# SYNCHRONY AND STABILITY OF ARTICULATORY LANDMARKS IN ENGLISH AND MANDARIN CV SEQUENCES

Benjamin M. Kramer[1], Michael C. Stern[1], Yichen Wang[2], Yuyang Liu[1], Jason A. Shaw[1]

[1]Yale University, [2]Michigan State University
ben.kramer@yale.edu, michael.stern@yale.edu, wangy176@msu.edu, yuyang.liu.yl2472@yale.edu, jason.shaw@yale.edu

## ABSTRACT

Theories of articulatory timing differ in whether the onsets or endpoints of movements are coordinated. We assessed these alternatives in word-initial CV sequences in English, a language without lexical tone, and Mandarin, a language with lexical tone, as measured using electromagnetic articulography. We found that, on average, the vowel target and consonant offset were achieved near-synchronously in both English and Mandarin, while the onsets of the consonant and vowel movements showed a temporal lag, a result consistent with endpoint-based coordination. The timing of the near-synchronous landmarks, however, was also more variable than the timing of onsets, suggesting a need for closer examination of synchrony- vs. stability-based metrics of temporal coordination.

**Keywords**: speech timing, Articulatory Phonology, endpoint-based timing, General Tau theory

## 1. INTRODUCTION

Languages are known to differ in how articulatory gestures are coordinated (e.g., [1]). Even similar sequences of segments can have different patterns of gestural timing across languages, which can be conditioned by phonological structure, such as syllables [2]–[5]. While the question of how articulatory gestures are coordinated in time has received substantial attention in the Articulatory Phonology (AP) literature [6], competing approaches have also been proposed (e.g., [7], [8]).

A key dimension of variation across theories is whether gestural coordination is based on the *onset* or the *endpoint* of movement. In the coupled oscillator model, as implemented by Nam and Saltzman [9], only the onsets of articulatory gestures are temporally coordinated. Moreover, the onsets of consonant and vowel gestures in a CV syllable are coordinated in-phase, which, in the absence of other influences, predicts temporal synchrony. Such CV synchrony has been reported for English [10] and German [11].

Tone languages, however, tend to show a positive *CV lag*, meaning that the onset of the vowel gesture occurs after the onset of the consonantal gesture. Gao [12] proposed that positive CV lag, first identified in Mandarin, derives from competitive coupling between the consonant gesture, the vowel gesture, and a tonal gesture, which is absent in languages without lexical tone. Similar positive lags have been found in other tonal languages, such as Thai [13] and varieties of Tibetan [14], [15]. There is, however, some conflicting evidence. Liu et al. [8] found consonant and vowel movement onsets to be synchronous (i.e., no CV lag) in Mandarin using a different methodological approach. In fact, methodological variation across studies could account for the difference in whether C and V gestures are found to be synchronous or not; a study of Swedish [16], a pitch accent language, reports a CV lag commensurate with tone languages when using the methods of Mücke et al. [11] but different results when using those of Löfqvist and Gracco [10].

Turk and Shattuck-Hufnagel [7] propose an alternative to onset-based coordination. They theorize that articulatory movement endpoints are involved in temporal coordination in speech, citing greater stability of movement endpoints compared to movement onsets observed across a range of motor functions, including speech [17]. In addition, Shaw and Chen [18] find on-average synchrony of the offset of consonant release movement and the achievement of vowel target in CV syllables in Mandarin. This finding is possibly consistent with endpoint-based models of timing [7], particularly if we consider the movement away from the consonant constriction to be independently controlled [19].

To evaluate onset- and endpoint-based proposals for gestural coordination, we used electromagnetic articulography to record productions of word-initial CV sequences in both English and Mandarin and investigated the synchrony and stability of key intervals bounded by either (1) movement onsets, (2) movement endpoints, or (3) a combination. We considered three landmarks to be movement endpoints: achievement of target of the consonant (Ctar), achievement of target of the vowel (Vtar), and the offset of movement away from consonant constriction (Coff). We investigated four intervals defined by these landmarks, as well as the consonant and vowel movement onsets (Cons, Vons), listed here: (1) Cons–Vons; (2) Vons–Ctar; (3) Ctar–Vtar; and (4) Vtar–Coff. A schematic depiction of the intervals for the syllable /ba/, defined over a lip aperture (LA) trajectory for the consonant and a tongue dorsum (TD) trajectory in the vertical dimension for the vowel, is provided in Figure 1.

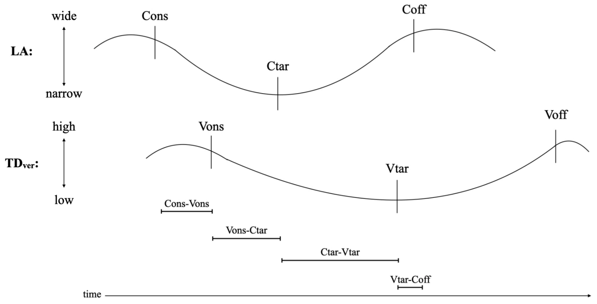

**Figure 1:** Example diagram of intervals of interest (/ba/).

## 2. METHODS

### 2.1 Participants

Acoustic and articulatory data was collected from 12 native speakers of American English (8 female, 4 male; ages 19–28, *M* = 20.75) and 12 native speakers of Mandarin Chinese (7 female, 4 male, 1 non-binary; ages 19–33, *M* = 24). No participants reported speech or hearing impairments, and none of the English speakers spoke or had studied a lexical tone language.

### 2.2 Materials

We elicited productions of eight word-initial CV sequences in each language, where the initial consonant is bilabial—either /b/ or /m/—and the vowel is either low back /ɑ/ or high front /i/. Target sequences containing the vowel /i/ were immediately preceded by the vowel /ɑ/, and sequences containing the vowel /ɑ/ were immediately preceded by the vowel /i/. This ensured that the magnitude of the tongue body movement towards the target vowel was similar across /i/ and /ɑ/ targets. All Mandarin target syllables bore a falling tone (T4) and were preceded immediately by a low tone (T3). Each target syllable was produced in two carrier sentences, occurring once in an informationally prominent position and once in a less prominent position. To encourage natural speech, each carrier sentence was preceded by a question, which served to provide context for the target sentences.

### 2.3 Data acquisition

Presentation of materials was controlled using E-Prime. On each trial, an audio recording of a question was played. The question was also displayed in text on the screen for 5000 ms. Participants were instructed to listen to the question and to read aloud the answer that followed.

In total, each participant produced 128 tokens (8 items × 2 carrier sentences × 8 repetitions) across four blocks of 32 items each. Within each repetition block, stimuli were presented in a randomized order.

The NDI Wave Speech Research System was used to record movements of nine sensors attached to the articulators and head at a sampling rate of 100 Hz. High-viscosity PeriAcryl was used to attach three sensors to the tongue: tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), placed ~1 cm, ~3 cm, and ~5 cm from the tip of the tongue. In order to track movements of the jaw, one sensor (lower incisor; LI) was attached to the hard tissue of the gum directly below the left incisor. Two sensors were attached at the vermillion border of the upper lip (UL) and lower lip (LL). Reference sensors were attached on the left and right mastoids, and on the nasion or bridge of the nose. Measurements of the occlusal plane and a midsagittal palate trace were also collected. A Sennheiser shotgun microphone collected acoustic data at a sampling rate of 22,050 Hz.

### 2.4 Data processing

Articulatory data was rotated to the occlusal plane and corrected for head movement computationally. Articulatory gestures were parsed from sensor trajectories in MVIEW [20]. Gesture *onset* and *offset* were measured as the timepoints at which an articulator's tangential velocity exceeded or sank below, respectively, a 20% threshold of a manually selected velocity peak. *Vowel target* was measured as the timepoint of minimum velocity between the onset and offset, and *consonant target* was measured as the timepoint at which tangential velocity sank below 20% of the onset velocity peak. Gestures associated with /b/ and /m/ were extracted from measurements of lip aperture (LA), calculated as the Euclidean distance between the UL and LL sensors. Gestures associated with the vowel /ɑ/ were parsed from the trajectory of the TD sensor. Gestures associated with the vowel /i/ were parsed from the trajectory of the TB sensor if this sensor was judged to be closest to the palate at the point of maximum /i/ constriction for a participant, and from the TD sensor otherwise.

Out of the 3072 tokens elicited, a total of 556 tokens (18.10%) were eliminated from analysis for the following reasons: data storage failure (18 tokens), failure of the gesture parsing tool to extract the consonant gesture, the vowel gesture, or both (378 tokens); disfluency (5 tokens); or failure of the participant to produce contrastive focus on the informationally prominent syllable, as judged by the experimenters (155 tokens).

## 3. RESULTS

Distributions of the four temporal intervals of interest are presented in Figure 2. The figures and all subsequent analysis exclude 69 tokens, for which at least one of the four intervals was greater than three standard deviations from the mean. Table 1 presents the mean by-subject duration and standard deviation of each interval.

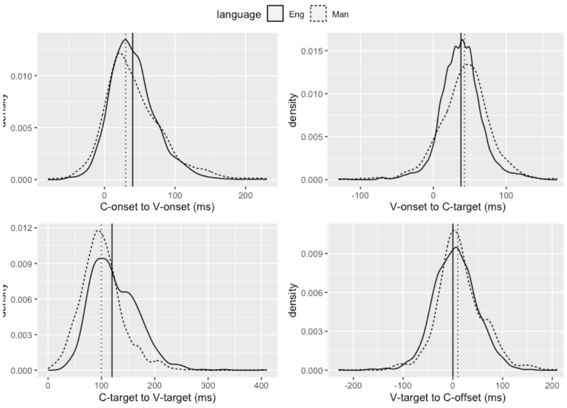

**Figure 2:** Distribution of key intervals across subjects. Vertical lines indicate the median.

| **Interval** | **Mean (ms)** | **St. Dev. (ms)** |
|---|---|---|
| **Cons-Vons** | 40.19 (Eng.)<br>41.17 (Man.) | 29.51 (Eng.)<br>39.59 (Man.) |
| **Vons-Ctar** | 37.94 (Eng.)<br>42.85 (Man.) | 25.88 (Eng.)<br>32.38 (Man.) |
| **Ctar-Vtar** | 123.81 (Eng.)<br>98.98 (Man.) | 40.80 (Eng.)<br>36.00 (Man.) |
| **Vtar-Coff** | 4.69 (Eng.)<br>15.33 (Man.) | 42.80 (Eng.)<br>44.95 (Man.) |

**Table 1:** By-subject mean and standard deviation of key intervals.

First, regarding synchrony, the mean duration of Vtar–Coff is much closer to zero than the other intervals, a pattern that also holds for each subject individually (Figure 3). To further probe this interval, we fit a linear mixed effects model with fixed effects of language, target syllable prominence, and vowel (all sum-coded) and random intercepts for subject and item. The effects of prominence and target vowel were small and not statistically significant (Table 2). The effect of language was significant, with Vtar–Coff longer in Mandarin than in English. The intercept, representing the grand mean, is 10.60 ms, indicating that the consonant offset occurs, on average, 10 ms after the vowel target. Language has a -5.74-ms effect, indicating that the English mean is ~6 ms shorter than the grand mean (i.e., the consonant offset comes just 4 ms after vowel target). Since English and Mandarin have been predicted to differ in Cons–Vons, we also fit a linear mixed effects model with the same structure as the one reported in Table 2 to this interval. Language was not a significant predictor of Cons–Vons duration, as determined by nested model comparison ($\chi^2(1) = .02$, $p = .88$).

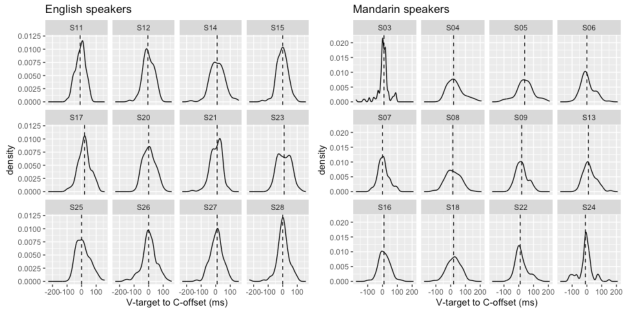

**Figure 3:** Distribution of Vtar-Coff by subject.

| | Estimate | Std. Error | *df* | *t* val. | *p* val. |
|---|---|---|---|---|---|
| (Intercept) | 10.60 | 3.54 | 10.94 | 3.00 | 0.01 |
| Language | -5.74 | 2.09 | 21.27 | -2.75 | 0.01 |
| Prominence | -0.62 | 0.88 | 2418.19 | -0.70 | 0.48 |
| Vowel | -3.67 | 2.99 | 5.95 | -1.23 | 0.27 |

**Table 2:** Results of linear mixed effects model of Vtar–Coff.

Second, regarding stability, the standard deviations in Table 1 indicate that Vons–Ctar, an interval defined by one onset-based landmark and one endpoint-based landmark, was the most stable. The Vtar–Coff interval, which was the closest to zero, was more variable than the others. Levene's tests confirm that the Vtar–Coff interval is significantly more variable than Cons–Vons ($F = 71.33$, $p < .001$) and Vons–Ctar ($F = 243.18$, $p < .001$); the difference with Ctar–Vtar was not significant ($F = 1.63$, $p = .20$). In addition, Welch's t-tests indicate that by-subject standard deviations for the Vtar–Coff interval are significantly higher than for the other three intervals (every $p < .001$).

## 4. DISCUSSION

In order to assess proposals for the coordination of consonant and vowel gestures, we evaluated the synchrony and stability of four sets of gestural landmarks, corresponding to either movement onsets or movement endpoints. The landmarks that showed greatest synchrony were Coff (endpoint of consonant release movement) and Vtar (endpoint of vowel movement to target) [mean duration = 4.69 ms (Eng.), 15.33 ms (Man.)]. The inter-landmark interval that showed the greatest temporal stability was the interval between Vons (onset of vowel movement) and Ctar (endpoint of consonant closure movement) [*SD* = 25.88 ms (Eng.), 32.38 ms (Man.)]. Notably, the landmarks that were most nearly synchronous were different than those that were most stable.

It is also noteworthy that we did not find a significant effect of language on the interval Cons–Vons, since it has been proposed that tonal gestures in languages with lexical tone bring about different patterns of temporal coordination compared to languages without lexical tone [12], [21]. We found a delay (i.e., a CV lag) between Cons and Vons not only in Mandarin (*M* = 41.70), but also in English (*M* = 40.57), a language in which onsets of C and V gestures have been theorized to be synchronized [9]. At least as indicated by the Cons–Vons interval, there was no difference in coordination between languages. Furthermore, the difference between languages was relatively small—only 10.64 ms (approximately one sample of our EMA data)—for the most synchronous interval, Vtar–Coff [mean duration = 4.69 ms (English), 15.33 ms (Mandarin)].

While consonant offset and vowel target may be near-synchronous in English and Mandarin, the question of whether these articulatory landmarks are coordinated (i.e., controlled by the speech production system such that they occur at the same time) remains open. A hypothesized coordination of consonant offset and vowel target represents a departure from the coupled oscillator model of gestural timing, in which only gesture onsets are eligible for coordination, with timing of a trajectory beyond onset simply following from the gesture's dynamic parameters [6], [7]. Gestural alignment patterns referencing targets have, however, featured in other AP research [22]. Phonological control of consonant offset and vowel target is also consistent with the proposal by Turk and Shattuck Hufnagel [7] that movement endpoints—not onsets—are controlled in speech timing, as long as it is assumed that consonant offset is the endpoint of a release gesture (cf. [19]) and vowel target is also the endpoint of a gesture. The interval we found to be most stable, Vons–Ctar, is also bounded by an endpoint (Ctar) on one end. Vons–Ctar stability is consistent with sequential timing (180-degree phasing) in AP, whereby the vowel gesture starts when the consonant gesture achieves its target; this model, however, also predicts temporal synchrony, which was not observed.

As suggested by Turk and Shattuck-Hufnagel [7], a possible theoretical implementation of endpoint-to-endpoint coordination is provided by General Tau theory [23], which allows for two movements to be *tau-coupled*, meaning that continuous, mutual sensory input guides the rate of closure (to a *goal*) such that the goals of each movement are achieved simultaneously. Temporal coordination of gesture endpoints (consonant release and vowel) may also be compatible with an Articulatory-Phonology-style framework if the stiffness parameter is allowed to be modulated (cf. [24]) continuously throughout the duration of the gesture to achieve synchronized gap closure. Synchronizing vowel target achievement with the offset of the consonant release gesture may serve the function of maximizing the acoustic salience of the vowel target by reducing acoustic interference from an ongoing consonantal constriction. This synchronized event may also correspond with the "peakRate" event in the acoustic signal, which is privileged in neural processing of speech [25].

In interpreting these results, it is important to note that we used similarly structured materials, the same method of gesture parsing, and the same methods of analysis across languages, which mitigates the issue of drawing cross-linguistic conclusions without methodological unity [16]. In undertaking a more direct comparison across languages, we found that there is, in fact, less variation in CV timing than expected based on past work. Both English and Mandarin showed a similar CV onset lag. Moreover, in both languages, Vons–Ctar was the most stable interval and Vtar–Coff was the closest to zero, indicating near synchrony of the endpoints of the consonant release movement and the vowel opening movement. Further research should aim to reconcile the metrics of synchrony and temporal stability as indicators of coordination, which paint different pictures in this data.

## 5. CONCLUSION

To assess theoretical proposals for the temporal coordination of gestures across languages with and without lexical tone, we investigated the synchrony and stability of articulatory landmarks based on both movement onsets and movement endpoints. CV timing was similar across English and Mandarin in several respects. Both languages had a positive CV lag, a pattern thought to be characteristic of languages with lexical tone but not of those without lexical tone. Additionally, the offset of controlled movement of the consonant occurred around the same time as the vowel target in both languages, in line with previous findings for Mandarin CV syllables [18]. These synchronous landmarks, however, were less stable than the others investigated, indicating that more work is required to assess the validity of synchrony and stability as indicators of coordination.

## 6. REFERENCES

[1] J. A. Shaw, "Micro-prosody," *Lang. Linguist. Compass*, vol. 16, no. 2, pp. 1–21, 2022, doi: 10.1111/lnc3.12449.

[2] J. A. Shaw and A. I. Gafos, "Stochastic time models of syllable structure," *PLoS ONE*, vol. 10, no. 5, pp. 1–36, 2015, doi: 10.1371/journal.pone.0124714.

[3] L. Goldstein, I. Chitoran, and E. Selkirk, "Syllable Structure as Coupled Oscillator Modes: Evidence from Georgian vs. Tashlhiyt Berber," *Proc. XVI Int. Congr. Phon. Sci.*, no. August, pp. 241–244, 2007.

[4] A. Hermes, D. Mücke, and B. Auris, "The variability of syllable patterns in Tashlhiyt Berber and Polish," *J. Phon.*, vol. 64, pp. 127–144, 2017, doi: 10.1016/j.wocn.2017.05.004.

[5] A. I. Gafos, J. Roeser, S. Sotiropoulou, P. Hoole, and C. Zeroual, "Structure in mind, structure in vocal tract," *Nat. Lang. Linguist. Theory*, vol. 38, no. 1, pp. 43–75, Feb. 2020, doi: 10.1007/s11049-019-09445-y.

[6] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, May 1992, doi: 10.1159/000261913.

[7] A. Turk and S. Shattuck-Hufnagel, *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford: Oxford University Press, 2020.

[8] Z. Liu, Y. Xu, and F.-F. Hsieh, "Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics," *J. Phon.*, vol. 90, 2022, doi: 10.1016/j.wocn.2021.101116.

[9] H. Nam and E. Saltzman, "A Competitive, Coupled Oscillator Model of Syllable Structure," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003, pp. 2253–2256.

[10] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1864–1876, 1999, doi: 10.1121/1.426723.

[11] D. Mücke, H. Nam, A. Hermes, and L. Goldstein, "Coupling of tone and constriction gestures in pitch accents," *Consonant Clust. Struct. Complex.*, pp. 205–230, 2012, doi: 10.1515/9781614510772.205.

[12] M. Gao, "Mandarin Tones: An Articulatory Phonology Account," Doctoral dissertation, Yale University, 2008.

[13] R. Karlin and S. Tilsen, "The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai," vol. 060006, no. 2014, p. 060006, 2015, doi: 10.1121/2.0000089.

[14] F. Hu, "Tones are not abstract autosegmentals," *Proc. Int. Conf. Speech Prosody*, vol. 2016-Janua, no. September, pp. 302–306, 2016, doi: 10.21437/speechprosody.2016-62.

[15] C. Geissler, J. A. Shaw, F. Hu, and M. Tiede, "Consistent C-V timing across speakers of diaspora Tibetan with and without lexical tone contrasts," in *Proceedings of the 12th International Seminar on Speech Production*, 2021.

[16] M. Svensson Lundmark, J. Frid, G. Ambrazaitis, and S. Schötz, "Word-initial consonant-vowel coordination in a lexical pitch-accent language," *Phonetica*, vol. 78, no. 5–6, pp. 515–569, 2021, doi: 10.1515/phon-2021-2014.

[17] J. S. Perkell and M. L. Matthies, "Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2911–2925, May 1992, doi: 10.1121/1.403778.

[18] J. A. Shaw and W. R. Chen, "Spatially Conditioned Speech Timing: Evidence and Implications," *Front. Psychol.*, vol. 10, no. December, pp. 1–17, 2019, doi: 10.3389/fpsyg.2019.02726.

[19] H. Nam, "Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structuer," *Lab. Phonol.*, vol. 9, pp. 483–506, 2007.

[20] M. Tiede, "MVIEW: Software for visualization and analysis of concurrently recorded movement data." Haskins Laboratories, New Haven, CT, 2005.

[21] M. Zhang, C. Geissler, and J. Shaw, "Gestural Representations of Tone in Mandarin: Evidence From Timing Alternations," *Int. Congr. Phon. Sci. ICPhS 2019*, no. August, pp. 1803–1807, 2019.

[22] A. I. Gafos, "A grammar of gestural coordination," *Nat. Lang. Linguist. Theory*, vol. 20, no. 2, pp. 269–337, 2002, doi: 10.1023/A:1014942312445.

[23] D. N. Lee, "Guiding Movement by Coupling Taus," *Ecol. Psychol.*, vol. 10, no. 3–4, pp. 221–250, 1998, doi: 10.1080/10407413.1998.9652683.

[24] K. D. Roon, P. Hoole, C. Zeroual, S. Du, and A. I. Gafos, "Stiffness and articulatory overlap in Moroccan Arabic consonant clusters," no. 2004, pp. 1–23, 2021.

[25] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Sci. Adv.*, vol. 5, no. 11, pp. 1–14, 2019, doi: 10.1126/sciadv.aay6279.