Gestural timing in Mandarin tone sandhi

Hao Yi  
*Linguistics; Cornell University; Ithaca, NY; hy433@cornell.edu*

Sam Tilsen  
*Linguistics; Cornell University; Ithaca, NY; tilsen@cornell.edu*

Recently tones have been analyzed as articulatory gestures, which can be coordinated with segmental gestures (Gao, 2008). Using Electromagnetic Articulometry (EMA), this paper investigates the timing patterns in Tone3 variants under the framework of Articulatory Phonology (AP). Mandarin third tone sandhi (Tone3 → T3S / ___ Tone3) is purportedly neutralizing in that the sandhi output (T3S) is acoustically very similar to Tone2. Despite that they have similar F0 contours, there exist subtle acoustic differences. However, the differences in the underlying representation between T3S and Tone2 remain unclear. Analyses of the timing of tonal gestures and articulatory gestures show that purported neutralized phonological contrast between T3S and Tone2 can nonetheless exhibit timing differences in Mandarin tone sandhi. Furthermore, it is likely that interactions between underlying and derived variants are the source of the incomplete neutralization.
1 Introduction

Mandarin third tone sandhi (i.e. Tone3 → T3S / ___ Tone3) is purportedly neutralizing in that T3S is acoustically very similar to Tone2. Figure 1 shows the F0 contours of Tone2, T3S (output of third tone sandhi), and T3H (surface form of Tone3 preceding a non-Tone3 in connected speech). Both Tone2 and T3S are rising tones, whereas T3H is a low tone. Despite the similarity between Tone2 and T3S, there exist subtle differences. Previous work has investigated the differences from an acoustic perspective. For example, the peak of the T3S contour does not reach as high as that of Tone2 (Chen and Yuan, 2007). However, the differences in underlying representation between T3S and Tone2 remains unclear. The aim of this paper is to offer an explanation for such differences from an articulatory perspective by investigating the timing of tonal gestures and articulatory gestures in Mandarin tone sandhi.

Under the framework of Articulatory Phonology (AP), articulatory gestures are proposed to be basic units of phonological structure (Browman and Goldstein, 1989, 1990, 1992). Gestures are one-dimensional dynamical systems that are associated with target values of vocal tract geometry that are achieved by the coordinated movements of articulators. For example, a bilabial closure gesture is associated with three articulators, namely the upper lip, the lower lip, and the jaw; movements of these articulators are coordinated so as to achieve a negative value of the tract variable, in this case the lip aperture.

Two or more gestures can be organized together in a specific way to form larger structures. There are two preferred ways in which a pair of gestures can be coupled: in-phase and anti-phase. According to hypothesis, an onset consonant (henceforth C) gesture is in-phase coupled to the vowel (henceforth V) gesture whereas the C gestures in an onset cluster are anti-phase coupled to each other. Therefore, in an English CCV syllable, the collective force of the coupling relations results in a pattern in which the onsets of the two 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).

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, and they behave like onset C gestures. Therefore, a C-center effect emerges in a Mandarin T3H-bearing CV syllable: the V gesture is activated halfway between the onsets of the C gesture and the T (L in this case) gesture, as illustrated in Figure 2.
in-phase coupling

anti-phase coupling

Figure 2: Analogy in the coupling relations (top) and the gestural score (bottom) between an English CCV syllable and a Mandarin T3H-bearing CV syllable. The T (L) gesture of T3H behaves like an onset C gesture. Both syllables display a C-center effect.

Gao (2008) also found that Tone2, hypothesized to be a rising tone composed of two T gestures (L and H), displays a C-center effect. Instead of being activated in a sequential fashion, the two T gestures are activated in synchrony with each other and function as one additional onset C gesture. As a result, Tone2 is no different from T3H in terms of the relative timing of the V gesture with respect to the C-center. An analogy can be drawn between a Tone2-bearing syllable in Mandarin and, for example, the CCV syllable smee in English, where [s] has a tongue tip gesture, [m] has two C gestures, namely a bilabial gesture and a velum gesture, and [i] has a tongue body gesture. The V (tongue body) gesture of [i] is activated halfway between the C1 (tongue tip) gesture of [s] and the C2 (both bilabial and velum) gestures of [m]. This rules out the possibility that the two C2 gestures are anti-phase coupled to each other. Instead, the two C2 gestures are activated in synchrony with each other; that is, they act like one onset C gesture.

Following Gao (2008), Hsieh (2011) proposed that T3S arises out of the re-organization of the T gestures in lexical Tone3. In this view, lexical Tone3 is a falling-rising tone and its H gesture acts as a coda consonant. That is, the H gesture is only anti-phase coupled to the V gesture, but bears no coupling relation to the C gesture or the L gesture. During the application of third tone sandhi, the H gesture of Tone3 somehow undergoes a qualitative shift, ‘advancing’ to be in-phase coupled to the L gesture, resulting T3S. However, one problem with Hsieh’s proposal is that the change in phasing is quite drastic: the analogous change of a coda C gesture becoming in-phase coupled to an onset C gesture is typically not observed in phonological alternations. Another issue that arises 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.

The current study investigates the timing of the tonal gestures and segmental gestures in Tone3 variants and aims to offer explanations for the difference between T3S and Tone2 from an articulatory perspective. First, in contrast to Gao (2008; cf. 3a), we hypothesize that the H gesture of Tone2 is coupled to the C, V,
and L gestures, as illustrated in Figure 3b.\(^1\) Provided that T3H (a single L gesture) is the baseline of the C-center effect, this hypothesis predicts differences between Tone2 and T3H in terms of the relative timing of the C, V, and T gestures. 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.

![Figure 3: Comparison of the coupling relations of a Tone2-bearing syllable between Gao’s (2008) proposal (left) and the current study (right).](image)

### 2 Method

#### 2.1 Task and participants

Four native speakers of Mandarin Chinese participated in the current study. They were born and raised in Beijing, and were graduate students of 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 presented.\(^2\)

There were four di-tone sequences: 1. Tone3 + Tone3, which produces a rising tone T3S followed by Tone3; 2. Tone2 + Tone3, which does not induce tone sandhi; 3. Tone3 + Tone2, which produces a low tone T3H followed by an L gesture; 4. Tone3 + Tone4, which produces a low tone T3H followed by an H gesture. The first tone of each di-tone sequence is the target tone, while the second tone provides the conditioning environment for tone sandhi. All target sequences were preceded by Tone2 in the same carrier sentence. This offers a contrasting tonal environment since Tone2 ends with a high offset and all the target tones start with a low onset.

<table>
<thead>
<tr>
<th>Di-tone sequence</th>
<th>Target tone</th>
<th>Tone3 + Tone3</th>
<th>Tone2 + Tone3</th>
<th>Tone3 + Tone2/Tone4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T3S</td>
<td>T20</td>
<td>T3H</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Stimuli used in the experiment — 1. Tone3 + Tone3; 2. Tone2 + Tone3; 3. Tone3 + Tone2; 4. Tone3 + Tone4. The last two di-tone sequences are lumped together because they display similar results.

The target syllable contained a labial consonant [m] and a low vowel [a], and was preceded by [li] and followed by [ni] (which bore the second tone of the di-tone sequence). The adjacent syllables each contained a coronal sonorant ([l] or [n]) and a high front vowel [i]. This ensures a clear observation of

\(^1\)Note that Figure 3 only represents the generic coupling relations, i.e. no specific coupling parameters are stipulated.

\(^2\)Data of the other two speakers are not presented here because frequent use of creaky voice precludes reliable analysis of tone gesture timing.
articulatory movement that are associated with consonants and vowels. The speakers were told that the
disyllabic sequences (i.e. [ma.ni]) were made-up names. Following Gao (2008), focus on the target syllable
was avoided by topicalizing of the subject in the following syntactically well-formed carrier sentence:³

\[ \text{sz}^4 \ w\text{ɔ}^3 \ \text{jau}^4 \ \text{lj\text{ɔ}u}^2 \ \text{li}^2 \ \ldots \ \text{i}^4 \ \text{tcia}^1 \]

\[ \text{be} \ I \ \text{want} \ \text{Liu Li (Proper Name)} \ \ldots \ \text{one family} \]

‘It is I that want the whole family of Liu Li and \ldots’

The speech material was presented on a monitor, which was approximately 1.5 m away from the speaker.
The experiment was organized into blocks of 16 trials. Both speakers completed 20 blocks. On 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: fast and slow. It took 1 second for the fast-moving
box to move from the bottom to the top of the screen, whereas 2.5 seconds for the slow-moving box. The
speakers were instructed to read out the sentence on the screen at the indicated speech rate after the red box
disappeared. The stimuli, as well as the speech rate, were presented in random order.

2.2 Data processing and data analysis

Articulatory data were collected by an NDI WAVE Electromagnetic Articulograph (EMA). For the purpose
of this study, eight sensors were used in each experiment: three as reference points attached to the nasion,
the left, and the right mastoid process, one each to the jaw under the lower teeth (JAW), the upper lip (UL),
the lower lip (LL), the tongue tip (TT), and the tongue body (TB). The TT sensor was placed about 1 cm
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 in MATLAB. The gestures that were involved in the target
syllable [ma] were a bilabial closure of [m] and tongue root retraction of [a]. The bilabial closure was
measured with the time course of the lip aperture (henceforth LA), the vertical distance between UL and LL;
the tongue root retraction was measured with the time course of the tongue body height (henceforth TBY),
the vertical displacement of 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. The F0 contours were extracted 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.

After the landmarks on the kinematic and F0 trajectories were recorded, temporal lags between onsets
of different trajectories were computed for further analysis. As shown in Figure 4, the CV lag is defined as
the temporal lag between the onset of the C gesture and the onset of the V gesture; the VT1 lag is defined
as 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 expect both the CV lag and the VT1 lag
to be positive because the V gesture is activated halfway between the C and T gestures (Gao, 2008).

³The numerical value following the syllable corresponds to Mandarin tone number.
Figure 4: Overlay of the normalized trajectories in a T3H-bearing syllable. The vertical line marks the onset of the gesture. The C, V, and L gestures are coded in blue, orange, and green, respectively.

The phase of the V gesture relative to the CT1 lag (henceforth 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; values larger than 50% indicate that the V gesture is activated after the C-center, whereas values smaller than 50% indicate that the V gesture is activated before the C-center.

3 Results

Patterns of the articulatory timing between tonal gestures and segmental gestures show a significant difference between Tone2 and T3H; this supports the hypothesis that the additional H gesture in Tone2 (compared to T3H) influences articulatory timing. Regarding the second hypothesis, the results were split: S1 showed incomplete neutralization, consistent with the idea that the coupling structure of T3S is influenced by its relation to Tone3; S2 showed no significant difference between T3S and Tone2.

3.1 Comparison between Tone2 and T3H

The simplicity of T3H — it consists of a single L gesture — renders it the baseline of the C-center effect. Analyses of T3H show that the V gesture is activated approximately halfway between the C gesture and the L gesture for both speakers in the current experiment. However, Tone2 is significantly different from T3H in terms of the relative timing of the C, V, and L gestures. Figure 5 shows the CV% of Tone2 and T3H for both speakers.
Figure 5: Comparison of the CV% between T3H and Tone2 (T20 in above figures) for S1 (left) and for S2 (right). The CV% difference between T3H and Tone2 is significant for S2, and marginally significant for S1.

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, whereas it takes up 57% in Tone2. The difference in CV% between Tone2 and T3S is significant for S2 (t(103) = 3.46, p < 0.001), but non-significant for S1 (t(109) = 1.18, p < 0.23 n.s.). However, a closer examination of S1’s data shows that the CV% difference is still marginally significant in fast speech (51% in T3H, 64% in Tone2, t(52) = 1.89, p < 0.07), but non-significant in slow speech (54% in T3H, 51% in Tone2, t(55) = – 0.41, p < 0.68 n.s.). For S2, the CV% difference between Tone2 and T3H is significant at both speech rates (t(48) = 3.23, p < 0.004 in fast speech; t(53) = 2.18, p < 0.04 in slow speech).

The above observations of Tone2 and T3H generally agree with the first hypothesis — the T2 (H) gesture is coupled to segmental gestures, acting as a third onset C-like gesture (even though so far we cannot explain the non-significant difference between Tone2 and T3H in S1’s slow speech).

3.2 Comparison between T3S and Tone2

Figure 6 displays the comparison of the 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. The CV% difference between Tone2 and T3S is significant for S1 (t(67) = 2.27, p < 0.03), but is non-significant for S2 (t(63) = 0.04, p < 0.98 n.s.). However, a closer examination of S1’s data shows that the CV% difference is significant in fast speech (t(31) = 2.08, p < 0.05), but not significant in slow speech (t(34) = 1.19, p < 0.25 n.s.). For S2, the CV% difference is non-significant at both speech rates (t(33) = 0.02, p < 0.98 n.s. in fast speech; t(28) = – 0.14, p < 0.89 n.s. in slow speech).

The above results show that for S1, the V gesture is activated closer to the onset of T1 (L) gesture in T3S, compared to in Tone2, whereas for S2, the CV% of T3S is approximately the same as that of Tone2. Therefore, there exist two patterns of coupling structures in T3S: for S1, T3S is articulatorily different from Tone2; for S2, T3S and Tone2 share the same coupling structure.
4 Discussion

Patterns of the articulatory timing between tonal gestures and segmental gestures indicate that the H gesture in Tone2 introduces more coupling interactions, therefore resulting in timing differences that distinguish Tone2 from T3H. Furthermore, the bias towards the underlying tone (lexical Tone3), which can be further explained by the coupling interactions introduced by the T2 (H) gesture, is responsible for the incomplete neutralization between Tone2 and T3S for S1.

4.1 Modeling Tone2

The comparison of the CV% between Tone2 and T3H indicates that the same coupling structure is not shared by the two tones: the V gesture is activated closer to the T1 (L) gesture in Tone2 than in T3H for both speakers. The difference in coupling structure between Tone2 and T3H is that the former contains a T2 (H) gesture besides the shared T1 (L) gesture. We argue that the additional T2 (H) gesture is coupled to the T1 (L) gesture as well as the C gesture and the V gesture, introducing more coupling interactions in a Tone2-bearing syllable.

An analogy can be drawn between lexical Tone3 and Tone2. We propose that Tone3, similar to Tone2, consists of two T gestures, namely L and H; both T gestures act as onset C gestures (cf. Figure 7a).4 Both T gestures act as onset C gestures. Thus, both T gestures and the onset C gesture are anti-phase coupled to each other, and they are in-phase coupled to V gesture. The collective force in a Tone3-bearing syllable renders a near synchronization of the V gesture and the T1 gesture (the CV% is close to 0 in Tone3). In other words, the V gesture is activated after the gestural midpoint between the C gesture and the T1 gesture due to the additional coupling interactions introduced by the T2 (H) gesture in Tone3. Similarly, due to the additional coupling interactions introduced by the T2 (H) gesture in Tone2, the V gesture is activated after the gestural midpoint between the C and T1 (L) gestures in Tone2.

However, the extent to which the V gesture is close to the T1 (L) gesture is different between Tone3 and Tone2.

---

4Gao (2008) proposed a similar coupling structure for Tone4, where two T gestures, namely H and L, both act as onset C gestures.
In Tone3, there is a near synchronization of the V gesture and the T1 (L) gesture, whereas in Tone2, there is still a temporal lag between the V gesture and the T1 (L) gesture. Therefore, the T2 (H) gesture in Tone2 should be regarded as an onset C gesture that does not bear as strong of a coupling with the V gesture as the T1 (L) gesture. To put it differently, in Tone2, the T2 (H) gesture is still in-phase coupled to the V gesture, but the T2 (H)-V coupling is weaker compared to the T1 (L)-V coupling or the C-V coupling, which results in the difference in the CV% between Tone2 and Tone3.

4.2 Modeling incomplete neutralization in T3S

The comparison of the CV% between Tone2 and T3H shows that, for S1, the V gesture is activated closer to the T1 (L) gesture in T3S than in Tone2, inducing incomplete neutralization between T3S and Tone2. One possible explanation for this pattern is a stronger coupling of T2 (H)-V in T3S than in Tone2. In T3S, the T2 (H)-V coupling is still weaker compared to the T1 (L)-V coupling, but the difference of coupling strength between T2 (H)-V and T1 (L)-V is smaller in T3S than in Tone2. Because of this, the CV% is larger in T3S compared to Tone2. Also note that because there is no coupling between the V gesture and the T2 (H) gesture in T3H, i.e. zero coupling strength, the V gesture of T3H is activated farther to the T1 (L) gesture in T3H than in Tone2. Hence, S1 displays a gradual decrease in the VT1 lag (thus in the CV%) across the three tones (T3S > Tone2 > T3H).

Figure 7 shows the proposed coupling relations for lexical Tone3, T3H, and T3S. In producing T3H (cf. Figure 7b), speakers decohere the coupling between the T2 (H) gesture and the V gesture. As a result, the two in-phase coupled gestures are un-coupled, which gives rise to T3H. The T2 (H) is anti-phase coupled to T1 (L) only. This causes the CV% of T3H to fluctuate around 50%. In producing T3S (cf. Figure 7c), speakers adjust the T2 (H)-V coupling in Tone3 to assimilate to that in Tone2 in order to preserve the structure of the Mandarin tone inventory, i.e. to avoid creating a new tone category. During the assimilation to Tone2, speakers are biased by the underlying coupling structure of Tone3, the underlying tone. Therefore, the coupling strength between the T2 (H) gesture and the V gesture in T3S is not adjusted to be exactly the same as that in Tone2, but instead might resemble that in Tone3 or lie between that in Tone2 and Tone3.

Independent evidence of a bias toward Tone3 in producing is found by Chen et al. (2011): T3S is stored in the mental lexicon as Tone3, despite that it is Tone2 that it is assimilated to. Under the framework of the form preparation paradigm, the sequences Tone3 + Tone3 and Tone3 + Tone2 showed 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.

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. This explains why the difference between T3S and Tone2 is significant for S1 whereas S2 displays no such difference between the two rising tones.

There are alternative parameterizations of coupling structure that may result in timing patterns similar to those observed in Tone2 and T3S (cf. Yi, 2014); future research should investigate these alternatives and develop methods to evaluate them empirically. It is also likely that speakers have learned different coupling structures for derived patterns such as T3S, so collecting data from more speakers will help shed light on the range of such patterns.
in-phase coupling
relatively weak in-phase coupling
anti-phase coupling

![Diagram](image.png)

Figure 7: Proposed derivations of T3H and T3S. The T2 (H) gesture acts as an additional onset gesture in the underlying coupling structure of lexical Tone3. The double solid line indicates a stronger coupling than the single solid line within a coordinative unit. The dashed arrow indicates potential task-specific bias.

## 5 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 can offer 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 subcategorical, gradient differences in articulatory timing. It is important for speech production models 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 phonological status of the difference between T3S and Tone2 as well as the inter-speaker variations from an articulatory perspective.

## References


