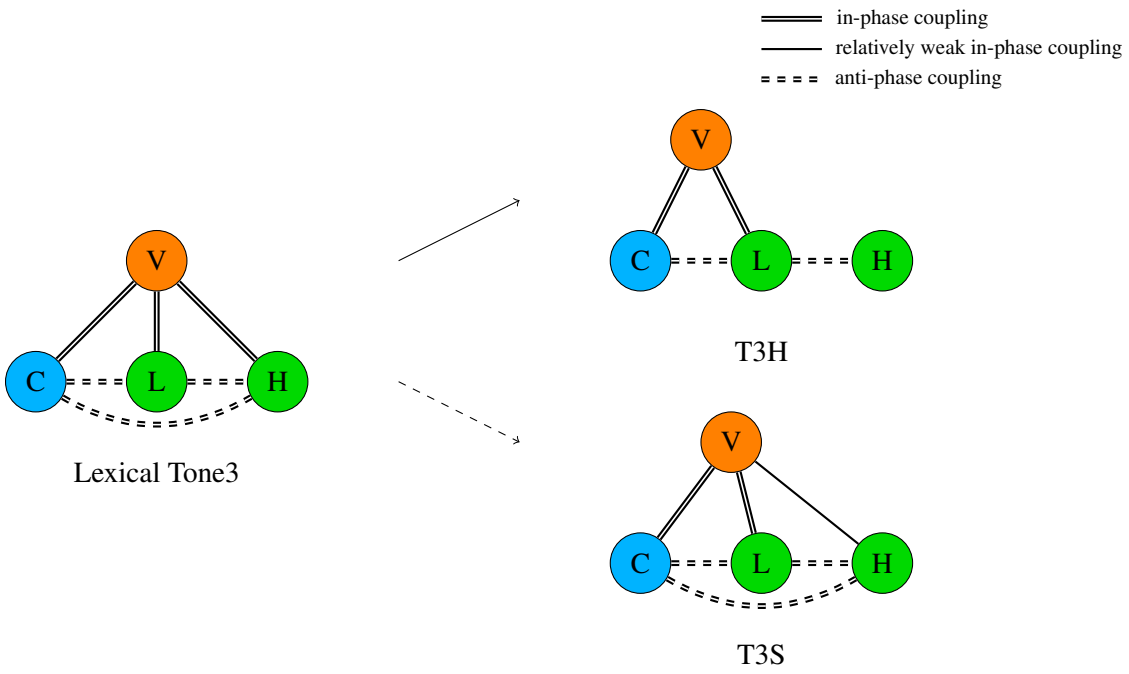


Figure 23: Proposed model of both half tone sandhi and third tone sandhi. The H gesture acts as an additional onset gesture in the underlying form of Tone3. Double solid lines indicate stronger coupling (in-phase) than single solid lines within the coordinative unit. Dashed arrow indicates potential task-specific bias (see text below).

of the Mandarin tone inventory, i.e. to avoid creating a new tone category. Two T gestures have been selected and sequenced before they are further coordinated. Speakers have to perform the sandhi with what they have in hand without making drastic structural changes. In other words, they have to make adjustments to the coupling interactions rather than to the selection or the sequence of the gestures. For example, the output of the third tone sandhi could not be a Tone4 because that would require the reversal of the sequence of the two T gestures, namely from L-H to H-L. Therefore, in third tone sandhi, Tone3 is assimilated to Tone2 rather than to Tone1 or Tone4.

During the assimilation to Tone2, speakers might be biased by Tone3, the underlying tone. Specifically, the coupling strength between the T2 (H) gesture and the V gesture in T3S might not be adjusted to be exactly the same as in Tone2, but instead still resembles that in Tone3 or simply lies between Tone2 and Tone3. The bias towards the underlying tone (i.e. Tone3) is the source of the incomplete neutralization between T3S and Tone2.

Here we view Tone2 and Tone3 (lexical) as two distinct tone categories that stand out on the spectrum of tones that are sequentially composed of both L and H gestures. AP states that dynamic parameters like constriction degree and relative phasing do not inherently define categorically distinct classes. It is assumed that there are stable ranges of parameters that tend to contrast with one another repeatedly in languages to form categories (Browman and Goldstein 1992). Thus we assume that the relative phasing in Tone2 and Tone3 are stable enough to contrast with one another, despite both of them being composed of both L and H gestures. That is, native speakers of Mandarin can utilize these stable ranges of parameters, such as the strength of coupling, to contrast Tone2 and Tone3. During the

application of third tone sandhi, speakers adjust the phasing principles, i.e. the coupling strength, to assimilate Tone3 to Tone2, for the sake of preserving the tone inventory of the language.

The effect of the bias towards Tone3 is confirmed by the evidence that T3S is still stored in the mental lexicon as Tone3, despite being assimilated to Tone2 (Chen et al. 2011). Under the framework of the form preparation paradigm, the sequences Tone3 + Tone3 and Tone3 + Tone2 were found to show a preparation effect of both segment and tonal sharing, while Tone3 + Tone3 and Tone2 + Tone3 only showed a preparation effect of segment sharing. Therefore, T3S and T3H share the same underlying form, i.e. lexical Tone3, in the mental lexicon. That is, for both T3S and T3H, both the L gestures and the H gesture are selected and sequenced before entering coordination.

However, the non-neutralized T3S is produced in a task-specific manner, i.e. T3S is only gesturally different from Tone2 in a subset of productions. Different speakers may also vary in their strategies of producing T3S. Therefore, distributionally speaking, T3S might be a weighted average of Tone2 and Tone3 in terms of gestural phasing. This explains why the difference between T3S and Tone2 in S1 is significant whereas S2 displays no such difference between T3S and Tone2.

Characterizing gestures as the primitive phonological units allows AP to capture both categorical and gradient information. In our case, the distinction between Tone2 and Tone3 (and T3H) is categorical distinction because they contrast in the underlying 'input' structures. Therefore, the categorical distinctions are made by 'turning on' or 'turning off' certain gestures, or setting parameters within a stable range that makes contrasts in a particular language. The distinction between Tone2 and T3S falls into the more gradient scale. The gradient difference can be captured by quantitative variation in the 'input' parametric specification of gestural organization, such as the phasing principles. For S1, such variations in the underlying structure stand out to be significant, whereas no such differences were observed for S2. Therefore, it is likely that speakers have learned different coupling structures for derived patterns such as the sandhi tone, and so collecting data from more speakers will help shed light on the range of such patterns.

7.3 Conclusion

This paper argues that the Tone2 articulatorily differs from T3H because the additional coupling interactions introduced by the T2 (H) gesture, which has been ignored by previous work. More importantly, the T2 (H) gesture plays an important role by offering explanations for the incomplete neutralization between Tone2 and T3S, which can be further attributed to a bias towards the underlying Tone3.

A key contribution of the current study is the finding that phonological patterns typically thought to be categorical, such as Mandarin tone sandhi, can exhibit sub-categorical, gradient differences in articulatory timing. It is important for models of speech production to be able to accommodate and explain such variation. We have proposed that gradient differences in gestural coupling strength parameters (perhaps resulting from interactions between underlying and derived variants) are a possible source of the variation. We have also suggested that these coupling strength parameters can differ across speakers. Future studies are called for to look into the phonological status of the difference between T3S and Tone2 as well as the inter-speaker variations from an articulatory perspective.

References

Brookes, Mike. 2005. VOICEBOX: Speech Processing Toolbox for MATLAB. Web page.

Browman, C. P., and L. Goldstein. 1989. Articulatory Gestures as Phonological Units. *Phonology* 6:201–251.

Browman, Catherine P., and Louis Goldstein. 1990. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics* 18:299–320.

Browman, Catherine P., and Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49:155–180.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, Yiya, Rachel Shen, and Niels Schiller. 2011. Representation of allophonic tone sandhi variants. In *In Proceedings of AMLaP 2011*.

Chen, Yiya, and Jiahong Yuan. 2007. A corpus study of the 3rd tone sandhi in standard chinese. In *In Proceedings of Interspeech 2007*.

Gao, Man. 2008. Mandarin Tones: An Articulatory Phonology account. Doctoral dissertation, Yale University.

Hsieh, Fang-Ying. 2011. A gestural account of mandarin tone 3 variation. In *In Proceedings of ICPhS XVII*.

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

Peng, Shu H. 2000. Lexical versus 'phonological' representations of Mandarin sandhi tones. In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, ed. M. B. Broe and Janet B. Pierrehumbert, 152–167. Cambridge, UK: Cambridge University Press.

Yip, Moira. 2002. *Tone*. Cambridge, UK: Cambridge University Press.

Zhang, Jie, and Yuwen Lai. 2010. Testing the role of phonetic knowledge in mandarin tone sandhi. *Phonology* 27:153–201.

Department of Linguistics
Cornell University
Ithaca, NY 14853
hy433@cornell.edu