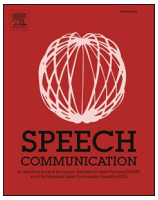
Speech Communication 163 (2024) 103121

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom



# Tone-syllable synchrony in Mandarin: New evidence and implications

Weiyi Kang *, Yi Xu

*University College London, United Kingdom*

ARTICLE INFO

*Keywords:*
Minimal contrast paradigm
Syllable boundary
Tone-syllable synchrony
GAMMs
Bayes factor
Synchronization model

ABSTRACT

Recent research has shown evidence based on a minimal contrast paradigm that consonants and vowels are articulatorily synchronized at the onset of the syllable. What remains less clear is the laryngeal dimension of the syllable, for which evidence of tone synchrony with the consonant-vowel syllable has been circumstantial. The present study assesses the precise tone-vowel alignment in Mandarin Chinese by applying the minimal contrast paradigm. The vowel onset is determined by detecting divergence points of F2 trajectories between a pair of disyllabic sequences with two contrasting vowels, and the onsets of tones are determined by detecting divergence points of $f_0$ trajectories in contrasting disyllabic tone pairs, using generalized additive mixed models (GAMMs). The alignment of the divergence-determined vowel and tone onsets is then evaluated with linear mixed effect models (LMEMs) and their synchrony is validated with Bayes factors. The results indicate that tone and vowel onsets are fully synchronized. There is therefore evidence for strict alignment of consonant, vowel and tone as hypothesized in the synchronization model of the syllable. Also, with the newly established tone onset, the previously reported 'anticipatory raising' effect of tone now appears to occur *within* rather than *before* the articulatory syllable. Implications of these findings will be discussed.

## 1. Introduction

The syllable is arguably one of the most fundamental units of speech (Coupé et al., 2019; Sun and Poeppel, 2023), but its nature has long been unsettled (Ladefoged, 1982). Recent research (Liu et al., 2022; Liu and Xu, 2021, 2023; Xu and Gao, 2018) has produced fresh evidence for a particular view that conceives the syllable as a mechanism of synchronizing consonant and vowel (CV) articulation at their onsets (Goldstein et al., 2006; Kozhevnikov and Chistovich, 1965; Nam et al., 2009). Along the same line of thinking, a further hypothesis is that laryngeal gestures for tone and phonation register are also synchronized with the vowel and consonants (Ohala and Kawasaki, 1984; Xu, 2020; Xu and Liu, 2006). This hypothesis, however, has not been examined with the same vigor as done on CV synchrony (Liu et al., 2022; Liu and Xu, 2021, 2023; Xu and Gao, 2018). The present study is conducted to fill this gap.

### 1.1. In search of the nature of the syllable

The nature of the syllable has long been a mystery. As remarked by Ladefoged (1982: 220), "[a]lthough nearly everyone can identify syllables, almost nobody can define them". An early theory was proposed by Stetson (1951), which asserts that a syllable is related to a sequence of muscular activities of the chest that control expiration. This chest pulse theory was rejected by Ladefoged (1967), however, who showed that thoracic muscle activities during speech correspond to slow inhalation and exhalation rather than to fast occurring syllables. Some theories have taken syllable as the basic unit of speech, e.g., Stetson's motor phonetics (Stetson, 1951) and Fujimura's (1994) C/D model. In MacNeilage's (1998) frame/content theory, the syllable is suggested to have evolved from the ingestion-related cyclicities of mandibular oscillation in chewing, sucking and licking. It has also been proposed that the syllable is a unit of stored motor programs (Dell, 1988; Levelt et al., 1999). On the other hand, the widely used theories of the syllable are mostly descriptive, concerned mainly with how consonants and vowels are arranged according to their sonority levels: high sonority ones in the syllable center while low sonority ones at the edges (Sievers,1893; Jespersen,1899; Whitney, 1865), or language specific phonotactics (Selkirk, 1982). There are no clear accounts of why sonority hierarchies are needed, or basic principles behind the phonotactic constraints. In particular, none of these theories has offered a detailed account of the articulatory dynamics of the syllable, especially in terms of the timing of the articulation of consonants, vowels and tones.

A very different account of the syllable comes from a somewhat forgotten line of work. Menzerath and de Lacerda (1933) observed in

* Corresponding author.
*E-mail address:* weiyi.kang.22@ucl.ac.uk (W. Kang).

https://doi.org/10.1016/j.specom.2024.103121
Received 16 January 2024; Received in revised form 5 April 2024; Accepted 29 July 2024
Available online 31 July 2024

German, based on the means of articulatory observation at the time, that the lip movement for the vowel /u/ in /pu/ starts at about the same time as the articulation of the initial consonant. They proposed that the phenomenon was due to a general organization principle of "koartikulation" for articulatory control, a term that later became popularized as the English word "coarticulation" (Kühnert and Nolan, 1999). A phenomenon of syllable-based coarticulation similar to the original one was again observed in Russian by Kozhevnikov and Chistovich (1965). They reported that in a syllable consisting of /u/ and an initial consonant cluster, the lip protrusion of the vowel begins at the same time as the articulatory activation of the first consonant, regardless of the number of consonants in the cluster. Based on this, they proposed the concept of articulatory syllable: the temporal domain where the initial consonant (s) and the vowel in a CnV structure are articulated.

The articulatory syllable hypothesis was soon challenged by the finding of cross-segment anticipatory coarticulation that goes across the syllable boundary marked by the acoustic consonant closure (Carney and Moll, 1971; Fowler, 1981; Mok, 2012; Öhman, 1966). Those findings were interpreted as suggesting that the articulatory domain of vowel is larger than, hence cannot be equivalent to, the acoustic syllable. Note that the reason behind the challenge is a lack of proposal by the articulatory syllable hypothesis as to what the *acoustic correlate* of the common onset of vowel and consonant should be at the start of the syllable. As a result, the formant movements in the direction of the second vowel in a VCV sequence found in Öhman (1966) are widely assumed as anticipatory coarticulation. And this is despite Ohman's (1966:165) remarkable proclamation in his paper that "a motion toward the final vowel starts not much later than, or perhaps even simultaneously with, the onset of the stop-consonant gesture." It was only until Xu and Liu (2006) that a suggestion was made that the acoustic onset of the initial consonant, hence that of the syllable, should be at the start of formant movements toward the initial consonant rather than at the start of the consonant closure. Xu and Liu (2006) further proposed that not only consonant and vowel are actually fully synchronized with each other at syllable onset, thus in agreement with the articulatory syllable hypothesis, but also lexical tone is synchronized with both consonant and vowel. Furthermore, the time-locking nature of anticipatory coarticulation observed by Bell-Berti and Harris (1979, 1981) suggests that the start of vowel articulation in a CV syllable is not very far ahead of the consonant closure, and the temporal scope coarticulation may not be as variable as popularly believed. Up till the proposal of the time-lock model, the influence of the conventional syllable remains notable. The articulatory gestures initiated before the onset of conventional syllables are consistently interpreted as anticipatory, which is now challenged by studies on the synchronization model (e.g. Liu et al., 2022). Goldstein et al. (2006) went a step further by proposing that vowel and consonant gestures are aligned 'in-phase' at syllable onset, while coda consonants are aligned 'anti-phase' to the nuclear vowel. In other words, the articulatory gestures are synchronized with syllable boundaries. However, Goldstein et al. (2006) approached the alignment problem from a phonological perspective, with limited emphasis on the details of strict temporal alignment. It therefore remained unclear whether articulatory gestures were synchronized with the conventional syllable boundaries or with a timepoint earlier, i.e., the boundaries of articulatory syllable. A similar idea was suggested in the time structure model of the syllable by Xu and Liu (2006), which proposes that not only consonant and vowel, but also tone, are all synchronized at the onset of the syllable.

Synchronization is a strong claim, however, for which solid evidence is needed to show that a clearly identifiable onset of one phonetic unit is fully aligned with the clearly identifiable onset of another phonetic unit. Evidence of synchronization has been difficult to obtain. Goldstein et al. (2006) have proposed, under the articulatory phonology framework, that vowel and consonant gestures are aligned 'in-phase' at syllable onset, while coda consonants are aligned 'anti-phase' to the nuclear vowel. But no clear evidence has been shown for the in-phase alignment. One of the reasons is that in studies under the framework of articulatory phonology, the onset of a gesture is determined as the point at which velocity has reached a threshold, typically 10 % or 20 % of the peak velocity of the gesture. With this method, consonant gestures have always been found to start earlier than vowel gestures (Gao, 2009; Shaw and Chen, 2019; Yi and Tilsen, 2016). This has led to the suggestion that synchronization is no longer assumed in articulatory phonology (Tilsen, 2020). Xu (2007) proposed a minimal pair method with four trisyllabic phrases: [ni ji wei], [lou ji wei], [mao ji wei], and [ma ji wei], to separately determine the onsets of consonants and vowels in Mandarin. The paradigm was an extension of the minimal contrast method used in Bell-Berti and Harris (1979, 1981) and Gelfer, Bell-Berti and Harris (1989), with the added capability to directly compare the timing of consonant and vowel in each syllable. The method was applied in Xu and Gao (2018) and preliminary evidence of CV synchrony was shown. Applying the same method, Liu et al. (2022) compared the timing of consonants and vowels in both articulatory (Electromagnetic Articulography or EMMA) and formant data, and found clear evidence of CV synchrony in Mandarin based on generalized additive mixed models (GAMMs) and Bayesian statistics. More evidence of CV synchrony was further found in English syllables with initial consonant clusters (Liu and Xu, 2021) and cases of resyllabification in English (Liu and Xu, 2023).

The synchronised onsets of C and V determined in these recent studies are temporally well ahead of the conventional syllable onset defined as the onset of initial consonant closure (Lehiste and Peterson, 1961; Turk et al., 2006). For the vowels, the newly determined onset would correspond to what is reported by Öhman (1966) as the start of the cross-syllable anticipatory vowel-to-vowel coarticulation. Fig. 1 shows an example of formant trajectories of a VCV sequence from Öhman (1966), in which F2 in [aby] begins to anticipate [a] well before the stop closure. This "anticipation" was reinterpreted by the synchronization model as the onset of [y] based on the target approximation mechanism of all articulatory gestures (Xu and Liu, 2006; Xu and Wang, 2001). Given also that similar "anticipation" of the [b] closure happens at roughly the same time, as indicated by F1 in Fig. 1, Öhman's (1966) finding could have been viewed as evidence of C-V co-onset rather than long-distance anticipatory coarticulation, and the point of CV co-onset before the conventional syllable boundary would be redefined as the onset of the syllable. As a result, the articulation of both C and V would happen within rather than before the domain of the syllable, which is in line with the hypothesis of Kozhevnikov and Chistovich (1965) that the domain of a syllable is where the consonant cluster(s) and the vowel are articulated.

The syllable is made up of not only consonants and vowels, but also laryngeal activities that generate voicing with fundamental frequency ($f_0$) patterns. So, a complete model of the syllable needs to also have an account of the laryngeal activities during its articulation. Ohala and Kawaski (1984:123) hypothesized that "the division of sound sequences into syllables" is "for the sake of synchronizing the segmental and suprasegmental articulations". Ohala (1992:335) further speculates that "syllable chunking may be done for the sake of synchronizing suprasegmental and segmental events or to accommodate neuromotor

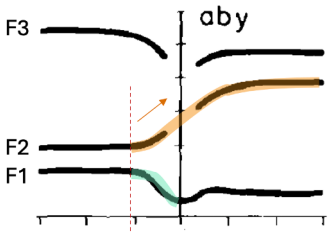



**Fig. 1.** Formant trajectories of [aby] from Öhman (1966). The start of the middle gap (where F2 and F3 end before the vertical axis) is conventionally considered as boundary between the first and second syllables. The dotted vertical line on the left represents the beginning of anticipatory coarticulation (Öhman, 1966).

constraints." The synchronization model of the syllable (Xu, 2020; Xu and Liu, 2006) hypothesizes that suprasegmental elements like tone and phonation register also start from syllable onset, just like consonant and vowel. Xu (2020) proposes further that the synchronization of consonant, vowel and laryngeal element is for the sake of eliminating most of the temporal degrees of freedom to make it possible for the central neural system to coordinate multiple articulators during speech.

There has been some preliminary evidence for the hypothesized synchrony of tone with the syllable (Ohala & Kawaski, 1984; Ohala, 1992; Xu, 2020; Xu and Liu, 2006). First, $f_0$ movements toward the underlying tonal targets appear to always occur within the temporal scope of the (conventional) syllable (Xu, 2005). Second, the timing of the tonal target approximation movement is not affected by voiceless consonants in Mandarin (Xu and Xu, 2003) or English (Xu and Xu, 2021). Finally, the timing of the tonal target approximation is not affected by the presence of nasal coda in Mandarin (Xu, 1998). However, the observed tone-syllable synchrony in these studies is with the conventional syllable, namely, temporal intervals demarcated by consonant closures (to be discussed in detail in 1.2). The newly established CV synchrony syllable raises the question as to whether tone is actually synchronized with the newly established articulatory syllable rather than the conventional acoustic syllable.

### 1.2. Tone and tonal alignment

Tones are fundamental frequency patterns used to differentiate words or morphemes that are otherwise identical in segmental composition in many of the world languages (Yip, 2002; Hyman, 2011; Xu, 2019). Among the many tonal issues that have been researched is the alignment of tone with segmental events. Prompted by a rich variety of contextual variations in many African tone languages, Goldsmith (1976) developed autosegmental phonology, which represents tones in an autonomous tier separate from the segmental tier. Segments and tones are then linked to each other based on language-specific association rules. The free and flexible alignment of tone and segment is considered as an essential and advantageous feature of this framework, as the independence of tone would allow freedom of association of tonal and segmental features for explaining various tonal variations (Goldsmith, 1976). There have also been experimental studies that look for evidence of contrastive $f_0$ peak alignment within the syllable in some tone languages (DiCanio et al., 2014; Remijsen and Ayoker, 2014; Zsiga and Nitisaroj, 2007). Xu (2005) argues, however, that surface $f_0$ alignment cannot be directly specified by phonology, because underlying tonal targets have to be articulatorily realized through target approximation (TA) (Xu and Wang, 2001), as schematized in Fig. 2. TA is similar to the task dynamic (TD) model for segmental gestures (Saltzman and Munhall, 1989) in that both assume that the basic mechanism of speech production is to approach linguistically specified targets (gesture scores or pitch targets). TA differs from TD, however, in assuming that targets themselves can be dynamic, as illustrated by the first target in Fig. 2. The rising or falling tones are commonly understood as having two tonal targets (rising: low-high; falling: high-low) in other frameworks, including TD. Also, as a model of tone, TA explicitly assumes that the temporal scope of target approximation is the syllable.

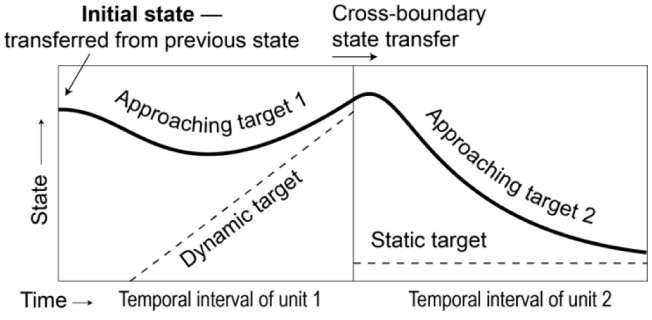



**Fig. 2.** A generalized Target Approximation (TA) model for tonal as well as other phonetic units. The dashed lines are underlying targets that are either static or dynamic, and the solid curves are the results of asymptotic approximation of the underlying targets. The final state of the first TA movement is transferred to the second movement to become its initial state, and the delayed peak into the second syllable is the result of inertia due to the state transfer at the syllable boundary.

Like TD, TA currently encompasses only a single articulatory mechanism, namely, to approach the target by overcoming inertia. But there are other known articulatory effects on the realization of tone as well, most notably, an effect known as anticipatory raising, pre-low raising, anticipatory dissimilation or H-raising. This effect was reported almost simultaneously for Thai (Gandour et al., 1992; Gandour et al., 1994), Yoruba (Laniran, 1992; Laniran and Clements, 2003), Igbo (Liberman et al., 1993) and Mandarin (Xu, 1993; 1997). The core phenomenon is that the high $f_0$ point of a tone becomes higher before a low tone than before a high or mid tone. For Mandarin, however, the raising effect is exerted also by a rising tone (Tone 2) and possibly by a falling tone as well (Xu, 1997). The articulatory mechanism of pre-low raising is still unclear, but it is likely related to the large downward movement of the larynx needed for pitch lowering, which involves the contraction of the external laryngeal muscles, particularly the sternohyoid (Atkinson, 1978; Erickson et al., 1995; Honda et al., 1999), as discussed in greater detail in Xu (2019).

Previous reports of pre-low raising have not attempted to identify the precise starting point of the effect, however. The substantial leftward shift of the boundary in the newly established articulatory syllable (Xu and Gao, 2018; Liu et al., 2022) raises the possibility that the raising effect happens within rather than before the articulatory syllable. If confirmed, this would remove another effect that has been considered anticipatory, just like most of the V-to-V anticipatory assimilation as discussed by Öhman (1966). But it would also introduce a significant complication, that is, the monotonous articulatory movements toward the target, labelled as gestures in TD or target approximation in TA, would include an optional initial movement in the opposite direction of the target if the syllable is assigned a low target, which would conflict with the unidirectional target approximation principle specified by the present version of TA and thus call for further update of relevant models including TA. But this potential complication is actually another reason why the determination of the precise tone-syllable alignment is of great importance, as it may force an update of TA or similar models to address a scenario that would be unimaginable based on only a vague notion of the temporal interval of tone.

The present study is therefore aimed at determining the precise onset of tone articulation, with two research questions: (1) Is tone articulation synchronized with the CV syllable? (2) Does 'anticipatory raising' occur before or within the articulatory syllable? Note, however, that the second question can be answered only if a clear answer to the first question is obtained.

## 2. Methodology

The basic method is to adopt the minimal contrast paradigm developed in Liu et al. (2022), i.e., to first determine tone and vowel onsets, respectively, by identifying trajectory divergence points in minimal contrast tone pairs and vowel pairs, and then compare their relative onset times. Here vowel onset is treated as equivalent to consonant onset as well as the onset of the articulatory syllable, based on the vowel-consonant synchrony in Mandarin found by Liu et al. (2022). The divergence point analysis is conducted with GAMMs; the timing difference of vowel and tone divergence is determined with LMEMs; and vowel-tone synchronization is verified by Bayes factors.

**Table 1**
Stimulus list.

| | Vowel group 1 | | | | Vowel group 2 | | |
|---|---|---|---|---|---|---|---|
| Tone group | IPA | Tone | Chinese | | IPA | Tone | Chinese |
| 1 | /ma.lu/ | 11 | 妈噜 | 17 | /ma.li/ | 11 | 妈哩 |
| 2 | /ma.lu/ | 21 | 麻噜 | 18 | /ma.li/ | 21 | 麻哩 |
| 3 | /ma.lu/ | 31 | 马噜 | 19 | /ma.li/ | 31 | 马哩 |
| 4 | /ma.lu/ | 41 | 骂噜 | 20 | /ma.li/ | 41 | 骂哩 |
| 5 | /ma.lu/ | 12 | 妈卢 | 21 | /ma.li/ | 12 | 妈梨 |
| 6 | /ma.lu/ | 22 | 麻卢 | 22 | /ma.li/ | 22 | 麻梨 |
| 7 | /ma.lu/ | 32 | 马卢 | 23 | /ma.li/ | 32 | 马梨 |
| 8 | /ma.lu/ | 42 | 骂卢 | 24 | /ma.li/ | 42 | 骂梨 |
| 9 | /ma.lu/ | 13 | 妈鲁 | 25 | /ma.li/ | 13 | 妈里 |
| 10 | /ma.lu/ | 23 | 麻鲁 | 26 | /ma.li/ | 23 | 麻里 |
| 11 | /ma.lu/ | 33 | 马鲁 | 27 | /ma.li/ | 33 | 马里 |
| 12 | /ma.lu/ | 43 | 骂鲁 | 28 | /ma.li/ | 43 | 骂里 |
| 13 | /ma.lu/ | 14 | 妈露 | 29 | /ma.li/ | 14 | 妈丽 |
| 14 | /ma.lu/ | 24 | 麻露 | 30 | /ma.li/ | 24 | 麻丽 |
| 15 | /ma.lu/ | 34 | 马露 | 31 | /ma.li/ | 34 | 马丽 |
| 16 | /ma.lu/ | 44 | 骂露 | 32 | /ma.li/ | 44 | 骂丽 |

These target words were to be embedded in a carrier phrase 'bǐ ___ kāng kǎi' (IPA: [bi _ kaŋ kaɪ]), meaning 'more generous than __'. With this carrier, the embedded pseudo words are understood as personal names.

### 2.1. Stimuli

The stimuli consist of 32 disyllabic pseudo words, as shown in Table 1, each in the form of $C_1V_1\#C_2V_2$, where # indicates syllable boundary. The words are divided evenly into two groups, which differ from each other only in the vowel of the second syllable: /ma.li/ vs. /ma.lu/. Within each group, a full set of bitonal combinations of the four Mandarin tones[1] are assigned. The use of pseudo words was to both minimize the word frequency effect (Wright, 1979) and solve the problem that many of the tonal combinations do not have real words.

### 2.2. Participants

5 females and 3 males were recruited for the experiment. All participants speak Mandarin Chinese as their native language and are from the northern part of China (3 from Beijing, 2 from Shandong, 2 from Shanxi and 1 from Henan). Data of 1 speaker was later removed due to constant oversmoothing problems detected in GAMMs. Although the participants also speak different dialects aside from Mandarin Chinese, they have been exposed to and frequently speaking Mandarin since early childhood. Also for each participant, the experimenter (the first author) made sure that their Mandarin had no detectable accent before being included.

### 2.3. Recording

Sound recording was done in a sound-proof recording booth in the Research Laboratory in Chandler House, University College London. All utterances were recorded with Praat (Boersma, 2001) at a sampling rate of 44.1 kHz. Before the recording, each participant was given some time to familiarize themselves with the stimuli. During the recording, one sentence was shown on a computer screen at a time. The participant read aloud the given sentence one at a time. They were instructed to speak at a normal speech rate. Each sentence was recorded 10 times in a computer-generated randomized order that was blocked by repetition. A total of 2560 utterances (32 sentences *10 repetitions * 8 speakers = 2560) were recorded.

The recordings were annotated manually with Praat as illustrated in Fig. 3. Intervals were marked at the conventional syllable boundaries (i.e., the landmark of abrupt spectral shift). The first boundary was placed at the onset of nasal murmur in /ma/, the second boundary at the onset of the /l/ closure, and the final boundary at the offset of voicing in /i/ or /u/. The $f_0$ and formant trajectories were extracted by a customized version of ProsodyPro (Xu, 2013) and FormantPro (Xu and Gao, 2018). Most of the parameters for ProsodyPro were set at the default values ($f_0$ range = 75Hz-600 Hz, Sampling rate = 200 Hz). For the $f_0$ trajectory extraction, manual corrections of the vocal pulse markings were made in cases where the automatic pulse marking by Praat was problematic.

Most of the parameters for FormantPro were also kept at the default setting (Max number of formants: 5; Window length: 0.025 s; Time step: 0.005 s; Smoothing window width: 0.10 s), except for maximum formant. To ensure the capturing of continuous F2, it was set at 6000 Hz for female speakers for /i/, where F2 could be almost as high as F3. For male speakers, it was set at 5500 Hz for /ma.li/ and 4500 Hz for /ma.lu/. The reason was that for /ma.lu/ produced by males F2 could be so low that it approaches F1, triggering Praat to mistake F3 for F2. The lowered maximum formant reduced such failures. A similar adjustment was not applied to /ma.lu/ produced by females because the improvement was negligible.

To ensure sufficient time resolution for testing the synchronization hypothesis, the time-normalized sampling rates of both $f_0$ and formant are set at 25 points per interval (syllable), which is much higher than the default ProsodyPro (10) and FormantPro (20) values.

### 2.4. Analysis

#### 2.4.1. Locating divergence points

The basic design of the stimuli was to enable the onset of the vowel and tone to be indicated by the point of departure between the trajectories of the contrasting vowel or tone in the second syllable, e.g., /maliXX/ and /maluXX/ for vowel (where X stands for any tone), and /malY11/ and /malY12/ for tone (where Y stands for either /i/ or /u/). More specifically, to locate the F2 divergence point for vowel, the tokens were organized into two groups: /mali/ and /malu/, regardless of the tones in each word. To locate the $f_0$ divergence point for tone, the tokens were grouped according to their tonal combinations, with a division of four minimal quadruplets, within each the second tones were in contrast, as shown in Table 2. For example, the first quadruplet (column 1 in Table 2) consists of Tone11/Tone12/Tone13/Tone14. In other words, within each quadruplet, combinations ending with Tones 2, 3, and 4 were compared with the ones ending with Tone 1 respectively. Also, because Tone33 undergoes tone sandhi, which renders it similar to Tone23 (Chao, 1968), this combination was compared with Tone21 rather than Tone31 (Columns 2 and 3 in Table 2).

To detect the divergence points, there is a likely confounding factor that needs to be taken into consideration. In a contrasting bi-tonal pair, the $f_0$ trajectories under comparison can be moving either in similar directions, as in T11 (/high # high/) vs. T14 (/high # fall/), or in opposite directions T11 (/high # high/) vs. T12 (/high # rise/). It is conceivable that the divergence point is detected later between two similar-direction trajectories than between two opposite-direction trajectories. To test whether this is the case, an additional grouping strategy, shown in Table 3, was applied for some of the comparisons during data analysis.

Both F2 and $f_0$ data underwent z-score transformation (Kingston, 2007; Lobanov, 1971) before further analysis, to make sure that data collected from different speakers are comparable with each other. Following Liu et al. (2022), the z-score transformed trajectories were then fitted into GAMMs, which model non-linear trajectories with random effects. For F2, one model was fitted for /mali/ versus /malu/. For $f_0$, one model was fitted for all tonal combinations, with each tonal combination as a separate level. The models were constructed for each speaker so that the variance between trials can be considered, yielding more accurate fits. Per requirement of GAMMs, 25 data points were extracted from each syllable, rendering 50 data points from the two

[1] The four Mandarin tones are Tone 1 (high level), Tone 2 (rising), Tone 3 (low), and Tone 4 (falling).

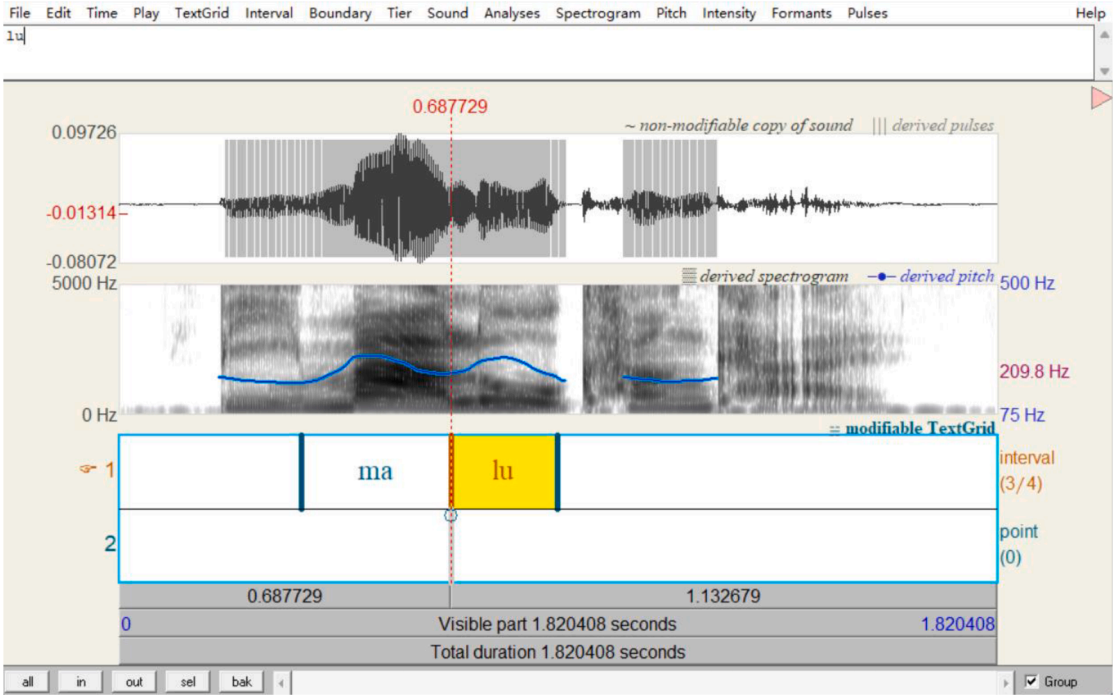



**Fig. 3.** An example of manual annotation. Interval boundaries are set at the onsets of the nasal and lateral closures as the conventional syllable boundaries. The $f_0$ track is displayed by the Praat setting, but was not used in the analysis. The $f_0$ contours used in the analysis were generated by ProsodyPro as described in the text.

**Table 2**
Contrastive quadruplet bi-tonal combinations for detecting $f_0$ divergence points. Each column shows a quadruplet (although Quadruplets 2 and 3 consist of 5 and 3 tone combinations, respectively, due to Tone 3 sandhi). The numbers refer to Mandarin tone names, and the lines depict the schematic underlying pitch shapes. For the real $f_0$ trajectories, see Fig. 7 in the Result section.

| Initial tone | Quadruplet 1 | Quadruplet 2 | Quadruplet 3 | Quadruplet 4 |
|---|---|---|---|---|
| 1 | 11 ¯ ¯ | 21 / ¯ | 31 _ ¯ | 41 \ ¯ |
| 2 | 12 ¯ / | 22 // | 32 _ / | 42 \ / |
| 3 | 13 ¯ _ | 23 / _, 33 _ _ | | 43 \ _ |
| 4 | 14 ¯ \ | 24 / \ | 34 _ \ | 44 \ \ |

target syllables for both F2 and $f_0$ trajectories.[2] To make the normalized time scale largely comparable to real time to make interpretation easy, the total normalized time was set to 0.5, so that the first conventional syllable spans from 0 to 0.25 and the second syllable from 0.25 to 0.5.

The analysis was done with the *mgcv* (Wood, 2017) and *itsadug* (van Rij et al., 2022) packages in R. GAMMs were fitted with the *bam*() function. The dependent variable was $f_0$ and F2, with vowel and tone as the main effects, respectively. The models included a by-vowel or a by-tone smooth function through time to track F2 or $f_0$ changes over time. To prevent inaccuracies caused by autocorrelation, a rho value extracted from *acf_resid*() was added to the model.[3] Based on the diagnosis report produced by *gam.check*(), the value of *k*, a parameter that controls the flexibility of each smooth, was set to 50 for formant analysis and 15 for $f_0$ analysis. Speaker 8 was excluded from the GAMM analysis because the models for his F2 were always reported as over-smoothed, even though *k* has been set to maximum.

The divergence point was identified with *plot_difference*(), which produces a difference plot of the minimal pairs and lists all time points where the difference reaches statistical significance. Fig. 4 is an example of the F2 trajectories modelled with GAMMs and its difference plot. Region(s) of significant difference was marked by red dotted vertical lines. The beginning of the significance period that lasted for more than 0.05 norm-time is recorded as the divergence point, which is roughly equivalent to the 40 ms threshold for avoiding Type I error in Liu et al. (2022).

Determining divergence points was more complex for $f_0$ models. As $f_0$ contours of tone minimal pairs often cross each other (Xu, 1999), it is natural for the significant difference to be interrupted by intersections with the x-axis. Fig. 5 shows an example where the $f_0$ contours of a minimal pair crisscross each other several times. Since the first period of

[2] It was not possible to use real-time trajectories as in Liu et al. (2022) because the fastest speakers were almost twice as fast as the slowest speakers, making the GAMMs analysis impossible.

[3] R syntax for the GAMM: model <- bam(F0 ~ Tone + s(Time, by=Tone, k = 15) + s(Time, Trial, by=Tone, bs='fs', m = 1), rho=f0acf_n[2], AR.start = data $start.event, nthreads=3)

**Table 3**
Bi-tonal minimal pairs for detecting $f_0$ divergent points. The top group consists of pairs where $f_0$ trajectories are expected to move in similar directions at syllable boundary. The bottom group consists of pairs where $f_0$ trajectories are expected to move in opposite directions at syllable boundary. The numbers refer to Mandarin tone names, and the lines schematically depict the underlying tone shapes. For the real $f_0$ trajectories see Fig. 7.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group 1 | 11 ¯ ¯ | | 21 / ¯ | | 31 _ ¯ | | 41 \ ¯ | |
| Similar direction | 14 ¯ \ | | 24 / \ | | 34 _ \ | | 41 \ \ | |
| Group 2 | 11 ¯ ¯ | 11 ¯ ¯ | 21 / ¯ | 21 / ¯ | 31 _ ¯ | 21 / ¯ | 41 \ ¯ | 41 \ ¯ |
| Different directions | 12 ¯ / | 13 ¯ _ | 22 / / | 23 / _ | 32 _ / | 33 _ _ | 42 \ / | 43 \ _ |

significant difference already exceeded the 0.05 norm-time threshold, the beginning point of this period was recorded as the diverging point, irrespective of later returns below the significance level. The difference analysis produced 91 onset times (1 F2 onset + 12 $f_0$ onset) * 7 speakers) in total.

#### 2.4.2. Comparing divergence timing in $f_0$ and F2 trajectories

To compare the timing of vowel and tone, the *lme4* R package (Bates et al., 2015) was used to produce linear mixed effect models (LMEMs). The type contrast between $f_0$ and F2 was modelled as fixed effect, and speaker and minimal pair were considered as random intercepts.[4] Random slopes were excluded to prevent singular fits, following Liu et al. (2022). A nested model that did not include type contrast was also fitted for ANOVA comparison, with which the significance of the type contrast can be quantified. A significant $p$ value would suggest that the onset times of $f_0$ and F2 were significantly different.

#### 2.4.3. Establishing divergence synchrony with Bayes factor

With inferential statistics produced with LMEMs, a nonsignificant comparison can indicate either the difference between two models is negligible or the data is insufficient to reach a firm conclusion. Non-significance alone therefore cannot provide direct evidence for the synchronized onset of F2 and $f_0$ in the articulatory syllable. Following Liu et al. (2022), to ascertain tone-vowel synchrony, Bayes factor was calculated to evaluate whether the data support the null model (i.e., onset synchrony) or the alternative model (i.e., onset asynchrony) (Stone, 2013).

Bayes factor is calculated by dividing posterior odds by prior odds, as shown in formula (1). As F2 and $f_0$ onset synchrony has not been fully established, the prior odds were set to 1 (Dienes, 2016; Liu et al., 2022), indicating that synchrony and asynchrony are equally likely in prior belief. The posterior odds were calculated by dividing the posterior distribution of the null model with that of the alternative model. As previous studies suggest, replication of studies is prone to fail when Bayes factor is about 1, but a Bayes factor higher than 3 can be considered as valid support for the null model (Dienes, 2016; Lakens et al., 2020). Therefore, this study adopted a threshold value of 3, that is, synchrony is validated only if Bayes factor exceeds 3.

$$BF_0 = \frac{Posterior\ odds}{Prior\ odds} \tag{1}$$

The calculation was conducted with the *brms* package in R (Bürkner, 2021). A non-linear model was fitted with the *brm*() function with a syntax similar to LMEMs.[5] The fixed effect and random intercept settings were identical to the LMEM. A dataset following normal distribution with a mean of zero was used as the model priors in model training (Nalborczyk et al., 2019). Each model undergoes 2000 iterations for warmup and 5000 iterations for training. A nested model that does not consider the fixed effect (i.e., type contrast) was also fitted. As iterations may be corrupted and become divergent transitions due to sampling errors, model diagnosis was conducted with *nuts_params*() to highlight the divergent transitions in the parallel coordinate plots. The models are considered acceptable as long as the divergent transitions do not show a particular pattern (Gabry et al., 2019). Bayes factor was then calculated using the *bayes_factor*() function to compare the full model against the nested model.

## 3. Results

Figs. 6 and 7 show mean F2 trajectories of /mali/ and /malu/ and mean $f_0$ contours of all the minimal bi-tonal pairs, respectively, averaged across 7 speakers. The center of the trajectories shows the means, and the shaded ribbons represent the 95 % confidence interval. The x-axis is normalized time with 0.5 representing the full length of both syllables, and 0.25 at the conventional boundary of the two syllables. The scale of the y-axis is z-score normalized F2 and $f_0$. The vertical dash-dot line in each plot represents the divergence point determined by GAMMs analysis to be discussed later.

### 3.1. F2 and $f_0$ divergence points

The fitted GAMMs provided informative descriptions of F2 and $f_0$ changes over the two syllables. As reported by the model summaries, approximately 84.41 % of F2 deviance and 88.57 % of $f_0$ deviance were explained by the models. The percentage of explained deviance of each speaker is displayed in Fig. 8. It is true that the percentage was exceptionally low for Speakers 5 and 6, but the $k$-indexes were constantly higher than 1 and the $p$ values were nonsignificant in the model check. So the models are unlikely to be oversmoothed, and the results from Speakers 5 and 6 were not excluded.

The divergence points determined by the GAMMs analysis from F2 and $f_0$ trajectories were similar to previous reports (Liu et al., 2022; Xu and Gao, 2018; Xu and Liu, 2007). On average, F2 started to diverge for the second syllable at 0.137 norm-time[6] (95 % CI = ±0.032, SD = 0.035), which is 0.1294 s before the conventional boundary in real time, as shown in Fig. 9. The divergence of $f_0$ contours was also similar. The point of divergence occurred at 0.145 norm-time on average (95 % CI = ±0.018, SD = 0.082), which is 0.1202s before the conventional boundary. Overall, the divergence timings of both F2 and $f_0$ were almost identical across speakers. Divergence may occur at any point within 42

[4] R syntax for the LMEM: model<- lmer(time~type + (1|speaker)+ (1| contrast), data= data, REML= FALSE)

[5] R syntax for the Bayesian model: model <- brm(onset~type + (1| contrast) + (1|speaker), data=data, family=gaussian(), prior=prior, warmup=2000, iter=7000, save_pars=save_pars (all=TRUE), control=list (adapt_delta =.99).

[6] The conversion rate of norm-time: 0.01 norm-time = 4% syllable length. Norm-time data in this study use 'norm-time' as the unit and follows the specified ratio.

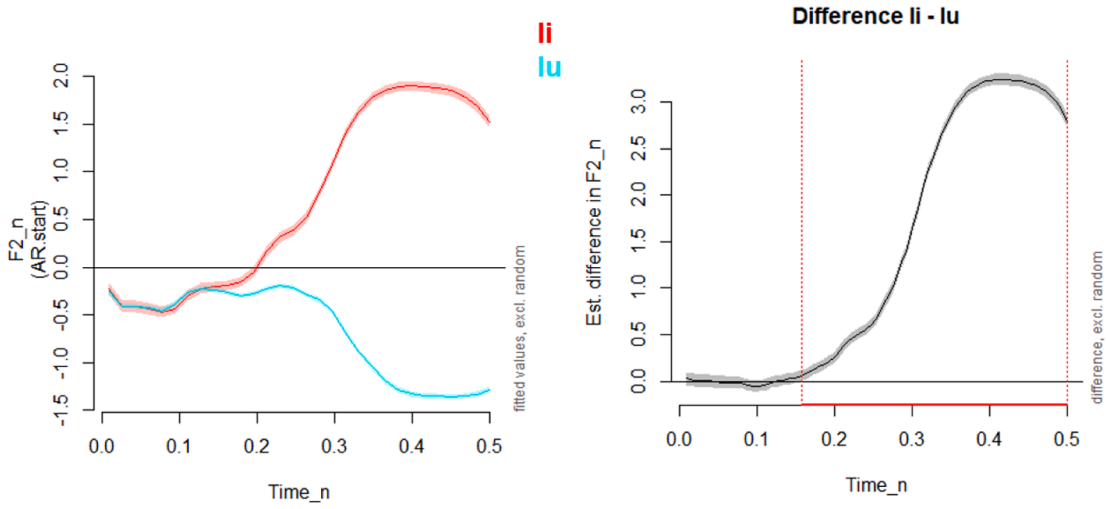



**Fig. 4.** Left: GAMMs fitted for /mali/ and /malu/ produced by Speaker 3. The shaded portions represent 95 % confidence intervals. Right: F2 difference between /mali/ and /malu/ over time. The red dotted vertical lines marked region(s) of significant difference.

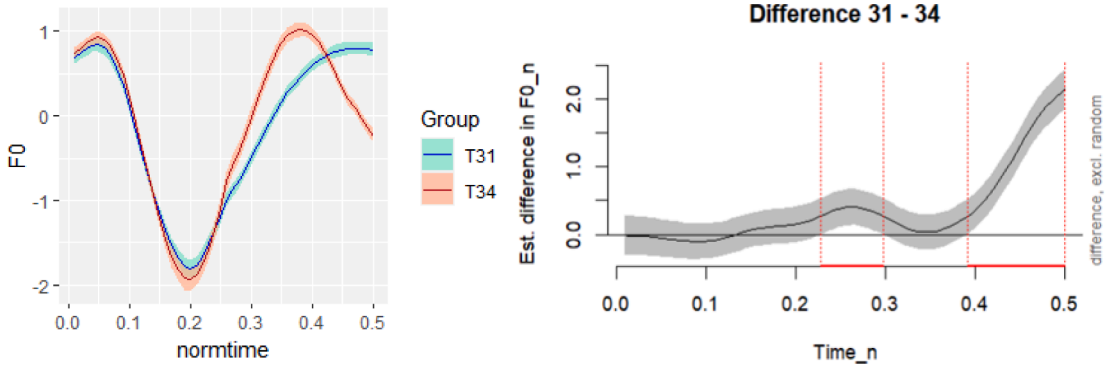



**Fig. 5.** Left: mean $f_0$ contours of Tone 31 and Tone 34. The shaded portions represent 95 % confidence intervals. Right: difference contour of Tone 31 vs Tone 34 produced by Speaker 6.

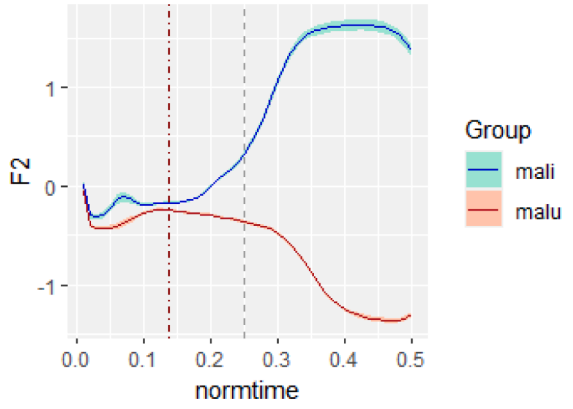



**Fig. 6.** Time-normalized mean F2 trajectory of the contrastive disyllabic vowel sequences by 7 speakers plotted on a normalized time scale. The mean contours are z-score normalized and averaged across all repetitions by each speaker. The shaded ribbons around the mean trajectories mark the 95 % confidence intervals. The vertical dash-dot line marks the divergence point detected by the GAMMs analysis; the dashed grey line marks the conventional boundary between the two syllables.

%−67 % of the first conventional syllable, and there was much overlap in the timing of F2 and $f_0$ onsets despite some differences. Whether such difference was statistically significant is explained in the following section.

### 3.1.1. *Comparison of F2 and $f_0$ divergence points*

The timing difference between $f_0$ and F2 divergence points was nonsignificant as found by LMEMs, with the $t$ value of −0.233. Given that tonal combinations that end with Tone 4 involve two similar-direction trajectories, while F2 movement in /li/ and /lu/ are always in opposite directions, another comparison was made between F2 and $f_0$ divergence points with the exclusion of those tone pairs where $f_0$ toward the second tone move in similar directions (T), so as to reduce uncertainty. The timing difference remained nonsignificant. The LMEM produced a $t$ value of 0.288, and the ANOVA test rendered a Chi-square value of 0.083 and a $p$ value of 0.774.

As the divergence can also be grouped according to the directional difference at syllable boundary, as shown in Table 3, a comparison of onset time was also made between similar-direction and opposite-direction $f_0$ trajectories. Fig. 10 shows the divergence points of the similar direction and opposite direction groups. The average divergence time was at 0.177 norm-time (for the similar direction group (95 % CI = ±0.034, SD = 0.087), which is 0.0836 s before the conventional boundary, and 0.129 norm-time for the opposite direction group (95 % CI = ±0.018, SD = 0.071), which is 0.1385s before the conventional boundary. The $t$ value reached 2.74 in the LMEM, with Chi-square at 6.953 and $p$ value at 0.008 in ANOVA. Hence, although the difference between $f_0$ and F2 onset was nonsignificant overall, the directional difference at syllable boundary did have a significant effect on the estimated divergence point.

Interestingly, the type of tonal target (dynamic vs. static) in syllable 1 did not have a significant effect. The $t$ value of the LMEM model was

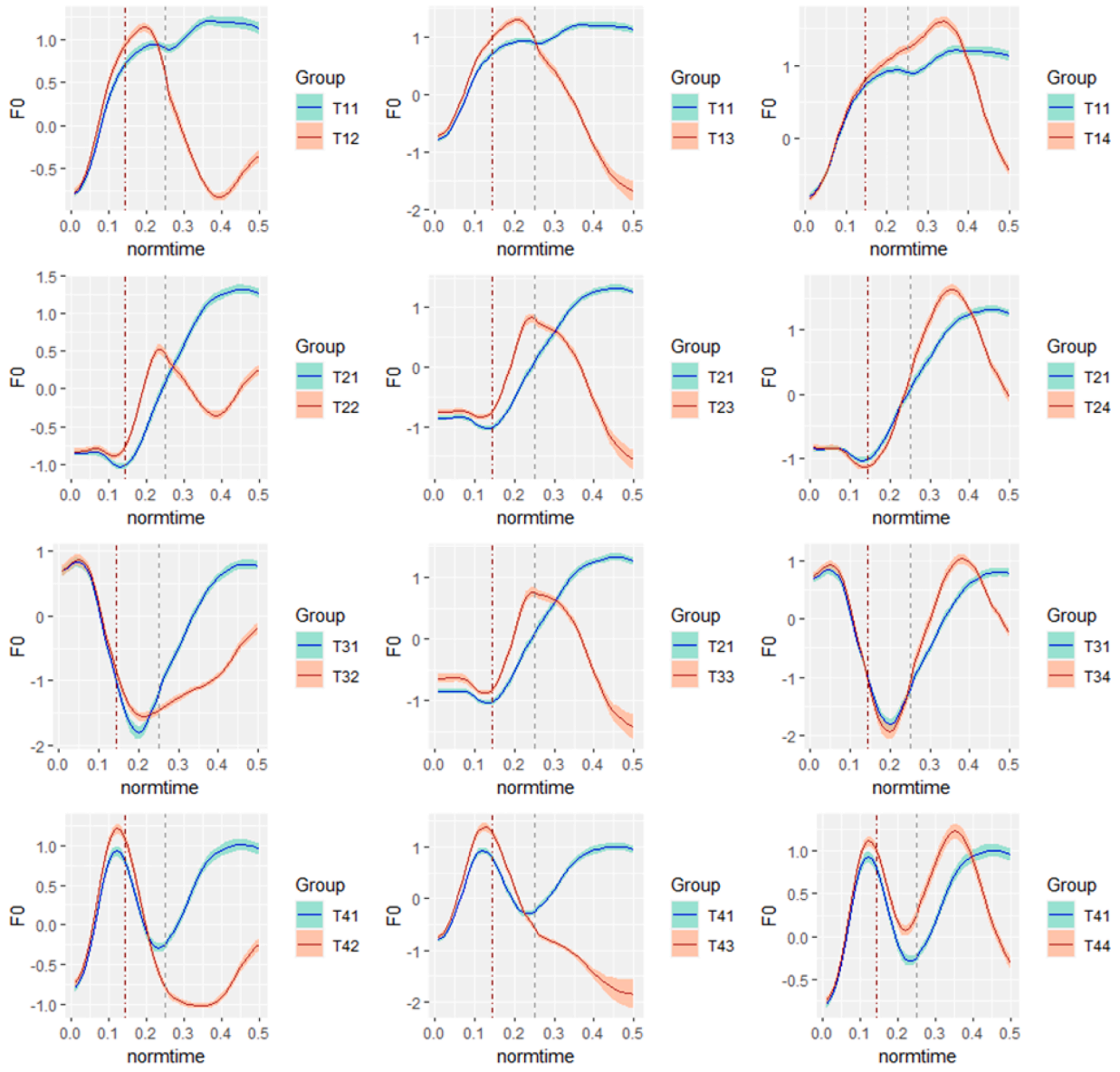



**Fig. 7.** Mean $f_0$ contours of the contrastive disyllabic tone sequences plotted on a normalized time scale. The mean contours are z-score normalized and averaged across all repetitions by each of the 7 speakers. The shaded ribbons represent 95 % confidence intervals. The vertical dash-dot line marks the divergence point detected by the GAMMs analysis; the dashed grey line marks the conventional boundary between the two syllables.

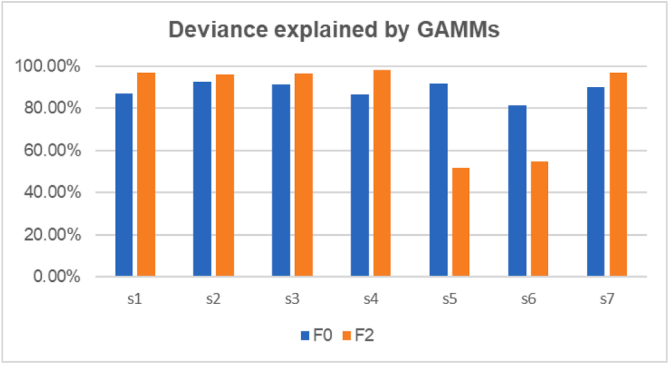



**Fig. 8.** Percentage of explained deviance by fitted GAMMs for each speaker.

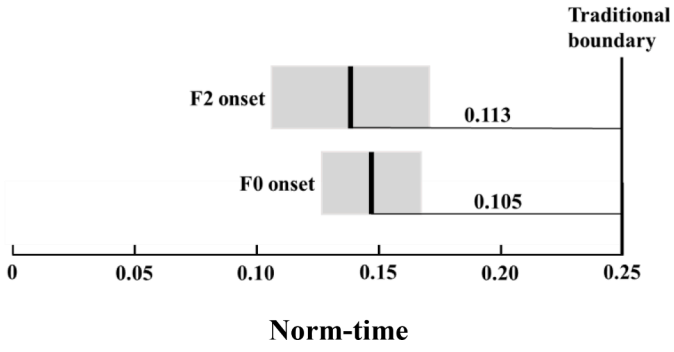



**Fig. 9.** F2 and $f_0$ divergence points on normalized time. The bold vertical bars in the middle of the shaded boxes represent the means, while the widths of the boxes represent the 95 % confidence intervals. The numbers above the horizontal lines show distances between mean onset times and conventional syllable boundaries.

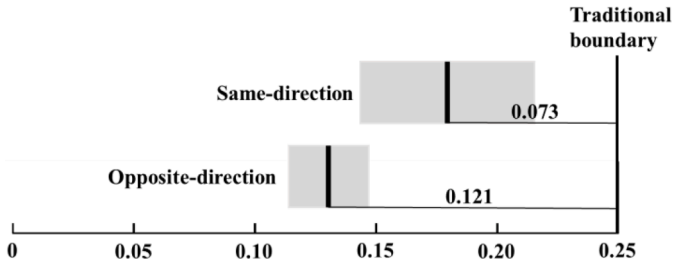



**Fig. 10.** Divergence points of similar direction and opposite direction groups plotted on the normalized time scale. The bold vertical lines represent the means, and the grey boxes represent the 95 % confidence intervals. The numbers above the horizontal lines show the distance between the mean onset time and the conventional onset of syllable 2.

−0.124, and ANOVA comparison produced a nonsignificant $p$ value of 0.9011. However, divergence time differed significantly between dynamic and static targets in the second syllable. The $t$ value was −2.593, and the $p$ value was 0.013, indicating that divergence points occurred later on the normalized time scale for the dynamic targets (0.160, SD = 0.082, 95 % CI = ±0.022) than for the static targets (0.113, SD = 0.083, 95 % CI = ±0.032).

### 3.1.2. Bayesian analysis of F2 and $f_0$ alignment

Fig. 11 shows the results of the Bayesian analysis for the alternative model, i.e., onset asynchrony (top panel), and the null model, i.e., onset synchrony (bottom panel). There were no divergent transitions reported in the modelling and diagnosis process (which would have been displayed as green lines in the right panels (Bürkner, 2021)). So the models for Bayesian Analysis are successfully constructed. The Bayesian analysis showed that the differences between $f_0$ and F2 had an estimate of 0.01 (b_type_contrast in Fig. 11, 95 % CI = ±0.08). The intercepts for the null model and the alternative model were also very close. The estimate was 0.14 for both models. The 95 % CI was ±0.04 for the null model and ±0.02 for the alternative model. The Bayes factor was 252.84 in the comparison, which is well beyond the threshold of 3. The results therefore provide strong support for onset synchrony over the normalized time scale.

## 3.2. Verification of F2-$f_0$ synchrony in real time

The trajectory comparison done on normalized time may be questionable, however. Time normalization may have masked some of the variability in the original real time, especially given the cross-speaker variability in speech rate. To verify the validity of the results in normalized time, the GAMM-detected F2 and $f_0$ divergence points were restored back to real time using formula (2).

$$\begin{cases} t_R = \dfrac{t_N}{t_{NS}} \times t_{RS1} \quad (0 \le t_N \le t_{NS}) \\ t_R = t_{RS1} + \dfrac{t_N - t_{NS}}{t_{NS}} \times t_{RS2} \quad (t_N > t_{NS}) \end{cases} \tag{2}$$

where $t_R$ is real-time divergence time, $t_N$ is normalized divergence time, and $t_{NS}$ is the length of the normalized syllable. $t_{RS1}$ and $t_{RS2}$ are the durations of the first and second target syllable in real-time. The conversion was applied to every repetition of every speaker, rendering 320 F2 divergence times and 240 $f_0$ divergence points. Thus, a total of 560×7 = 3920 divergence times were restored.

### 3.2.1. F2 and $f_0$ divergence points

The restored real-time divergence points were similar to the normalized ones, as shown in Fig. 12. The average divergence time was 0.1568 s (95 % CI = ±0.001, SD = 0.030) for F2, which is 0.1294 s before the conventional syllable boundary (95 % CI = ±0.002, SD = 0.049), and 0.1660 s (95 % CI = ±0.005, SD = 0.099) for $f_0$, which is 0.1202 s before the conventional syllable boundary. These values are very close to the mean divergence points based on normalized time presented earlier. There was a marked decrease in the confidence interval of both F2 and $f_0$ divergence times compared to those of normalized time in Fig. 8, so that the time ranges of F2 and $f_0$ were no

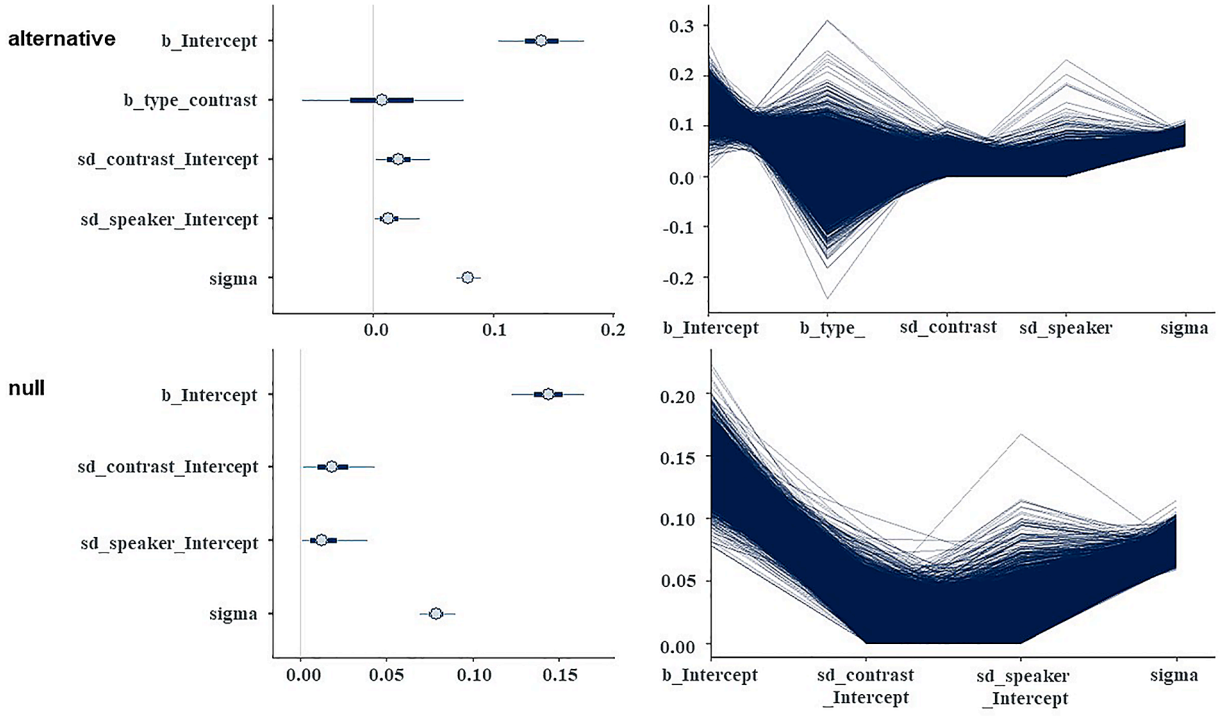



**Fig. 11.** Left: results of Bayesian analysis. The light horizontal bars represent 50 % confidence intervals, and the dark horizontal bars represent the 95 % confidence interval. The means are represented by unfilled circles. Right: Parallel coordinate plots, with each line representing one iteration.

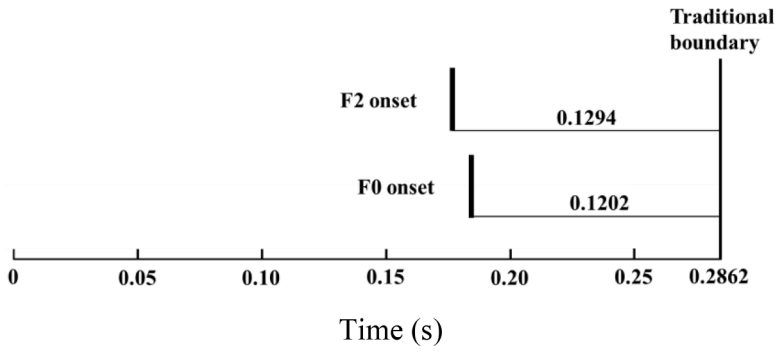



**Fig. 12.** Real-time F2 and $f_0$ divergence points. The bold vertical bars represent the means. The 95 % confidence intervals, however, are too narrow to be visible. The numbers above the horizontal lines show distances between the mean onset time and the conventional syllable boundaries.

longer overlapped.

*3.2.2. Alignment of $f_0$ and F2 divergence points*

Despite the non-overlap of $f_0$ and F2 divergence points, the statistical difference between them was nonsignificant. The LMEMs produced a $t$ value of –0.41, and the Chi-square and $p$ values were 0.167 and 0.682, respectively. Similarly, timing difference was also nonsignificant for tone groups involving movement of opposite directions. The $t$ value was 0.002, while Chi-square and $p$ values were at 0 and 0.998, respectively. Overall, the results in real and normalized times are consistent with each other, and no significant differences in the timing of divergence points were found in either time scale.

The difference between the similar direction and opposite direction groups shown in Table 3 was also calculated on the real time scale. Fig. 13 shows the divergence points of the two direction groups. The divergence points in the opposite direction group are much earlier than those of the similar direction group. The mean divergence points in the same and opposite direction groups were 0.2026 s (95 % CI = ±0.008, SD = 0.099) and 0.1477 s (95 % CI = ±0.002, SD = 0.060), respectively. The $t$ value of the difference between the two groups was 2.61, and the ANOVA tests showed a Chi-square value of 5.456 and a $p$ value of 0.019, indicating that the timing difference between the two direction groups is significant. This is in line with the results on the normalized time reported in 3.1.2.

Similar to the results on normalized time, the type of tonal target (dynamic vs. static) in syllable 1 also did not have a significant effect. The $t$ value was −0.147, and the corresponding $p$ value was 0.883. It was the target type of syllable 2 that led to a significant difference. The average divergence time for dynamic and static targets were 0.1825 s (SD = 0.099, 95 % CI = ±0.006) and 0.131 s (SD = 0.098, 95 % CI = ±0.008), respectively. The divergence points of static targets in syllable 2 were significantly earlier, with a $t$ value of –2.500 and a $p$ value of 0.026.

*3.2.3. Bayesian analysis of F2 and $f_0$ divergence points in real time*

Fig. 14 shows the results of the Bayesian analysis for the alternative model, i.e., onset asynchrony (top panel), and the null model, i.e., onset synchrony (bottom panel). Three divergent transitions were reported in the diagnosis for the null model (the green lines (Bürkner, 2021) in the bottom right panel). The number of divergence transitions was low, and did not indicate a single-point convergence of the estimated parameters (Gabry et al., 2019), the null model was considered reliable. These results of the Bayesian analysis of real-time data were in line with those of the norm-time data. The estimated effect of onset type was also 0.01 (95 % CI = ±0.09), very close to 0. The estimated intercepts were 0.15 (95 % CI = ±0.05) and 0.16 (95 % CI = ±0.03) for the alternative and null models, which suggests a limited difference. The Bayes factor was 229.79, which again well exceeds the threshold of 3, indicating strong support for tone-vowel synchrony.

## 4. Discussion

The results of the study have provided answers to the two research questions raised in the Introduction: (1) Is tone articulation synchronized with the CV syllable? (2) Does 'pre-low raising' of tone occur before or within the articulatory syllable? For the first question, we used disyllabic sequences consisting of one vowel minimal pair and 12 tone minimal pairs to determine the articulatory onsets of vowel and tone, respectively. The vowel minimal pair contrasted the vowels of the second syllable (/mali/ vs. /malu/), while the tone minimal pairs contrasted the tones of the second syllable (Tables 2-3). GAMMs were fitted to detect the time point when F2 and $f_0$ start to diverge towards the vowel and tone targets, respectively. LMEMs were then applied to compare the F2 and $f_0$ divergence points detected by GAMMs. ANOVA comparison of LMEMs found the timing difference between F2 and $f_0$ divergence points to be nonsignificant. To assess if the non-significant difference between the vowel-tone divergence points is merely null results due to insufficient data or due to underlying synchrony, Bayes

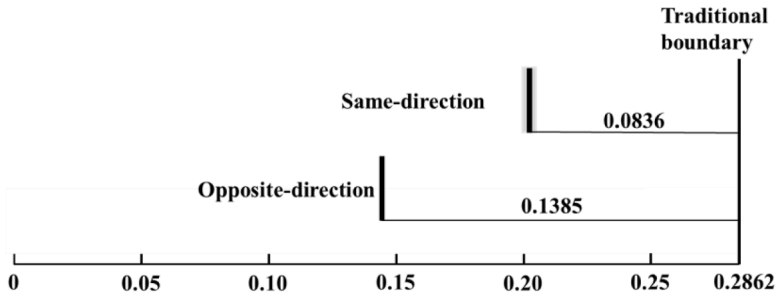



**Fig. 13.** Divergence points of similar direction and opposite direction groups plotted on real time scale. The bold vertical lines represent the means, and the grey boxes represent the 95 % confidence intervals (invisible for the opposite direction group because it is too narrow). The numbers above the horizontal lines show the distance between the mean onset time and the conventional onset of syllable 2.

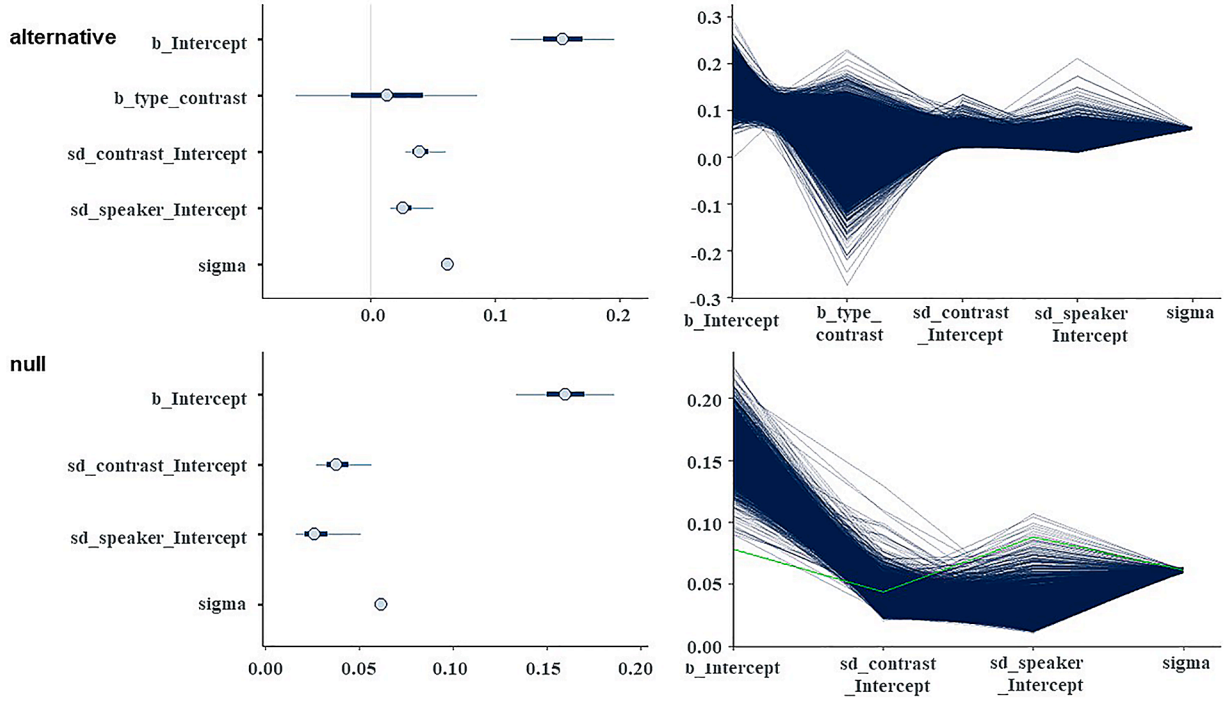



**Fig. 14.** Left: results of Bayesian analysis. The light horizontal bars represent 50 % confidence intervals, and the dark horizontal bars represent 95 % confidence intervals. The means are represented by the unfilled circles. Right: Parallel coordinate plots, with each line representing one iteration. The green lines in the null model represent the divergent transitions reported in the diagnosis (Bürkner, 2021).

factors were calculated, and the results showed that the synchrony hypothesis was over 200 times more likely for both time-normalized and real time divergence points. Therefore, following the first quantitative evidence for CV co-onset (Liu et al., 2022), we now also have quantitative evidence for onset synchronization of vowel and tone. And given the synchrony of vowel and consonant onsets (Liu et al., 2022; Liu and Xu, 2021, 2023), the current results suggest that tone onset is synchronized with the consonant onset as well. So, the answer to the first research question is affirmative. In other words, there is now quantitative evidence for full CVT co-onset as proposed in the synchronization model of the syllable (Xu, 2020; Xu and Liu, 2006). Such fully synchronized CVT onset is consistent with the hypothesis that the nature of the syllable is to synchronize segmental and suprasegmental articulations (Ohala, 1992; Ohala and Kawasaki, 1984), and with the suggestion that such synchronization eliminates most of temporal degrees of freedom to make the central neural control of multi-dimensional articulation possible for human speech (Xu, 2020).

The tonal alignment results of the present study were obtained with a quantitative method that differs from previous methods used to assess tonal onset. Gao (2009), for example, used maximum and minimum $f_0$ points as indicators of the onset and offset of tonal gestures. As is already known, however, $f_0$ turning points are highly variable due to the influence of various confounding factors (Xu, 1997, 1998; Xu and Wang, 2001). The best way to control these confounding factors is to apply the principle of minimal contrast all the way down to a direct comparison of movement trajectories (Bell-Berti and Harris, 1979, 1981; Gelfer et al., 1989). In this way, the influence of the confounding factors, both known and unknown, would apply to all members of the contrasting pair, and the observed difference is safely attributable to the intrinsic differences. Also, because the same method was used in this study to assess both vowel and tone onset time, the current results are free of potential confounds due to the use of different assessment methods as seen in Gao (2009), where C and V onsets were assessed by velocity threshold, but tone onset was assessed by $f_0$ turning points.

Also compared to previous methods of using running *t*-tests to detect event onsets (Xu et al., 2004), there is an advantage of GAMMs. As pointed out by Liu et al. (2022), the former method is more prone to Type I error due to repeated *t*-tests. GAMMs, in contrast, only create one model per minimal pair, which would greatly reduce the risk of Type I error.

### 4.1. *Increased complexity of tone articulation*

In regard to the second research question, the common onset of vowel and tone determined in the present study has shifted the CVT synchronized syllable onset leftward from the conventional acoustic onset by over 125 ms in real time, or 43.6 % in normalized time. This has introduced a complication to the mechanism of tone articulation, as predicted in the Introduction. That is, the previously reported pre-low raising effect is no longer 'anticipatory' across syllables, as it is now shown to be mostly happening *after* rather than *before* the common onset of the CVT syllable, as can be seen in Fig. 7. The patterns can be divided into three different cases as specified below with reference to Fig. 7.

Case 1—Columns 1–2, rows 1–2. Tone of syllable 1 is High (T1) or Rising (T2): $f_0$ peak around normalized time 0.2 is higher when the tone of syllable 2 is Rising or Low than when it is High. But these peaks are well after the divergence points marked by the first vertical line, and so are no longer anticipatory.

Case 2—Row 3. Tone of syllable 1 is Low (T3). No raising effect can be seen, as the Low tone is not subject to anticipatory raising (Xu, 1997, 1999). The only exception is Column 2, Row 3, where the Low tone is changed into Rising tone due to tone sandhi (Chao, 1968).

Case 3—Column 3, Rows 1–3. Tone of syllable 2 is Fall (T4): The $f_0$ peak seems to be raised compared to the contrasting High tone (T1). This is a case previously reported to also show anticipatory raising (Xu, 1997). But given that the $f_0$ peak occurs rather late, around 40 % into the second conventional syllable, it now seems that the raised peak is more likely due to the intrinsically higher $f_0$ of T4 rather than T1 as is the case in the canonical forms of the two tones (Xu, 1997). In other words, thanks to the leftward shift of the syllable boundary, T4 no longer seems to exert a pre-low raising effect.

Case 4—Row 4. Tone of syllable 1 is Fall (T4): The $f_0$ peak is higher

when the tone of syllable 2 is Rise (T2—column 1) or Low (T3—column 2) than when it is High, indicating a raising effect. The peak, however, occurs before the tone divergence point marked by the first vertical line. This might suggest that the raising effect is anticipatory, but it needs to be pointed out that the divergence point is not necessarily the actual onset of the tone. As shown in Liu et al. (2022), the acoustically detected divergence point is roughly 60 ms later than the articulatorily detected divergence point based on EMMA data, which is likely due to the greater amount of noise intrinsic to acoustic data that would tend to delay the first detection of the divergence of the trajectories. So, a further leftward shift of the divergence point would mean that the raised $f_0$ peak is again within rather than before the CVT articulatory syllable.

Case 5—Column 3, Row 4. Here the early $f_0$ peak is slightly higher when the tone of syllable 2 is Fall than when it is High. But the height difference is small, so it is unclear whether there is a real raising effect. Data reported in Xu (1997, 1999) do not seem to show an anticipatory raising effect by a following T4 on the preceding T4. So this case needs to be clarified in future studies.

Overall, therefore, given the new syllable onset detected by the synchronized divergence points, pre-low raising now seems to occur within rather than before the syllable that carries the triggering tone. This would mean that the raising is part of the articulation of the tone itself rather than its preparation. This finding might have important implications for the modeling of tone articulation as a simple target approximation process (Xu and Wang, 2001). At first glance, a revision of the target approximation model seems to be needed. But another possibility is to assume that there can be an optional initial target added before the main target. This initial target is set to increase the distance of the articulation movement by modifying the target onset in the opposite direction of the final goal. As argued in Xu (2020), a movement requiring a high velocity as its goal, e.g., a badminton smash, may involve an initial move in the opposite direction of the final velocity to increase the traveling distance of the racket. In the case of pre-low raising, the need to increase the traveling distance is evoked by the extra effort required to lower $f_0$ beyond the medial level (Atkinson, 1978; Erickson et al. 1995; Ohala, 1972). However, the distance increasing strategy in a badminton smash may consist of two phases. The first is a slow backward movement of the whole arm, and the second is a rapid backward flexion of the wrist right before the forward movement of the racket. What the current finding suggests is that the laryngeal equivalence of this backward wrist flexion may occur within the temporal domain of a tone itself. To the extent that tone articulation is a motor skill just like other motor activities, the finding may have broader impications on motor control in general (Xu, 2020). But that is a topic beyond the scope of the present paper.

A further caveat is that, while the leftward shift of the syllable boundary from the conventional acoustic boundary may look remarkable —over 125 ms in real time, or 43.6 % in normalized time, it may still be short of revealing the true articulatory onset of either vowel or tone. As mentioned earlier, the articulatory divergence points based on EMMA data are roughly 60 ms earlier than the acoustically detected divergence point (Liu et al., 2022). In fact, even articulatory divergence is just the result of muscle activities that must have commenced sometime even earlier. Therefore, what the divergence points in the present study indicate is that both tone and vowel should have at least started by that time.

## 5. Conclusion

The present study has applied a recently developed segmentation method (Liu et al., 2022) to investigate the temporal alignment of tone and vowel in Mandarin syllables. We first determined the onsets of vowel and tone by detecting the divergence points in F2 and $f_0$ trajectories of contrastive vowel and tone pairs using GAMMs, and then established their synchrony by LMEMs as well as Bayes factor analysis. The results show clear evidence of full synchrony of articulatory onsets of tone and vowel in Mandarin syllables. Given the existing evidence of consonant-vowel synchrony at the syllable onset (Liu et al., 2020; Liu and Xu, 2021, 2023), the current results therefore provide evidence for full synchrony of consonant, vowel and tone onsets in Mandarin syllables. This finding lends further support for the synchronization model of the syllable that views it as a synchronization mechanism that freezes most of the temporal degrees of freedom to enable simultaneous articulation of consonantal, vocalic and tonal elements of speech (Xu, 2020).

A notable corollary of the current finding is that the newly determined common onset of tone and vowel has shifted the syllable boundary earlier from the conventional boundary by at least 125 ms in real time, or 43.6 % in normalized time. This change has relocated the previously observed 'anticipatory raising' effect, namely, the raising of $f_0$ peak before a low-pitched tone, to within rather than before the syllable that carries the low-pitched tone. This finding may have significant implications for the understanding of not only tone articulation, but also other motor skills that require similar preparatory actions.

## Author statement

We declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## CRediT authorship contribution statement

**Weiyi Kang:** Writing – original draft, Visualization, Formal analysis, Data curation. **Yi Xu:** Writing – review & editing, Supervision, Software, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgement

We would like to thank Dr. Zirui Liu for advice on model-fitting, Keyue Chen for helping to arrange the database, and Yiwen Fan for debugging the R program for Bayesian analysis.

## References

Atkinson, J.E., 1978. Correlation analysis of the physiological factors controlling fundamental voice frequency. J. Acoustical Soc. Am. 63, 211–222.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.
Bell-Berti, F., Harris, K.S., 1979. Anticipatory coarticulation: some implications from a study of lip rounding. J. Acoustical Soc. Am. 65 (5), 1268–1270.
Bell-Berti, F., Harris, K.S., 1981. A temporal model of speech production. Phonetica 38, 9–20.
Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5 (9/10), 341–345.
Bürkner, P., 2021. Bayesian item response modeling in R with brms and Stan. J. Stat. Softw. 100 (5), 1–54. https://doi.org/10.18637/jss.v100.i05.
Carney, P., Moll, K., 1971. A cinefluorographic investigation of fricative consonant-vowel coarticulation. Phonetica 23, 193–202.
Chao, Y.R., 1968. A Grammar of Spoken Chinese. University of California Press, Berkeley, CA.
Coupé, C., Oh, Y.M., Dediu, D., Pellegrino, F., 2019. Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. Sci. Adv. 5 (9), eaaw2594.
Dell, G.S., 1988. The retrieval of phonological forms in production: tests of predictions from a connectionist model. J. Mem. Lang. 27 (2), 124–142.

DiCanio, C., Amith, J.D. and Castillo García, R. (2014). The phonetics of moraic alignment in Yoloxóchitl Mixtec. TAL 2014. Nijmegen: 203–210.
Dienes, Z., 2016. How Bayes factors change scientific practice. J. Math. Psychol. 72, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003.
Erickson, D., Honda, K., Hirai, H., Beckman, M.E., 1995. The production of low tones in English intonation. J. Phon. 23, 179–188.
Fowler, C., 1981. Production and perception of coarticulation among stressed and unstressed vowels. J. Speech. Hear. Res. 24, 127–139.
Fujimura, O., 1994. C/D Model: a computational model of phonetic implementation. In: Ristad, E.S. (Ed.), Language and Computations. American Math Society, Providence, RI, pp. 1–20.
Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A., 2019. Visualization in Bayesian Workflow. J. R. Statistic. Soc. Series A: Statis. Soc. 182 (2), 389–402. https://doi.org/10.1111/rssa.12378.
Gandour, J., Potisuk, S., Dechongkit, S., Ponglorpisit, S., 1992. Anticipatory tonal coarticulation in Thai noun compounds. Linguistics Tibeto-Burman Area 15, 111–124.
Gandour, J., Potisuk, S., Dechongkit, S., 1994. Tonal coarticulation in Thai. J. Phon. 22, 477–492.
Gao, M., 2009. Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. Chin. J. Phonetics 2, 43–50.
Gelfer, C.E., Bell-Berti, F., Harris, K.S., 1989. Determining the extent of coarticulation: effects of experimental design. J. Acoustical Soc. Am. 86 (6), 2443–2445.
Goldsmith, J.A., 1976. Autosegmental Phonology. MIT. Ph. D. dissertation.
Goldstein, L., Byrd, D., Saltzman, E., 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. Action Language via Mirror Neuron Syst. 215–249.
Honda, K., Hirai, H., Masaki, S., Shimada, Y., 1999. Role of vertical larynx movement and cervical lordosis in $f_0$ Control. Lang. Speech. 42, 401–411.
Hyman, L.M., 2011. Tone: is it different? In: Goldsmith, John, Riggle, Jason, Yu, Alan C. L. (Eds.), The Handbook of Phonological Theory, 2nd ed. Blackwell, Oxford, pp. 197–239.
Jespersen, O., 1899. Fonetik: En systematisk Fremstilling Af Læren Om Sproglyd. Det Schøbergse Forlag, København.
Kingston, J., 2007. Segmental influences on F0: automatic or controlled? Tones Tunes 2, 171–210.
Kozhevnikov, V.A., Chistovich, L.A., 1965. Speech: Articulation and Perception. Translation by Joint Publications Research Service. JPRS, Washington, DC, p. 30543.
Kühnert, B., Nolan, F., 1999. The origin of coarticulation. In: Hardcastle, W.J., Newlett, N. (Eds.), Coarticulation: Theory, Data and Techniques. Cambridge University Press, Cambridge, pp. 7–30.
Ladefoged, P., 1967. Three Areas of Experimental Phonetics. Oxford University Press, London.
Ladefoged, P., 1982. A Course in Phonetics. Hartcourt Brace Jovanovich, New York.
Lakens, D., McLatchie, N., Isager, P.M., Scheel, A.M., Dienes, Z., 2020. Improving inferences about null effects with bayes factors and equivalence tests. J. Gerontol. - Series B Psychol. Sci. Soc. Sci. 75 (1), 45–57. https://doi.org/10.1093/geronb/gby0653.
Laniran, Y., 1992. Intonation in Tone Languages: The phonetic Implementation of Tones in Yorùbá. Cornell University. Ph.D. Dissertation.
Laniran, Y.O., Clements, G.N., 2003. Downstep and high raising: interacting factors in Yoruba tone production. J. Phon. 31, 203–250.
Lehiste, I., Peterson, G.E., 1961. Some basic considerations in the analysis of intonation. J. Acoustical Soc. Am. 33, 419–425.
Levelt, W.J., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Behav. Brain Sci. 22 (1), 1–38.
Liberman, M., Schultz, J.M., Hong, S., Okeke, V., 1993. The phonetic interpretation of tone in Igbo. Phonetica 50, 147–160.
Liu, Z., Xu, Y., 2021. Segmental alignment of English syllables with singleton and cluster onsets. In: Proceedings of Interspeech 2021. Brno, Czech Republic.
Liu, Z., Xu, Y., 2023. Deep learning assessment of syllable affiliation of intervocalic consonants. J. Acoustical Soc. Am. 153 (2), 848–866.
Liu, Z., Xu, Y., Hsieh, F., 2022. Coarticulation as synchronised CV co-onset – parallel evidence from articulation and acoustics. J. Phon. 90 https://doi.org/10.1016/j.wocn.2021.101116, 101116–.
Lobanov, B.M., 1971. Classification of Russian vowels spoken by different speakers. J. Acoust. Soc. Am. 49 (2B), 606–608. https://doi.org/10.1121/1.1912396.
MacNeilage, P.F., 1998. The frame/content theory of evolution of speech production. Behav. Brain Sci. 21, 499–546.
Menzerath, P., de Lacerda, A., 1933. Koartikulation, Seuerung und Lautabgrenzung. Fred. Dummlers.
Mok, P.P.K., 2012. Effects of consonant cluster syllabification on vowel-to-vowel coarticulation in English. Speech. Commun. 54 (8), 946–956. https://doi.org/10.1016/j.specom.2012.04.001.
Nam, H., Goldstein, L., Saltzman, E., 2009. Self-organization of syllable structure: a coupled oscillator model. Approaches to Phonological Complexity. De Gruyter Mouton.
Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., Bürkner, P.-C., 2019. An introduction to Bayesian multilevel models using brms: a case study of gender effects on vowel variability in standard Indonesian. J. Speech, Language, Hear. Res. 62 (5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006.
Ohala, J.J., 1972. How is pitch lowered? J. Acoustical Soc. Am. 52, 124.
Ohala, J.J., 1992. Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In: Proceedings of Chicago Linguistic Society 26. Chicago, pp. 319–338.
Ohala, J.J., Kawasaki, H., 1984. Prosodic phonology and phonetics. Phonology. 1, 113–127.
Öhman, S.E., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoustical Soc. Am. 39 (1), 151–168. https://doi.org/10.1121/1.1909864.
Remijsen, B., Ayoker, O.G., 2014. Contrastive tonal alignment in falling contours in Shilluk. Phonology. 31 (03), 435–462.
Saltzman, E.L., Munhall, K.G., 1989. A dynamical approach to gestural patterning in speech production. Ecol. Psychol. 1 (4), 333–382. https://doi.org/10.1207/s15326969eco0104_2.
Selkirk, E., 1982. The syllable. In: Hulst, H.V.d., Smith, N. (Eds.), The Structure of Phornological Representations (partII). Foris, Dordrecht, The Netherlands, pp. 337–383 pp.
Shaw, J.A., Chen, W.R., 2019. Spatially conditioned speech timing: evidence and implications. Front. Psychol. 10, 2726. https://doi.org/10.3389/fpsyg. 2019.02726.
Sievers, E., 1893. Grundzüge Der Phonetik zur Einführung in Das Studium der Lautlehere der Indogermanischen Sprachen, 4th ed. Breitkopf & Härtel, Leipzig.
Stetson, R.H., 1951. Motor Phonetics: A study of Speech Movements in Action. North Holland, Amsterdam.
Stone, J.V., 2013. Bayes' rule: A tutorial Introduction to Bayesian analysis. Sebtel Press.
Sun, Y., Poeppel, D., 2023. Syllables and their beginnings have a special role in the mental lexicon. Proc. Nat. Acad. Sci. 120 (36), e2215710120.
Tilsen, S., 2020. Detecting anticipatory information in speech with signal chopping. J. Phon. 82, 100996.
Turk, A., Nakai, S., Sugahara, M., 2006. Acoustic segment durations in prosodic research: a practical guide. In: Sudhoff, S., Lenertová, D., Meyeret, R. (Eds.), Methods in Empirical Prosody Research. De Gruyter, Berlin, New York, pp. 1–28 al.pp.
van Rij J., Wieling M., Baayen R., van Rijn H. (2022). "itsadug: interpreting Time Series and Autocorrelated Data Using GAMMs." R package version 2.4.1.
Whitney, W.D., 1865. The relation of vowel and consonant. J. Am. Oriental Soc. 8, 357–373.
Wood, S., 2017. Generalized Additive Models: An Introduction With R, 2 ed. Chapman and Hall/CRC.
Wright, C.E., 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. Mem. Cognit. 7 (6), 411–419. https://doi.org/10.3758/Bf03198257.
Xu, C.X., Xu, Y., 2003. Effects of consonant aspiration on Mandarin tones. J. Int. Phon. Assoc. 33, 165–181.
Xu, Y., 1993. Contextual Tonal Variation in Mandarin Chinese. The University of Connecticut. Ph.D. dissertation.
Xu, Y., 1997. Contextual tonal variations in Mandarin. J. Phon. 25, 61–83.
Xu, Y., 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. Phonetica 55, 179–203.
Xu, 1999. Effects of tone and focus on the formation and alignment of $f_0$ contours. J. Phon. 27 (1), 55–105. https://doi.org/10.1006/jpho.1999.0086.
Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech. Commun. 46 (3–4), 220–251.
Xu, Y., 2007. Speech as articulatory encoding of communicative functions. In: Proceedings of The 16th International Congress of Phonetic Sciences, pp. 25–30. Saarbrucken:.
Xu, Y., 2013. ProsodyPro — a tool for large-scale systematic prosody analysis. In: *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), pp. 7–10. Aix-en-Provence, France.
Xu, 2019. Prosody, tone, and intonation. The Routledge Handbook of Phonetics, 1st ed. Routledge, pp. 314–356. https://doi.org/10.4324/9780429056253-13.
Xu, Y. (2020). Syllable is a synchronization mechanism that makes human speech possible. *PsyArXiv*. https://doi.org/10.31234/osf.io/9v4hr.
Xu, Y., Gao, H., 2018. FormantPro as a tool for speech analysis and segmentation /FormantPro como uma ferramenta para a análise e segmentação da fala. Revista De Estudos Da Linguagem 26 (4). https://doi.org/10.17851/2237-2083.26.4.1435-1454.
Xu, Y., Larson, C.R., Bauer, J.J., Hain, T.C., 2004. Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. J. Acoustical Soc. Am. 116, 1168–1178.
Xu, Y., Liu, F., 2006. Tonal alignment, syllable structure and coarticulation: toward an integrated model. Italian J. Linguistics 18, 125–159.
Xu, Y., Liu, F., 2007. Determining the temporal interval of segments with the help of $f_0$ contours. J. Phon. 35, 398–420.
Xu, Y., Wang, Q.E., 2001. Pitch targets and their realization: evidence from Mandarin Chinese. Speech. Commun. 33, 319–337.
Xu, Y., Xu, A., 2021. Consonantal F0 perturbation in American English involves multiple mechanisms. J. Acoust. Soc. Am. 149 (4), 2877–2895.
Yi, H., Tilsen, S., 2016. Interaction between Lexical tone and intonation: an EMA Study. In: Proceedings of Interspeech 2016, pp. 2448–2452.
Yip, M., 2002. Tone. Cambridge University Press, Cambridge.
Zsiga, E., Nitisaroj, R., 2007. Tone features, tone perception, and peak alignment in Thai. Lang. Speech. 50 (3), 343–383.