GESTURAL REPRESENTATIONS OF TONE IN MANDARIN: EVIDENCE FROM TIMING ALTERNATIONS

Muye Zhang, Christopher Geissler, Jason Shaw

Yale University
muye.zhang@yale.edu

ABSTRACT

We tested predictions of the Articulatory Phonology analysis of tone as gesture, providing the first direct comparison of C-V timing between syllables with and without tone. In an EMA study of Mandarin Chinese, syllables with a full tone were compared to two types of segmentally-matched toneless syllables: lexically-toneless and contextually-reduced. The key prediction is that presence of tone will condition a larger lag between the onset consonant and the vowel. Four native speakers showed this pattern– a larger C-V lag for syllables with tone–providing evidence that tone conditions timing lag. We also compared f0 trajectories across the two types of toneless syllable. Contrary to past reports, we found individual differences in whether contextually-reduced syllables were produced with a full tone. In contrast, all speakers produced lexically-toneless syllables with f0 trajectories resembling linear interpolation between flanking tones.

Keywords: lexical tone, Mandarin Chinese, articulatory phonology, EMA, individual variation

1. INTRODUCTION

The primary argument that lexical tones are articulatory gestures, as in the phonological primitives of Articulatory Phonology [5], has come from the way that tones interact with other gestures. Specifically, lexical tone has been argued to condition patterns of relative timing between consonants and vowels based on evidence from kinematic data. In syllables with complex onsets (and no lexical tone), the onset of movement of a vowel tends to occur during the middle of the preceding consonant cluster, the so-called "c-center" effect [3, 11, 8, 14, 7].

Similarly, in languages with lexical tone, the vowel has been observed to begin movement around the midpoint between the onset consonant and the tone, a pattern reported in Mandarin Chinese [6], Thai [10], and Lhasa Tibetan [9]. Considered relative to a toneless CV syllable baseline, for which onset consonant and vowel begin movement at roughly the same time, it is striking that adding a consonant to the syllable to yield CCV and adding a lexical tone to yield, e.g., CV̀, result in the same pattern of relative timing—it is as if the tone is functioning as an onset consonant. Since [6], this gestural approach to tone has served as a working hypothesis within Articulatory Phonology and one that presents a viable alternative to the analysis of tones as autosegments. However, this analysis makes a key prediction, which has yet to be tested: within the same language, CV syllables with and without tone are predicted to differ in timing. Specifically, the lag between consonant and vowel gestures should increase in a syllable with tone relative to a closely matched syllable without lexical tone. In this paper, we present what is, to our knowledge, the first empirical evidence for tone-conditioned timing variation of the type uniquely predicted by the Articulatory Phonology approach to tone.

2. APPROACH

In order to assess the timing alternations between syllables with and without lexical tone, we make use of two types of "toneless" syllables in Mandarin Chinese: (a) certain grammatical morphemes have no tone specified, including the sentence-final question particle, and (b) certain productive compounding paradigms remove the tone of embedded non-head members [4]; specifically, compounding disyllabic words with a noun-head result in a toneless second syllable (/bō.lí/ *glass* + /bēi/ *cup* > [bō.li.bēi]), whereas compounding two monosyllabic words with a noun-head does not result in such tonelessness (/liáng/ *measuring* + /yóu/ *oil* + /bēi/ *cup* > [liáng.yóu.bēi]). We call the lexically-unspecified toneless syllables the Absent condition and the contextually-reduced toneless syllables the Reduced condition.

For Aim I, we compare both of these types of "toneless" syllables to matched full-tone targets, addressing the AP hypothesis regarding the timing alternations of syllables with and without lexical tone. Specifically, we calculate the C-V lag (of a CV syllable) by subtracting an index of consonant onset from an index of vowel onset. The key prediction is that

Table 1: Example stimulus set

Condition	Context	Carrier	Target
Full	*zhè yī lèi tùzi xǐhuān bǎ shù pí qǐ kāi, zhǎo chóngzi chī.*	*wǒmen gěi tā qǐmíng jiào*	*qǐ **mù** tù*
	This type of rabbit likes prying off tree bark to look for bugs to eat. We call it: a bark-prying rabbit.		
Reduced	*zhè yī lèi tùzi hěn xǐhuān ānjìng de kàn qítā de tùzi.*	*wǒmen gěi tā qǐmíng jiào*	*qǐ**mu** tù*
	This type of rabbit likes quietly watching other rabbits. We call it: an admiring rabbit.		
Absent	*zài yīgè jǐubā, xīn lái de fúwùyuán hěn bèn. lǎobǎn wèn:*	*píngzi gài dōu bù huì*	*qǐ **mə** tā*
	At a bar, there's a new employee who's incompetent. The manager asks: "he can't even open a bottle?"		

C-V lag will be longer in the Full-tone condition than the Reduced-tone or Absent-tone, due to the presence of the tonal gesture.

While the literature claims these two types of syllables are equally toneless, we are also addressing this claim, in Aim II, by evaluating f0 trajectories for tone specification, applying the method of [15]. Specifically, we calculate the posterior probability that an f0 trajectory was a smooth interpolation between two flanking tonal targets–the f0 targets (peaks) of the tones in the syllables preceding and following the target syllable. The reasoning goes: if the target syllable has no lexical tone, then its f0 contour will be statistically indistinguishable from a linear interpolation between the preceding and following tonal peaks, as often assumed [12, 1]. If, on the other hand, the target syllable does have a tone, then the f0 trajectory will be clearly distinguishable from linear interpolation.

3. METHODS

3.1. Materials

The target syllable for all conditions was /mu/; this syllable was chosen because it has a bilabial closure gesture for the consonant, and a maximal tongue dorsum retraction for the vowel. In the Full- and Reduced-tone conditions, the target was the middle syllable of a three-syllable compound; in the Absent-tone condition, the target was the middle syllable of a homophonous three-syllable sequence. The target syllable was always preceded by the high front vowel /i/ and followed by the voiceless stop /t/ to provide a clear acoustic cue to the end of the syllable. The target syllables were either Tone 3 (low) or Tone 4 (high-falling), and the preceding and following tones always formed discontinuous tonal contours. That is, Tone 3 (low) was never preceded by Tone 4 (high-falling) or another Tone 3, and were only followed by Tone 1 (high), or Tone 4 (falling).

Contexts were created to facilitate the novel compounds in the Full- and Reduced-tone conditions, with corresponding contexts in the Absent-tone condition, resulting in a total of eight sets of three sentences each; one example is shown in Table (1). All the sentences were judged by three native speakers to be acceptable. We consider these stimuli to be an improvement from existing studies because they comprise syllables that are ecologically valid in terms of Mandarin phonotactics as well as morphological, syntactic, semantic, and pragmatic structure.

3.2. Participants

Four males were recruited from the Yale University community; their ages ranged from 19 to 25 (mean = 21.3) years of age. All were native speakers of Mandarin Chinese (screened before participation), and one was also a native speaker of American English. The procedures were explained and the experiment was conducted in Mandarin by the first author.

3.3. Paradigm & Procedure

For each experimental item, each speaker read the context silently, pronounced the carrier phrase with the target, and then pronounced the target alone. Each block contained all eight sets of three conditions in a unique, random order for each participant, for a total of 24 sentences, and 48 pronunciations of the target (in the carrier phrases and in isolation). Each speaker completed 12-15 blocks. All stimuli were presented in Chinese characters, and speakers took short breaks as needed between blocks.

The data were recorded with an NDI Wave EMA system at a 100 Hz sampling rate to capture articulatory movement. Five NDI Wave 5DoF sensors were attached to the sagittal midline of the tongue and lips, along with one each to the jaw (under the lower incisor), the nasion, and the left and right mastoids. The anterior tongue sensor (tongue tip; TT) was attached 1 cm from the tongue tip, the posterior sensor (tongue dorsum; TD) was attached 5 cm behind the TT sensor, and the middle sensor (tongue body) was attached at the midpoint between the TT and TD sensors. Acoustic data were recorded at 22 KHz.

3.4. Post-processing & Analysis

The bite plane was recorded for each subject with three 5DOF sensors on a rigid object held between the subject's teeth. Head movements were corrected computationally using these sensors and the three non-oral sensors as references; this correction rotated the data to align the origin of the spatial coordinates with the occlusal plane at the front teeth.

In Aim I, for the consonant gesture, a lip aperture (LA) measure was computed as the Euclidean distance between the upper and lower lip sensors; for the vowel, the position of the tongue dorsum (TD) sensor was used directly. The gestural onsets of the consonant and the vowel were determined using a 20% threshold of peak velocity in the relevant trajectories (LA for the labial consonant and TD for the vowel). Each token's C-V lag was calculated by subtracting the timestamp of the achievement of target of the consonant gesture (the start of the gestural plateau) from the achievement of target of the vowel. We measured lag based on these landmarks instead of the gesture onset landmark directly because of recent work showing that the timing of the gesture onset in Mandarin is influenced by the spatial position of the articulators and also because we observed a high degree of temporal variability in gestural onsets in these data [13]. To account for variation in speech rate, each lag measure was normalized by dividing it by the duration of the entire syllable (the time from the consonant's onset of movement to the offset of movement of the vowel). Outliers to the model fit were removed following [2] (cf. *a priori* trimming). C-V lags were z-scored per subject and compared across conditions and subjects; statistical significance is reported on the basis of nested comparison of linear mixed-effects models with random intercepts for subjects, items, and tonal contour, and a by-subject random slope for condition.

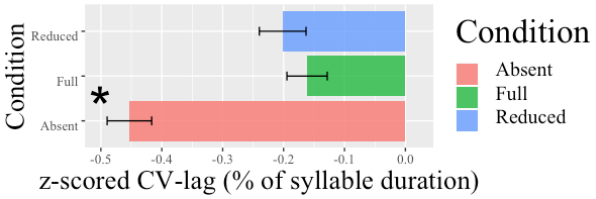
In Aim II, pitch tracking was conducted in MATLAB using the YAAPT algorithm [17]. The tonal peaks in the f0 contours of the syllables preceding and following the target were identified; these intermediate trajectories were compared to determine the presence or absence of a tonal gesture in the target. Following [15], we represented continuous f0 trajectories between the flanking tonal targets (the f0 peaks of the syllables preceding and following the target syllable) with four DCT components, which account for, on average, 96% of the variance in the data.

A Bayesian classifier was trained on the DCT representations of the Full-tone condition (non-linear trajectories due to the discontinuous tonal contour design of the stimuli) vs. a linear interpolation trajectory, following [15]; the results of this were used to reclassify tokens as Full, Reduced, or Absent tone.

4. RESULTS

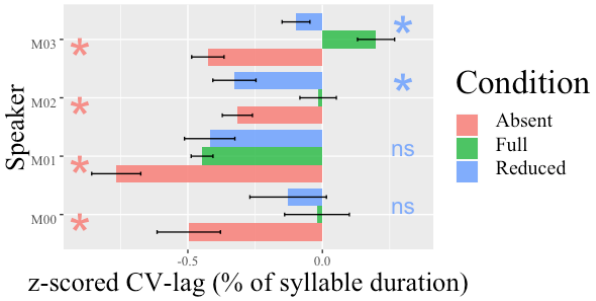
To address Aim I, normalized C-V lags were compared across speakers and are shown in Figure (1). A significant main effect of condition (p=.032) was observed; pairwise t-tests corrected with Holm's method show that the mean lag for the Full-tone condition was significantly greater than the lag in the Absent-tone condition (p<.001) but not so than the lag in the Reduced-tone condition (p=.22).

Figure 1: Aim I: C-V timing alternations. Whiskers indicate standard error of the mean.



This difference in C-V lag between the Full- and Absent-tone conditions provides evidence that the presence/absence of tone conditions intra-syllabic timing relations, consistent with the AP analysis of tone as gesture. However, notable individual differences were observed in these comparisons, so individual models were created to analyze the C-V lags within subjects, which are presented in Figure (2).

Figure 2: Aim I: C-V lags by speaker. Stars indicate significant differences vs. Full-tone tokens.

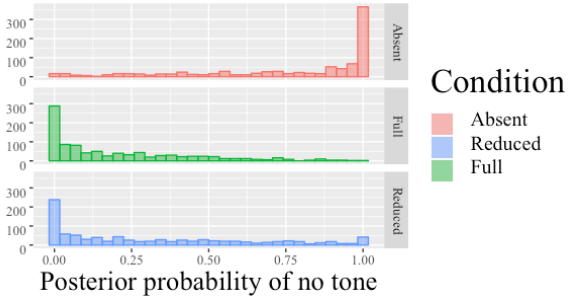


The predicted pattern–a greater C-V lag for Full-tone targets relative to Absent-tone targets–was observed for all speakers (p's<.04). However, the C-V lags for the Reduced-tone targets patterned differently across speakers: M00 and M01's Reduced-tone tokens were not significantly different than their Full-tone tokens (p's>.6), while M02 and M03's Reduced-tone were significantly shorter (p's<.04), suggesting that these two speakers did indeed show contextual reduction in tone, as indexed by C-V lags.

In Aim II, the target syllables' f0 trajectories were analyzed using a Bayesian classifier, which calculated posterior probabilities (PP) that an f0 trajectory was a linear interpolation between flanking tonal targets: specifically, PPs of 0 represented f0 trajectories like the full tone condition, whereas PPs of 1 represented f0 trajectories that interpolated between flanking tones.

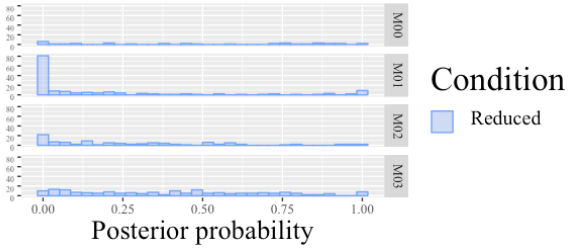
Firstly, Absent condition tokens were largely classified as having no tone, as their f0 trajectories were statistically indistinguishable from linear interpolation. Tokens in the Full condition were largely classified as having tone. Figure (3) shows the PPs for the Reduced-tone tokens for all speakers; this pattern indicates that so-called "Reduced-tone" tokens were largely pronounced with tones; that is, there are few tokens that are intermediate between full tone and linear interpolation.

Figure 3: Aim II: posterior probability of no tone.



Altogether, these results show a situation approximating complete categoricality. However, individual speakers pattern differently; PPs for the Reduced condition tokens are shown in Figure (4).

Figure 4: Aim II: posterior probability of no tone.



Again, the results show individual variation: M01 showed no reduction in C-V lags or f0, while M00 showed some reduction across tokens in f0 (but not in C-V lags) and M02/M03 show some reduction in both measures. These speakers therefore exhibit the putative tonal alternation, in which the exact same syllable /mu/ is produced both with and without tone, as facilitated by the different experimental contexts.

5. DISCUSSION

Overall, the results presented here are positive evidence for a gestural representation of tone; specifically, our findings suggest that tonal "gestures" interact with consonant and vowel gestures such that syllables with and without lexical tone have different C-V timings. This result would be unexpected in an autosegmental analysis of tone.

We also investigated whether the two types of toneless syllables in Mandarin Chinese were of one category or not. By comparing lexically-absent tone (in grammatical particles) with contextually-reduced tone (in certain compounding paradigms), we found that the lexically toneless tokens were toneless, as expected, but that for half the speaker sample, the contextually toneless tokens showed f0 trajectories indistinguishable from full-tone tokens. This finding contrasts the report in [4], suggesting variability in the degree to which this tonal-reduction paradigm is present across the broader speech community.

Taken together, the C-V lag and f0 measures suggest that some degree of tonal reduction is indeed taking place, independent of speech rate (due to normalization by syllable duration). However, the measures were not entirely consistent, as some speakers showed reduction in one measure without a corresponding reduction in the other. This could indicate that the relation between tone presence/absence and C-V lag is mediated by a third factor which affects each measure on different timescales.

One major theme emerging from this research is the importance of identifying variability in the data. While the C-V lag results show a consistent result between the Absent-tone and Full-tone conditions across speakers, the Reduced-tone tokens showed two different patterns within the sample. Furthermore, the C-V lags in the Reduced-tone condition pattern differently across speakers, despite a relatively uniform production of tone in that condition.

Future work will address possible factors influencing individual differences. A possible tool for this is the Autism Quotient questionnaire, a measure of context-sensitivity already used by researchers in phonetics [16] as well as in developmental pragmatics [18] and neurolinguistics [19]. Additional subjects tested in this paradigm would elucidate the role of such a domain-general cognitive dimension in this context-modulation task.

Overall, identifying alternations in tone presence and absence is key to evaluating a gestural account of lexical tone. We carried out this study in Mandarin building on [6], but it is important to consider more cases of true tone alternations in more languages to evaluate a gestural representation of lexical tone.

6. REFERENCES

[1] Arvaniti, A., Ladd, D. R. 2015. Underspecification in intonation revisited: a reply to Xu, Lee, Prom-on and Liu. *Phonology* 32, 537–541.

[2] Baayen, R. H., Milin, P. 2010. Analyzing Reaction Times. *International Journal of Psychological Research* 3(2), 12–28.

[3] Browman, C. P., Goldstein, L. 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45(2-4), 140–155.

[4] Chen, Y., Xu, Y. 2006. Production of weak elements in speech - Evidence from F0 patterns of neutral tone in Standard Chinese. *Phonetica* 63(1), 47–75.

[5] Gafos, A., Goldstein, L., Côté, M.-H., Turk, A. 2012. Organization of Phonological Elements: Articulatory Representation and Organization. In: Cohn, A. C., Fourgeron, C., Huffman, M. K., (eds), *The Oxford Handbook of Laboratory Phonology*. 1–23.

[6] Gao, M. 2008. *Mandarin Tones: An Articulatory Phonology Account*. Phd thesis Yale University.

[7] Hermes, A., Mücke, D., Auris, B. 2017. The variability of syllable patterns in Tashlhiyt Berber and Polish. *Journal of Phonetics* 64, 127–144.

[8] Hermes, A., Mücke, D., Grice, M. 2013. Gestural coordination of Italian word-initial clusters: The case of 'impure s'. *Phonology* 30, 1–25.

[9] Hu, F. 2016. Tones are not abstract autosegmentals. *Speech Prosody 2016* 302–306.

[10] Karlin, R., Tilsen, S. 2015. The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai. *Proceedings of Meetings on Acoustics* 22(060006), 1–9.

[11] Marin, S., Pouplier, M. 2010. Temporal Organization of Complex Onsets and Codas in American English: Testing the Predictions of a Gestural Coupling Model. *Motor Control* 14(3), 380–407.

[12] Pierrehumbert, J. B., Beckman, M. 1988. Japanese tone structure. *Linguistic Inquiry Monograph Series* Cambridge, MA. MIT Press.

[13] Shaw, J. A., Chen, W.-r. 2018. Variation in the spatial position of articulators influences the relative timing between consonants and vowels: evidence from CV timing in Mandarin Chinese. *16th Conference on Laboratory Phonology* Universidade de Lisboa, Lisbon Portugal.

[14] Shaw, J. A., Gafos, A. I. 2015. Stochastic time models of syllable structure. *PLoS ONE* 10(5), 1–36.

[15] Shaw, J. A., Kawahara, S. 2018. Assessing surface phonological specification through simulation and classification of phonetic trajectories. *Phonology* 35, 481–522.

[16] Yu, A. C. L. 2010. Perceptual Compensation Is Correlated with Individuals' "Autistic" Traits : Implications for Models of Sound Change. *PLoS ONE* 5(8).

[17] Zahorian, S. A., Hu, H. 2008. YAAPT pitch tracking MATLAB function. *The Journal of the Acoustical Society of America* 123, 4559–4571.

[18] Zhang, M., Piñango, M. M., Davidson, K. 2017. The development of metonymic processing as the growth of context-construal ability. *Poster session presented at the 42nd Annual Boston University Conference on Language Development (BUCLD42)* Boston, MA.

[19] Zhang, M., Piñango, M. M., Deo, A. 2018. Real-time roots of meaning change: Electrophysiology reveals the contextual-modulation processing basis of synchronic variation in the location-possession domain. Kalish, C., Rau, M., Zhu, J., Rogers, T. T., (eds), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* Austin, TX. Cognitive Science Society 2783–2788.