Journal of Phonetics 108 (2025) 101389

Contents lists available at ScienceDirect

Journal of Phonetics

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

Thai speakers time lexical tones to supralaryngeal articulatory events

Francesco Burroni [a,*], Sam Tilsen [b]

[a] Ludwig-Maximilians-Universität München, Institute for Phonetics and Speech Processing (IPS), Spoken Language Processing Group, Akademiestraße 7, 80799 Munich, Germany
[b] Cornell University, Department of Linguistics, Morrill Hall 203, 14850, Ithaca, NY, USA

ARTICLE INFO

Article history:
Received 25 September 2023
Received in revised form 16 December 2024
Accepted 26 December 2024
Available online 10 January 2025

Keywords:
Articulatory timing
Segmental anchoring
Lexical tone
Stability
Variability
Mutual information
Thai

ABSTRACT

What do speakers do when they produce tones? Do they aim for the synchronization of f0 targets with segmental acoustic events, or do they execute a routine in which f0 changes and oral articulations are precisely coordinated? This paper explores these questions in Thai using acoustic and electromagnetic articulography data from eight speakers. Drawing on analyses of variability, stability, and informativity, our findings indicate that the timing of the onsets of tonal and oral articulatory gestures is generally more stable than the timing of tonal and oral targets, both in articulatory and acoustic measurements. When comparing the two modalities directly, we found that the lag between tonal onset and vocalic gesture onset exhibits the lowest variability and the highest mutual information among a large set of timing measures. Additionally, only articulatory lags remain stable under rate and context perturbations. To explain these findings, we propose that Thai tones are timed onset-to-onset with vocalic gestures and develop a model that formally implements this proposal. This model also accounts for otherwise puzzling acoustic patterns, such as a negative lag between tonal onset and acoustic syllable boundaries at slower speech rates. Further temporal patterns, such as surface non-zero time-locking rather than perfect synchrony of events, are also clarified. In sum, this work advances our understanding of tonal timing in Thai and outlines its implications for more general theories of phonology and speech production.

1. Introduction

Unlike speakers of non-tonal languages, speakers of tonal languages, such as Thai or Mandarin Chinese, need to produce changes in fundamental frequency (f0) over timescales roughly commensurate with the timescales of consonantal and vocalic articulation. Given the existence of lawful (relative) timing relationships between oral articulatory gestures in speech production (cf. Krause & Kawamoto, 2020 and references therein for an overview), we may reasonably hypothesize that f0 changes obey similarly stringent timing relationships. However, no consensus exists on the details of tonal timing in speech production.

Crucial questions that arise are the following: what other events are used to control the relative timing of tone production? Are these events associated with the acoustic signal, such as segmental boundaries? Or are they events associated with articulation, such as the onset of lingual movement for a vocalic posture?

Second, what specific (sub)-portions of events are coordinated with each other? Is it the initiation of tones and articulatory movements/acoustic signals or the achievement of their targets? Or both?

These issues pertain to an important area of investigation in the phonetic and phonological literature known variously as tonal “timing” or “alignment” or “anchoring” (Burroni, 2023a; Karlin, 2022; Xu, 1998; Yi, 2017; Zsiga, 2012). The study of tonal timing is dedicated to uncovering the strategies employed by speakers of different languages to synchronize modulations of f0, a main acoustic cue to lexical tones and pitch accents, with the unfolding of consonant and vowels constituting syllables and words.

In the phonological literature, tonal timing is treated as a problem of symbolic association between tonal primitives, such as high (H) and low (L) tones, with various phonological entities referred to as tone-bearing units (TBUs). These are syllables, vowels, or moras in different languages (Yip, 2002). However, in speech, phonological structures have to be realized in time, an issue that has quite naturally led researchers to probe the question of how tonal phonological associations are implemented phonetically in acoustic and

[*] Corresponding author.
E-mail address: francesco.burroni@phonetik.uni-muenchen.de (F. Burroni).

https://doi.org/10.1016/j.wocn.2024.101389
0095-4470/© 2024 The Author(s). Published by Elsevier Ltd.

articulatory signals (Burroni, 2023a; D'Imperio et al., 2007; Flemming & Cho, 2017; Karlin, 2022; Mücke et al., 2009; Niemann et al., 2011; Xu, 1998, 2009; Xu & Liu, 2006; Yi, 2017). In phonetic research, tones are often studied and equated with contours synchronized to syllable acoustic boundaries (Lee & Mok, 2021), yet a variety of phenomena like tonal coarticulation and peak delay are problematic under this widely held assumption because they challenge the synchronization of tones and acoustic syllable boundaries (Xu & Liu, 2006).

The question of the phonetic implementation of tonal timing embraces two different aspects in the literature. The first aspect, which we refer to as the modality of tonal timing, deals with whether tonal timing is controlled with respect to an acoustic or an articulatory modality in production. This question has received much attention, especially in the literature on pitch accents (e.g., D'Imperio et al., 2007b; Gao, 2008; Mücke et al., 2009; Niemann, 2016; Xu & Liu, 2006). The second aspect deals with what landmarks may be employed to time tones to other relevant speech events (Burroni, 2023a; Flemming & Cho, 2017; Gao, 2008; Karlin, 2022; Turk & Shattuck-Hufnagel, 2020b; Yi, 2014; Zsiga & Nitisaroj, 2007).

The issue of tonal timing modality and landmarks is also not isolated. Rather, the question of tonal timing can also help to assess more generally the role of acoustic and articulatory information in speech production; as well as how movements underlying speech production may be coordinated with each other. In this respect, tonal timing is a useful testing ground case for different theories of phonetic implementation, speech production, and timing in speech motor control.

In this paper, we attempt to answer the question of whether speakers of a tone language, Central Thai (tha, henceforth simply Thai), produce lexical tones by primarily controlling their timing relative to segmental acoustic or articulatory events, as well as how these events are coordinated to each other using sub-portions or landmarks.

To answer these modality and landmarks questions, we scrutinize a variety of possible articulatory and acoustic landmarks and coordination strategies that could be used by speakers to accomplish tonal timing. More specifically, we first investigate whether tonal events are more likely to start together with other acoustic or articulatory events, or to be produced such that speech production targets are achieved at the same time, or perhaps so that both onsets and targets are reached together, exhibiting full synchronization of both landmarks. Subsequently, we compare the most stable relative timing across modalities to probe which, if any, is more stably controlled by speakers. Answering these questions allows us to evaluate how tightly integrated tones may be in a feedforward plan routinely assumed for speech motor control and to probe the details of timing implementation across the oral and laryngeal domains. Additionally, evidence from tonal timing can inform and help refining the development of theories of speech production and speech motor control, which are often based on non-tonal languages.

The remainder of this paper is organized as follows. We first review how tones are produced via laryngeal adjustments that are largely independent of supralaryngeal adjustments, a consideration that leads us to consider the issue of how the two can be timed (1.1). We then discuss different phonological models that assume acoustic and articulatory timing (1.2) and their relationship to speech production models (1.3). In (1.4), we discuss how the issue of tonal timing modality can only be evaluated against specific predictions regarding which landmarks are involved in the control of timing. In subsection (1.5), we present our research questions and different hypotheses and predictions concerning tonal timing modality and landmarks. Specifically, we draw upon notions of variability, stability, and informativity partly established in previous work on tonal timing and partly extended to tonal timing in this article. Finally, in (1.6) we introduce the language investigated in this paper, Thai, and discuss why it represents a good testing ground for tonal timing. Section 2 presents our methodology. Section 3 presents our results comparing the stability and variability of landmarks within modality (3.1) and then across modality (3.2), where a mutual information analysis supplements our analyses of variability and stability. In Section 4, we discuss our findings and their implications for the question of tonal timing landmarks (4.1) and modality (4.2), where, in so doing, we also present a stochastic time model of tonal timing that formally implements our proposal and that can explain some otherwise puzzling patterns. We then discuss remaining issues regarding the role of synchronization in tonal timing (4.3) and the limitations intrinsic in this work (4.4). We conclude by summarizing our main findings (5).

1.1. Tone production and the issue of tonal timing

First, a question that we may ask is why speaker may even need to implement the timing of tones to other events. The answer is that vocal fold tension, underlying the production of tones, is controlled largely independently of oral articulatory movements underlying the production of consonants and vowels, an insight reflected in both the acoustic theory of speech production (Fant, 1971) and in autosegmental metrical phonology (Goldsmith, 1976). Given this independence, laryngeal and supralaryngeal articulation – which cannot be randomly timed in order for appropriate speech outputs to obtain – need to be coordinated by speakers in the production of any language, but especially in a tone language where each syllable or word comes with a lexical f0 specification.

We now present a brief overview of the mechanisms behind tonal production and then discuss their relationship to oral articulation (*cf.* Erickson, 2011; Fujisaki et al., 2004; Hirose, 2010; Honda, 2004; Story, 2015 for detailed overviews). Tones are produced by speakers through modulations of f0, an acoustic correlate of the rate of vocal fold vibration. Increases in vocal fold vibration rate result from a reduction in mass per unit area of the vocal folds and/or changes in transglottal pressure. Reduction of mass per unit area effectively means elongation or stretching of the vocal fold body and cover. Elongation of the vocal folds is produced primarily by contractions of an intrinsic laryngeal muscle, the cricothyroid muscle (CT). Contraction of the CT causes the thyroid cartilage to translate and rotate about the cricothyroid joint, resulting in an increased distance between the arytenoid and thyroid cartilages. Since the vocal folds insert anteriorly into the thyroid cartilage and posteriorly into the arytenoid cartilages, increasing their distance causes the vocal folds to stretch, effectively diminishing their mass per unit area and causing them to vibrate faster when excited by airflow.

The mechanisms behind decreases in f0 are more complex. Decreases in f0 can be brought about by intrinsic laryngeal muscle adjustments, specifically by a cessation of CT contraction that reduces the distance between the arytenoid and thyroid cartilages, causing the vocal folds to thicken and reduce their rate of vibration. Additionally, decreases in f0, especially in the lower f0 range, may also be brought about by changes in the vertical position of the larynx triggered by the contraction of external laryngeal muscles. Specifically, infrahyoid or strap muscles contraction causes a rotation of the cricoid cartilage about the cervical spine. This rotation reduces the distance between the arytenoid and cricoid cartilages, increasing the folds' mass per unit area, which lowers the frequency of vocal fold vibration (Honda et al., 1999). Although transglottal pressure could also be manipulated to effect changes in f0, there is little evidence that speakers directly use this alternative mechanism (Tanaka et al., 1997).

Regarding the timing of muscular adjustments and their acoustic consequences, it is important to note that laryngeal adjustments do not instantaneously affect f0. Electromyography (EMG) studies on speech and in-vivo studies on canine and feline excised larynxes suggest delays in the order of $\sim$ 50–100 ms for CT contraction and even longer ones for strap muscle contraction (*cf*. Erickson, 2011; Hoole, 2006 and references therein).

Given that both increases and decreases in f0 are produced primarily by intrinsic and extrinsic laryngeal muscle activity, their production can function largely, but not completely, independently of oral articulation of consonants and vowels, which involves primary muscles in the facial, oropharyngeal, and velopharyngeal regions. A small degree of interaction between laryngeal and supralaryngeal articulation does exist. These interactions are most clearly manifested in microprosodic perturbations (Hoole & Honda, 2011 and references therein) and in supralaryngeal adjustments due to tone production (Burroni et al., 2024; Erickson et al., 2004; Hoole & Hu, 2004; Shaw et al., 2016).

Despite these limited interactions between the two systems, however, it seems safe to conclude that laryngeal tension, reflected in f0, is mostly controlled by speakers of a tone language for the purposes of tone production and that this control is largely independent of supralaryngeal articulation. Given this substantial degree of autonomy between the two systems, speakers of a tonal language are faced with the task of coordinating laryngeal adjustments, underlying tones, with supralaryngeal movements, underlying the production of consonants and vowels, as these can (almost) freely combine.

1.2. Tonal timing modality: Phonological models

With respect to the modality of tonal timing, two hypotheses have been entertained regarding how speakers control it. These are segmental acoustic timing and gestural articulatory timing.

As noted in review work (Lee & Mok, 2021; Xu & Liu, 2006), most empirical phonetic research on lexical tone languages, like Mandarin Chinese (cmn) or Thai, studies speakers' production of lexical tones by extracting f0 contours over syllable-sized or vowel-sized spans. Tones are, thus, studied by (and equated with) f0 contours based on said segmental acoustic boundaries. For tonal timing, this means that full synchronization is simply assumed over a syllable or a word to which the tone is lexically associated. However, it is known that the alignment between tones and acoustic boundaries is imperfect due to effects such as peak delay into following syllables and coarticulatory effects between tones (Burroni, 2023a; Gandour et al., 1994; Xu, 2001; Xu & Lee, 2022; Xu & Liu, 2006); additionally, as already pointed out, EMG evidence also suggests an asynchronous CT or strap muscle activity onset (section 1.1) compared to the acoustic boundaries of a syllable.

In spite of the issues outlined above, the idea that tones are produced by aligning tonal onset or targets relative to acoustic boundaries was formalized in a framework known as the segmental anchoring hypothesis (SAH), which builds on earlier seminal work in autosegmental metrical phonology (Goldsmith, 1976; Pierrehumbert, 1980). The SAH is grounded in the empirical observation that acoustic studies on the timing of F0 peaks and valleys in tones and pitch accents have revealed systematic patterns in their alignment with anchoring sites within the (acoustic) segmental string. (e.g., Arvaniti et al., 1998; Atterer & Ladd, 2004; Ladd et al., 1999, 2000). Segmental anchoring as a timing pattern controlled by speakers was first studied in Arvaniti et al. (1998), who showed that Greek prenuclear accent targets, such as the f0 peak of a High (H) tone, are consistently aligned to an acoustic landmark, around $\sim$20 ms after the onset of the post-accentual vowel. These findings of precise alignment between tonal events and segmental anchors were later formalized as a hypothesis about control of the timing of tonal patterns in work by Ladd et al. (1999), where the authors showed that English rising-falling prenuclear accents, Low (L) + H, are well-aligned to specific anchors in the segmental string and that their slope is predictable from the alignment. Ladd et al. (1999) further reported findings in line with the notion that pitch accent alignment is stable across rates. Albeit primarily developed for post-lexical pitch accents, the idea of segmental anchoring was extended to languages with lexical pitch accents languages and tonal languages (e.g., Cho, 2010; Flemming & Cho, 2017). In fact, evidence compatible with the hypothesis of segmental anchoring has been reported in a wide variety of prosodic systems employing pitch accents and tones, for instance, in Dutch (Ladd et al., 2000; Schepman et al., 2006), German (Atterer & Ladd, 2004) Japanese (Ishihara, 2003), Italian (Petrone & Ladd, 2007), Mandarin (Cho, 2010; Flemming & Cho, 2017), Korean (Cho, 2010), Drehu (Torres & Fletcher, 2020), and Yoloxóchitl Mixtec (DiCanio et al., 2014), among others.

Despite wide adoption of the SAH, ensuing research into the SAH demonstrated that the timing of f0 movement onsets and targets to acoustic landmarks could be influenced by linguistic and paralinguistic factors such as syllable structure, speech rate, and diatopic variation (Atterer & Ladd, 2004; Flemming & Cho, 2017; Petrone & Ladd, 2007; Prieto & Torreira, 2007). Given these problems with maintaining an invariant notion of segmental anchoring, Ladd (2006) proposed to treat tonal f0 movements as gestures and their anchoring as part of a more general theory of gestural coordination. This proposal was also in line with mounting evidence that post-lexical pitch accents may be more stably anchored to oral articulatory gestures than acoustic targets in work that had been dedicated to answering this question (D'Imperio et al., 2007; Mücke et al., 2009; Niemann et al., 2011; Prieto et al., 2007).

A first attempt at modeling f0 changes as gestures timed to supralaryngeal articulatory gestures, in the sense of Articulatory Phonology (AP, Browman & Goldstein, 1989, 1992, 1986; Goldstein & Fowler, 2003), was attempted by Gao (2008) for Mandarin tones. In a gestural model, like the one proposed by Gao (2008), tones are conceptualized as gestures, that is, dynamical systems that specify time-varying f0 target values for an f0 tract variable (Burroni, 2023a; Gao, 2008; McGowan & Saltzman, 1995). Usually, two discrete levels (High and Low) are assumed, as in AM, however a third Mid level, perhaps representing a neutral value, has also been proposed in other analyses of tones in AP (Burroni, 2023a).

Despite the use of similar High and Low primitives, we must note that there exist similarities, as well as profound differences between AM and AP treatments of tone. In both AM and AP, tones are realized in a discretely separate "tier" or tract variable, the laryngeal tier in AM and an f0 tract variable in in the Task/Dynamic model of AP (Burroni, 2023a; Gao, 2008; McGowan & Saltzman, 1995). As a matter of fact, the very idea of separate tiers or tract variables, including to our knowledge the first notion of a gestural score (called "orchestral score") was first proposed in AM (Goldsmith, 1976) and the idea of introducing non-linear phonological primitives that isomorphically map to articulation was fully developed in AP by building on previous work, including AM (Browman & Goldstein, 1986).

Additionally, both AM and AP consider f0 the linguistically relevant "end effector" or task controlled by speakers. In other words, under both frameworks, speakers aim for particular f0 values, however, in AP, changes in an f0 tract variable are further realized by tone driving changes at the model articulator level, for instance, changes in glottal aperture, thyroid cartilage rotation, larynx height, laryngeal tension *etc*. (Burroni, 2023a; Gao, 2008; McGowan & Saltzman, 1995).

However, the idea of separation between laryngeal and supralaryngeal system the notion of an acoustic task (or goal) is where the similarities end. Profound differences exist between the tonal targets of AM and the tone gestures of AP.

First, in AM accounts, tones are conceptualized as pitch targets aligned to the segmental streaming that are phonetically realized via interpolating functions (Pierrehumbert, 1980). In AP, as noted, tone gestures are dynamical systems that act on the vocal tract forcing an f0 tract variable to time-varying f0 values over period of times. The evolution of f0 is, thus, determined by an f0 target as well as by other parameters (like stiffness) that are incorporated in the Task Dynamic model and potentially by blending with other (laryngeal) gestures (Burroni & Kirby, 2023, under review; Saltzman & Munhall, 1989).

A second crucial difference also exists. In AM, tones are instantaneous values aligned to **acoustic events** and used to approximate f0 using interpolation. In contrast, the tone gestures of AP are forces that shape the vocal tract over periods of time and that come with (lexical) relative timing specifications to other **articulatory events**, also conceptualized as forces that shape the vocal tract. Thus, the two theories crucially differ in what events are used by speakers to control the relative timing of tone production.

The differences between the two theories become especially pronounced when looking at cases like contour tone timing. In AM, contour tones are formed by two pitch targets (e.g., H and L for a Falling tone) that **need** to align independently to acoustic events in the segmental string. In AP, on the other hand, a contour tone is a period where forces drive the f0 tract variable to target states, like High and Low. These forces have lawful timing specification to other articulatory gestures produced at other tract variables, for instance with consonantal and vocalic gesture initiation, as well as an internal timing specification. Thus, for a Falling contour tone like HL the two gestures **need not** to be independently aligned, but the entire contour may come with a single timing specification to consonantal and vocalic (Burroni, 2023a; Gao, 2008) gesture, and a tract-variable internal anti-phase specification whereby the initiation of the Low gesture happens roughly around the completion of the High gesture. Much like the release of a consonant happens after the completion of a closure (Browman, 1994; Burroni, 2022; Nam, 2007b, 2007a), but the two cannot be independently timed.

Two findings of Gao (2008) in particular support the notion that the timing of tones is controlled in a gestural, articulatory sense, rather than being aligned to acoustic syllable boundaries.

First, Gao (2008) reported that Mandarin speakers do not initiate the consonantal onset and vocalic gesture of a syllable synchronously, but rather they initiate the vowel at the midpoint of the consonantal and tonal onsets, ~50 ms after the consonant (cf. also Yi, 2017). This puzzling relative timing pattern is reminiscent of the ones observed in English consonant clusters, where the vowel in a word-initial consonant cluster starts at the midpoint between the onsets of all the consonants in the cluster (Browman & Goldstein, 1988; Marin & Pouplier, 2010; Tilsen et al., 2012). In view of the observed C-V timing patterns in Mandarin, Gao (2008) proposed that Mandarin tones act like an onset consonant and are timed to the other supralaryngeal articulatory gestures, just like supralaryngeal articulatory gestures are timed to each other. Specifically, both the consonantal and the tone gestures experience forces that drive them to be initiated synchronously with the vowel, a pattern called in-phase coordination; yet the consonantal and tonal gestures are also exert forces upon one other so as to cause them to be initiated sequentially, a pattern called anti-phase coordination – directly parallel to the timing of consonantal gestures in an onset cluster. Model simulations of these timing specifications in a system of coupled oscillators (Browman & Goldstein, 2000; Burroni, 2023a; Gao, 2008; Nam, 2007a), showed that the outcome of these timing specifications is a compromise, where the onset of the consonant and the onset of the tone gestures are displaced (quasi-)symmetrically before and after the onset of the vocalic gesture, just like the C-center effect in English consonant clusters. Thus, tones have effects similar to those of an onset consonant on vowel timing. The wider implication is that tones can be equated, in terms of their effect on articulation, with other articulatory gestures.

A second important finding of Gao, which supports the idea that tones are tightly integrated into the articulatory plan, is that the initiation of tones relative to consonantal and vocalic gestures is not conditioned by syllable structure, specifically by the presence of a coda, in Mandarin. Thus, the lag between the initiations of a consonant and tonal gesture is identical for words like [ma] and [man]. The context-invariance of this timing patterns stands in contrast to variability of segmental anchoring that is observed across different syllable structures (Prieto & Torreira, 2007). In view of these findings, some have suggested that quasi-stable segmental anchoring is the byproduct of stable underlying timing regimes among articulatory gestures (Gao, 2008; Ladd, 2006).

However, it must be noted, crucially Gao (2008) did not test whether timing to acoustic or articulatory events reflects a more "stable" behavior of speakers, a crucial prerequisite to assess the predictions of acoustic and articulatory accounts of tonal timing (as noted by Gao, 2010). We take up this issue in the present paper.

Ever since this seminal treatment, gestural models assuming articulatory timing of tones have proven very attractive and have been further developed for Mandarin (Shaw & Chen, 2019; Yi, 2017; Yi & Tilsen, 2014), Tibetan (Hu, 2016), Thai (Burroni, 2023a; Karlin & Tilsen, 2014) and pitch accent languages, such as Swedish (Svensson Lundmark et al., 2021) and Serbian (Karlin, 2022). Even though some aspects of Gao's original proposal, especially the tonal c-center, have been called into question (Burroni, 2023a; Geissler, 2021; Kramer et al., 2023; Svensson Lundmark et al., 2021), the basic insight of tones being tightly integrated into articulation and timed to articulatory events still merits consideration and a direct comparison of acoustic and articulatory timing is essential to assess the merits of this proposal.

It is also important to note that a similar articulatory view of tonal timing is not limited to Articulatory Phonology, but it is also expressed in recent versions of the PENTA model, where tones are co-produced with the articulatory movements of consonants and vowels. Altogether these form a synchronized syllable-sized articulatory unit (Xu, 2020; Xu et al., 2022; Xu & Liu, 2006).

Given this background, we can appreciate that both acoustic and articulatory timing of tonal events explain some important properties of tonal phonology and production, and that it may be important to understand how the two are related. We now turn to examining in more detail the potential modalities in which the timing of tonal and oral articulatory/segmental events might be controlled, in order to clarify why both hypotheses are worth entertaining and what their wider implications are.

1.3. Tonal timing modality: Speech production models

Different hypotheses concerning tonal timing modality are also related to different models of speech production.

At first sight, given that tones are produced by articulatory adjustments, one may be tempted to assume that any apparent evidence for control of the timing of acoustic events is merely an artifact of the circumstance that articulation is the cause of acoustic change. Analysis of acoustic events may also simply be a matter of convenience, because articulation in general is more difficult to study. Laryngeal articulation in particular is hard to image and requires invasive methodologies to do so (Hirose, 2010; Hoole et al., 1999). However, such interpretations are limited in scope.

There is nowadays a general consensus in the speech production literature regarding the role of acoustic information to guide speech learning and production via feedback (e.g., Houde & Nagarajan, 2011; Parrell et al., 2019; Perkell, 2012 among many others). More specifically, there are models of speech production, for example the DIVA model (Guenther & Vladusich, 2012), where the acquisition of speech is based on tuning articulator-to-sound mappings, known as target regions, on the basis of acoustic targets. Tones may be particularly good candidates for speech events whose timing is controlled via acoustic targets. Even in the articulation-focused framework of Articulatory Phonology (Burroni, 2023a; Gao, 2008) f0 has been adopted as a state variable, when it comes to modeling tone production. Moreover, speakers are highly sensitive to feedback perturbations of f0 and are known to update their feedforward system in altered auditory feedback experiments (e.g,. Jones & Munhall, 2002; cf. Tang, 2024 for a detailed review).

In view of these models of speech production and the importance of acoustic information in tone production in particular, it seems tempting to assume that speakers of a tone language learn to coordinate f0 contours underlying tone production and segmental acoustic boundaries, while their underlying articulatory movements may exhibit more flexibility in view of the many non-linearities underlying articulatory to acoustic mappings (Moisik & Gick, 2017; Titze, 2000 Chapter 8; Weerathunge et al., 2022). Additionally, the timing of articulatory movements may be driven mostly by the need for acoustic landmark coordination to obtain, with a certain flexibility concerning the relative timing of the articulatory movements themselves, as long as an appropriate acoustic output is obtained. This is indeed also the position of models of lexical access based on acoustic "landmarks"(Stevens, 2002), who hypothesize that speech production is based on coordinating acoustic landmarks, including segmental anchoring of tones and intonation (Turk & Shattuck-Hufnagel, 2020b, p. 61).

Articulatory timing for tones, on the other hand, would entail that the laryngeal maneuvers regulating laryngeal tension and underlying f0 control have specific coordination regimes with the production of other articulatory gestures. These coordination regimes would be directly the goal of the speech production plan, rather than serving solely as a medium for an acoustic output to obtain. In terms of motor control, speakers of a tone language would thus have feedforward plans for the supralaryngeal articulation of sounds that incorporate timing regimes of such commands to the laryngeal commands underlying the production of lexical tones. In other words, in articulatory timing models, the timing regimes for tones are part of a single feedforward plan comprising laryngeal and supralaryngeal articulation, and, crucially, these plans are not yoked purely to produce an acoustic output; they have their own regularities and form a motoric routine that is a target for the speech production system. The articulatory gestures and their timing are the speech production goal in this case. This is essentially the position of gestural frameworks, like AP and the Task Dynamic (TD) model of speech production (Browman & Goldstein, 1992; Saltzman & Munhall, 1989), where the lexical representation specifies timing regimes of tones and consonantal and vocalic articulation (Burroni, 2023a; Gao, 2008). Note that in gestural frameworks like AP, tonal timing regimes are still inferred based on f0 contours, as f0 is the tract variable responsible for the linguistically meaningful goal of tone production. This abstraction of f0 over laryngeal articulation is similar to the one operating in inferring the production of a bilabial closure gesture, like /p/, from global lip closure rather than by individual movements of lower lip, upper lip, and jaw. Crucially, just like for other articulatory gestures, the f0 events are timed to the onset of other articulatory gestures, not to acoustic boundaries; additionally, f0 should drive adjustments, e.g., in laryngeal rotation angle and potentially other features at the model articulator level that we should be able to observe together with acoustic changes in f0 given appropriate imaging techniques (Burroni, 2023a; Gao, 2008; McGowan & Saltzman, 1995).

Given the background just outlined, it becomes clear that tonal timing is not simply a matter of practical implementation or of different phonological theories. Rather, tonal timing is a core issue that can be used to probe the very nature of tones as speech goals and how they are integrated into the more general process of speech production. Whether those goals are articulatory or acoustic in nature is a fundamental question that has yet to be resolved. Tonal timing also has broader ramifications for models of speech production, as it entails complex timing regimes that can shed light on how speakers produce speech by coupling the supralaryngeal and laryngeal systems.

1.4. Tonal timing landmarks: onsets, targets, and beyond

In addition to the modality of events, a comprehensive account of timing must also determine what sorts of events are timed relative to each other. A complex event that unfolds in time, such as a movement or speech segment, can normally be construed to have a beginning and end. Articulatory gestures are understood to have an onset – the time at which the gesture becomes active – and in most circumstances they achieve targets – the vocal tract reaches a particular state – at some time. Similarly, speech segments or phones are commonly assumed to have a start and an end (or onset and offset), which are the “boundaries” of the segmentation of the acoustic signal. For simplicity, we will use the term landmark to refer generally to any of these simple events that delimit the temporal intervals of gestures or segments.

A general issue in any type of coordinated speech movement is whether and how these landmarks are coordinated with each other (Gafos, 2002; Karlin, 2022; Nam, 2007a; Nam et al., 2009; Tilsen, 2022; Turk & Shattuck-Hufnagel, 2020a). As noted in the literature (Lee & Mok, 2021; Xu & Liu, 2006), tones, in the form of f0 contours representing the end effector of laryngeal adjustments, are empirically assumed to be synchronized to the acoustic boundaries of a syllable or word. Both their onset and their target/offset have been held to coincide with the beginning/end of the segments of a host syllable/word, or sometimes a mora. This idea has been captured differently in different models of f0 control. Yet, it must be noted that the notion that tone landmarks coincide with segmentally defined onsets/offsets is not very consistent with some empirical patterns such as perseverative coarticulation of tones across syllables, peak delay relative to segmental boundaries, and other phenomena (Xu, 2020; Xu & Liu, 2006).

To remedy these issues, some models allow for detailed specification of when onsets and targets/offsets occur relative to segmental boundaries, along with specification of rates of target achievement over an articulatory interval. For instance, in the PENTA model of f0 control, tonal targets are approximated throughout the boundaries of an articulatorily-defined syllable using an approximation rate (Xu, 2004; Xu et al., 2022; Xu & Liu, 2006). On the other hand, it has been noted that specifying timing to both syllable boundaries together with a slope or rate of approximation introduces overparameterization, as the slope of an f0 contour is mostly predictable from onset and target f0 values specified at syllable boundaries. This overparameterization creates a conflict that potentially needs to be solved by speakers (Flemming & Cho, 2017).

Other models propose more parsimonious representations that rely on onset and/or target coordination only, without approximation rates. Under the SAH, both tonal onsets and targets specifying f0 values are considered to be aligned to the acoustic segmental string (Flemming & Cho, 2017; Ladd, 2004), however targets are often attributed a privileged status and target alignment is studied much more widely than onset alignment (e.g., Prieto & Torreira, 2007).

In the standard model of AP and its coupled oscillator extension, which aims to account for control of intergestural timing (Goldstein et al., 2009; Nam, 2007a; Nam et al., 2009; Nam & Saltzman, 2003), only the onsets of articulatory gestures are coordinated with each other and movements evolve according to intrinsic-timing specifications that regulate the time course of gestural activation. These specifications include a stiffness parameter which controls the time to target (Saltzman & Munhall, 1989). When this model is applied to tone, it entails that tones are only specified to start relative to other gestures, but their target/end is not directly controlled; at least unless an additional mechanism suppressing groups of gestures is incorporated, as in Selection-Coordination (S-C) theory (Tilsen, 2016, 2018).

However, extensions to AP have also proposed richer landmark coordination topologies that are not onset-to-onset, for instance, target-to-release coordination (Gafos, 2002). A similar idea where tonal targets and vowel targets are coordinated with each other has recently been applied in an AP analysis of Serbian tones (Karlin, 2022). Such an idea that tones may be coordinated by specifying a target relative to another vocalic (or consonantal) target is also fully compatible with the recently proposed Timing-Extrinsic three-component model (XT/3C) (Turk & Shattuck-Hufnagel, 2020a). XT/3C proposes that the coordination of acoustic/articulatory events is controlled to achieve targets synchronously (Elie et al., 2023; Turk & Shattuck-Hufnagel, 2020a, 2020b) in line with Tau theory (Lee, 2011), a motor control approach originally developed for movements that have the goal of avoiding collisions with hard obstacles such as braking in car-driving or landing on a perch in animal flying movements. In the XT/3C model, speakers are assumed to control how their effectors close a gap between their initial state and target given a current closing rate. If the reasoning is applied to tones and articulatory gestures, these should reach their targets together. From the point of view of this framework, the onsets of synchronized movements are not as crucial as target achievement. In other words, onsets don't matter too much as long as they don't happen too late (Lee, 2011).

One empirical pattern to consider in relation to the question of onset vs. target control is that, while speech events invariably begin, it is not a given that they reach an underlying target. The fact that speech sounds are commonly reduced in rapid, spontaneous speech suggests that targets are not always achieved. This is commonly referred to as "target undershoot", and it is a likely articulatory basis for many common sound changes. The fact that speakers are willing to produce speech in which an underlying target is not achieved is an important challenge for models that privilege targets over onsets, like Tau theory where targets are both context-dependent and always achieved (Elie et al., 2023).

It is important to emphasize here that in all cases, models of timing make predictions about quantities that are not directly observable, in the sense that there is no ground truth procedure for locating them in observed signals. This holds in both articulatory and acoustic domains. Rather, the onsets and targets of gestures, along with the boundaries of acoustic segments, are theoretical constructs. There exists no theory-independent or assumption-free way of identifying them in articulatory or acoustics signals. Nonetheless, to make any theoretical progress, we must adopt heuristics that allow us to infer when the theoretical events occur (cf. Mücke et al., 2020 for an in-depth discussion of this point). The onset of a tonal event can be operationalized in various ways, such as the occurrence of a spike in neural populations that control laryngeal posture (Lu et al., 2023), or a peak in electromyographic signals representing collective action of muscle fibers (Erickson, 2011), or as inflection points in f0 contours obtained from autocorrelation or cross-correlation-based pitch estimation (Arvaniti et al., 1998; Flemming & Cho, 2017; Gao, 2008; Xu & Liu, 2006). Crucially, despite the fact that our measurements of the theoretical events are necessarily indirect, there are good reasons to assume that they are not wildly off. Nonetheless, it is important to bear in mind that any particular choice of estimation method may have consequences for the inferences we draw about theoretical events.

Another reason that caution is warranted in the study of timing has to do with the fact that the signals we observe can represent the simultaneous influences of multiple motor actions. This is particularly the case in studying timing of oral articulation, where there are biomechanical interactions between the jaw and tongue, and between the jaw and lips. The overlap of motor commands that govern the movements of articulators can result in target undershoot, and this can in turn lead to biases in estimates of when the theoretical events occur. Because of this, we should be careful in rejecting theories based on mismatches between phonetic measurements and theoretical predictions (Mücke et al., 2020). This is because, beyond the possibility of different ways to quantify events, the measurements themselves are also influenced by a variety of factors such as competing demands on articulators (Mücke et al., 2020) and intrinsic variability (Gafos et al., 2014; Shaw et al., 2011; Shaw & Gafos, 2015). Thus, it is always important to remember that the inferences we draw based on empirical measurements need to be taken with caution before they are held to invalidate a theory. In this paper, we use empirically estimated landmarks out of practical necessity, not because we attach theoretical value to the landmarks. This has been the practice of speech timing researchers since the field came into being, and we believe that, especially when segmental and tonal contexts are experimentally controlled, there are sufficient correspondences between empirical measures and theoretical events to draw inferences about the latter from the former.

In conclusion, the point that we wish to emphasize is that it is important to determine what types of landmarks speakers can and do control, even if our estimates of these are indirect. Specifically, is it the onsets of gestures/segments whose timing is controlled, or the targets/offsets of gestures and segments, or both? This landmark issue has important consequences when evaluating whether speakers rely preferentially on an acoustic or articulatory modality to accomplish tonal timing, as both modalities need are assumed to make use of landmarks to accomplish timing. In turn, the issue of modality bears on the core of how much tones are integrated into the articulation plan, as we have outlined above. To evaluate this issue of tonal timing modality, however, we also need to evaluate in parallel the issue of how that modality is instantiated using different landmarks.

1.5. Research questions, hypotheses, and predictions

Since laryngeal and supralaryngeal articulation are coupled yet largely independent systems, speakers of tone languages are faced with the task of adopting timing regimes that allow for precise f0 control over relatively short timespans of speech. Assessing the reliance of speakers on different modalities is a question that, however, cannot be investigated without preliminarily establishing what movement or event landmarks may underlie articulatory or acoustic timing of tones. Thus, we first illustrate possible hypotheses and predictions regarding the issue of tonal timing landmarks and subsequently move to the question of tonal timing modality.

The issue of **tonal timing landmark types** can be summarized in the following question: what is the basic coordination strategy adopted by speaker in controlling the timing of lexical tone production relative to segmental production? That is, what types of landmarks do speakers control in the timing of laryngeal to supralaryngeal events?

Some models of tonal timing, for example, standard formulations of AP, hypothesize that speakers control the timing of tones relative to supralaryngeal articulation by displaying more stable **onset-to-onset** coordination, just like they do for supralaryngeal articulatory events. The SAH hypothesis also considers onset-to-onset acoustic coordination but together with target-to-target coordination. On the other hand, some extensions of Articulatory Phonology, the SAH, and the XT-3C model hold that speakers should (also) display **target-to-target** coordination, so that tonal and articulatory/acoustic targets are reached at the same time. Finally, much of the literature on lexical tones, which is based on acoustic research, assumes a full synchronization and a comparable role for **both onsets and targets**, as both are involved in timing tones to syllable- or word-size units.

The predictions associated with the types of landmarks that are temporally coordinated involve variability. The temporal interval (lag) between any two events is expected to be less variable if speakers control the relative timing of those events, and more variable if they do not (Browman & Goldstein, 1988; Burroni, 2023a; Gafos et al., 2014; Honorof & Browman, 1995; Mücke et al., 2009, 2020; Shaw et al., 2009, 2011; Tilsen et al., 2012; Tilsen, 2017). The predictions based on variability are illustrated below, Fig. 1, which represents patterns of variability in onset-to-onset and target-to-target lags, associated with (a) onset-to-onset coordination, (b) target-to-target coordination, and (c) coordination of both onsets and targets (full synchronization).

The issue of **tonal timing modality** can be summarized in the following question: are the events whose timing is controlled defined an acoustic or articulatory domain?

Whereas AP, extensions of AP, and XT-3C conceptualize the relevant events in the articulatory domain, the SAH does so in the acoustic domain, as do theories of lexical access that are based on acoustic information. As is the case for hypotheses of landmark type, the hypotheses associated with landmark modality make predictions about patterns of variability: if the relevant events are articulatory in nature, then lags between articulatory measures should exhibit lower variability than ones between acoustic measures. Vice versa if the relevant events are acoustic. This approach to investigation of modality has been used in a number of previous studies of pitch accent timing (D'Imperio et al., 2007; Mücke et al., 2009; Niemann, 2016; Niemann et al., 2011; Prieto et al., 2007).

There are some important caveats to mention here, with respect to testing the above predictions. First, given the invasive nature of direct recording of laryngeal activity (Sections 1.1, 1.3), we can only practically obtain articulatory event measures from oral articulatory signals, not from laryngeal articulation. For this reason, we can only partly test the hypotheses in the sense that we can assess the variability of an f0 events measured acoustically in relation to segmental events measured acoustically vs. articulatorily. This simplification has been commonly adopted in previous work and is still capable of providing evidence for or against the relevance of a particular modality in temporal control. We also note that, theoretically, AP considers f0 the relevant state variable and articulation only specified at the model articulator level. F0 is also the relevant variable for acoustic theories. Thus, studying f0 is theoretically consistent.

Second, it is also possible that event time estimates from different modalities may be biased to have different levels of variability. For instance, articulatory landmarking may be more prone to measurement errors than acoustic segmentation, or vice versa. Since there is no method for obtaining ground truth measures of theoretical events (Section 1.4), this problem is unavoidable. As a result, it is not possible to reach definitive conclusions. This situation is not unique to our study, but rather, is universal.

Third, variability patterns alone may be misleading when there are additional influences on production that systematically influence timing. For instance, variation in speech rate can be expected to increase variability in event timing. An external source of variation of this sort may obscure differences in variability that would be more pronounced in the absence of the external influence. For this reason, and also to strengthen the inferences that can be drawn from estimates of event times, we assess two additional types of predictions beyond variability. We refer to these as stability and informativity.

By "stability", we mean the extent to which the relative timing intervals between events remain consistent despite experimental manipulations or perturbations; such as variations in speaking rate, syllable structure, and tonal context differences. The concept of stability is well-known from work on nonspeech

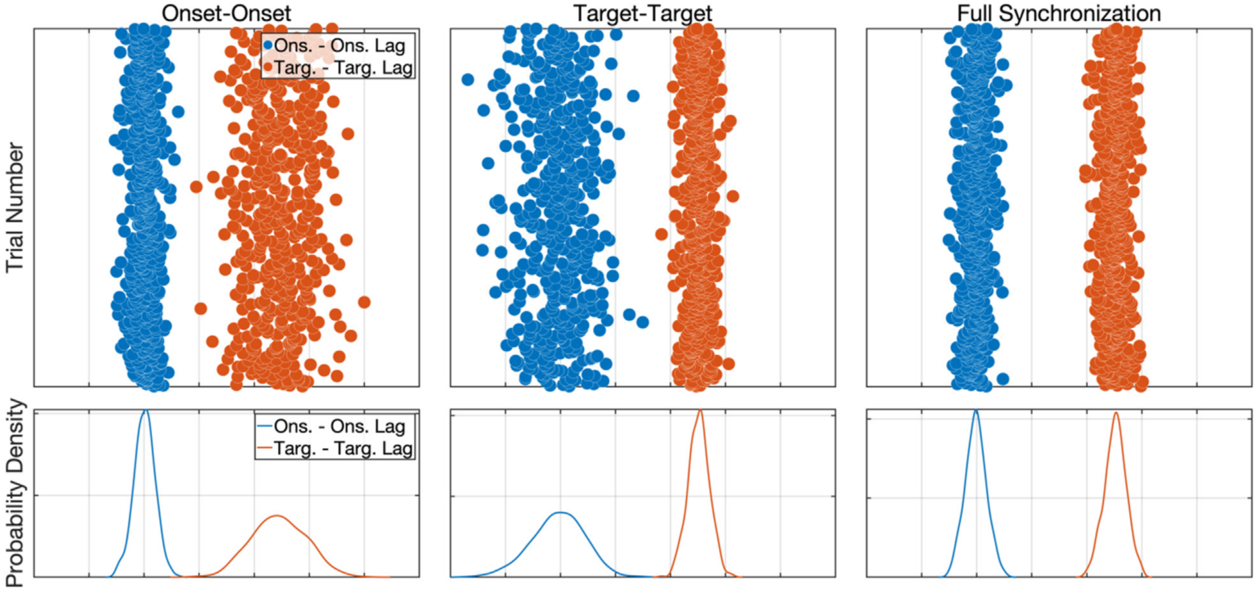

Fig. 1. Illustration of predicted variability under onset-to-onset (left), target-to-target (center), and full synchronization (right). Each pair of dots lying on the same y-value in the top panel indicates an observed onset-to-onset lag (blue) and target-to-target lag (orange) for a tonal event and an acoustic/articulatory event to which the tone is assumed to be timed. The two types of lags are offset from each other purely for illustration purposes. The bottom panels portray kernel density estimates of lag distribution, showcasing the difference in spread or variability between lags of landmarks that are assumed to be under speaker control. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

motor control where it is defined as the capacity to return to an equilibrium state after perturbations (Scholz & Schöner, 1999). The notion of stability we rely on in this paper aligns with the approach taken in the seminal work on the Segmental Anchoring Hypothesis (SAH) by Arvaniti et al. (1998). They showed that the relative timing of tonal targets and a segmental anchor remains stable despite changes in the duration of the relevant acoustic events due to speaking rate differences. For example, consider the timing relationship between a pitch peak (tonal target) and the onset of a syllable (segmental anchor). If the pitch peak consistently occurs at the same relative time during the syllable, regardless of whether the speaker is speaking fast or slow, this demonstrates relative timing stability. The relative invariance of the temporal interval indicates that the production system has the goal of maintaining specific timing patterns regardless of the external influences responsible for variation in syllable duration. Systematicity of this type is taken to indicate phonological control that should not be altered by changes in speaking rate, a paralinguistic variable. Indeed, using speaking rate as an experiment perturbation to test for control should be viewed as an important feature of our method, and it aligns with the widely held notion that phonological specifications should be insensitive to rate effects (*cf.* the thorough discussions in Bennett et al., 2023; Flemming & Cho, 2017).

The opposite pattern, where speaking rate has systematic effects on relative timing, entails a lack of stability. For example, the relative timing between a pitch peak and a syllable onset may increase or decrease with changes in speaking rate. This would be expected if the timing of the pitch peak is not coordinated with the syllable onset. The pitch peak might be timed to a later event or prevented from being delayed indefinitely due to other targets, such as a subsequent pitch fall. If such systematic effects of speaking rate are observed, we may conclude that the relative timing specifications between the peak and the syllable onset are either non-existent or weak. In other words, speaking rate alters the timing without the system attempting to return to its normal state or oppose the perturbation. When the system does not resist the change, both the pitch peak and the syllable onset are independently influenced by external factors like speaking rate. This results in their relative timing pattern exhibiting systematic effects, such as anticipation or delays. These predictions, where relative timing changes or remains stable as a function of speaking rate, are exemplified below in Fig. 2.

Note that stability predictions can also be used to compare different landmarks within the same modality, and thus, can be employed to address the question of tonal timing landmarks within the same modality.

Finally, “informativity” refers to the extent to which the timing of on event provides information about the timing of another. Informativity as we operationalize it, is an information-theoretic measure (mutual information) that is a generalization of correlation. Mutual information is the amount of information that is present in variables considered separately that is absent when they are considered together. Unlike linear correlation, mutual information captures non-linear relations between variables. It accomplishes this by calculating the difference between estimates of the entropy of a joint distribution of variables and the entropies of the variable distributions calculated independently. We use this measure of informativity to test both event type and event modality predictions. Specifically, relying on articulatory/acoustic modalities for tonal timing predicts that knowing when an articulatory or acoustic event occurs can help predict the initiation of tonal events. This is because the two are coordinated, creating a dependency between them.

As an example, consider again the timing relationship between a pitch peak (tonal event) and the onset of a syllable (articulatory/acoustic event). If the onset of the syllable consistently indicates when the pitch peak will occur, this demonstrates high informativity. In other words, knowing the timing of the syllable onset allows us to predict the timing of the pitch peak, indicating strong coordination and dependency between these events. To our knowledge, this notion of informativity has not been extensively explored in the literature on the timing of tones or other articulatory/acoustic events. However, it has been applied to spatial dependencies in production as a diagnostic for the degree of coarticulation (Chen et al., 2015; Iskarous et al., 2013). In the context of coarticulation, spatial informativity refers to how the spatial configuration of one articulatory event provides information about the configuration of another. For instance, the positioning of the tongue during the production of a consonant can predict its positioning during a following vowel if there is a strong spatial dependency, and vice versa. Consider, for instance, a fronted velar constriction in the environment preceding a front vowel as an example of the spatial dependency just described. The similarity between temporal and spatial dependencies lies in the concept of shared information, quantified via mutual information. In temporal dependencies, the timing of one event offers predictive information about the timing of another. Similarly, in spatial dependencies, the positioning of one articulatory event informs the positioning of another. Both types of dependencies reflect a coordinated and interdependent production system, whether in terms of time or space. Understanding these dependencies aids in diagnosing the degree of coarticulation and the overall timing organization of speech production systems.

We wish to stress that, compared to variability and stability, informativity is a more complex form of evidence for temporal control of events because it both incorporates external factors (implicitly via estimated probability distributions) and captures non-linearities in correlations between variables. We assess informativity predictions by comparing estimates of mutual information between distributions of timings of f0 events and acoustic/articulatory events, following previous work on spatial dependencies in coarticulation. The higher the mutual information between variables, the more likely it is that they are temporally coordinated. Mutual information for two completely independent and two highly dependent variables is illustrated in Fig. 3 and its mathematical formulation is presented in our methods (Section 2.3).

An important caveat regarding coordinated landmarks is that, in the case of contour tones, like complex falling-rising or rising-falling tones, potentially, the onset and target of individual raising and falling components could be analyzed separately. In practice, however, estimating the onset and target of the second component of tonal contours is difficult in view on contextual effects, as we detail in Section 2.2. Nonetheless, given that the onset of the second movement of a contour tone

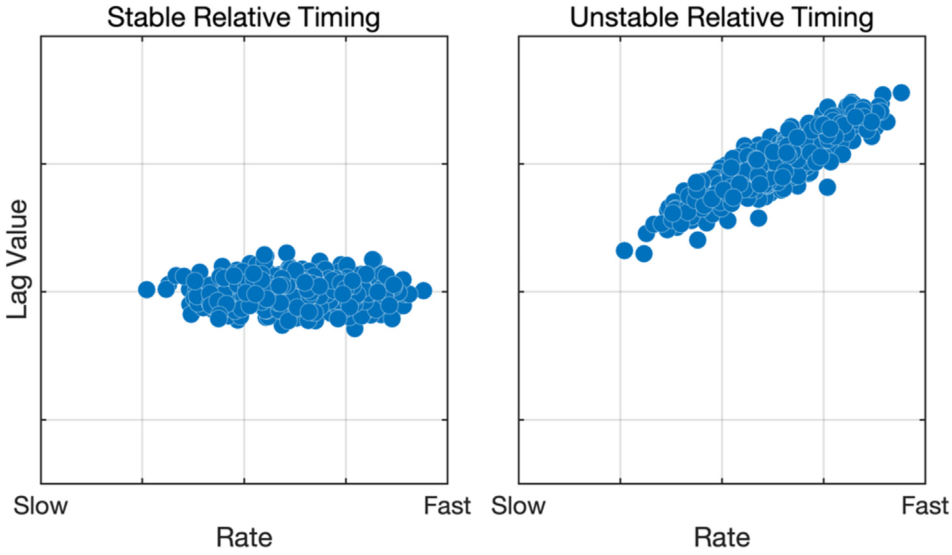

Fig. 2. Left: More stable relative timing where the lag is not systematically influenced by an external perturbation, such as a change in speech rate. Each dot represents an observation. Right: Less stable relative timing where events are independently influenced by rate, yielding systematic effects on the relative timing of two events.

is temporally close to the target of the first, a detailed analyses of the onsets and targets of the initial movement of a contour tone is likely to offer worthwhile information for the study of tonal timing.

To sum up, our predictions with respect to both tonal timing landmarks and modality are summarized in Table 1 below.

1.6. Thai: The language under investigation

Thai represents a good case study for investigation of tonal timing in view of its tonal inventory and established analyses in both AP (Burroni, 2023a, 2023b; Gao, 2008; Karlin, 2018; Karlin & Tilsen, 2014) and Autosegmental and Metrical Phonology / SAH (Gandour, 1974; Morén & Zsiga, 2006; Zsiga & Nitisaroj, 2007). Thai is almost universally described as a five-tone system (*cf.* Abramson, 1962 for alternative descriptions), Fig. 4.

Among the five tones, the Falling (F) and Rising (R) tones exhibit a complex trajectory during which f0 rises and then falls or falls and then rises, respectively. In this respect, the f0 contours of the F and R tones resemble an articulatory trajectory in that they invariably approach a High or a Low f0 target, respectively, before moving away from it. This happens regardless of the following tonal context (Burroni, 2022; Burroni & Kirby, 2023). The shape of the Falling and Rising tones makes them amenable to tracking their f0 onsets and targets in a way that is comparable to landmarking adopted for articulatory trajectories. This is a desirable property, given that some previous work on tones have been complicated by the fact that identical segmentation algorithms could not be applied to both the articulatory trajectories and the f0 trajectories (Gao, 2008).

The f0-based estimation of onset/target events is not as robust for other tones, like the Mid (M) and the Low (L) because onsets and final target will be heavily influenced by the f0 direction of the preceding tone, following tone, and speech rate. For instance, in M–L sequences, it would be impossible to determine the target of the first M tone and the onset of the L tone algorithmically as the f0 trajectory does not change direction and looks like one long f0 fall. Similarly, the High tone (H), albeit looking like a complex contour shape, would be harder to study compared to the F and R tones because it exhibits contextual variation due to tonal contexts, with the final fall being absent in certain tonal combinations (Burroni & Kirby, 2023). Additionally, the H tone is also subject to sociolinguistic (Teeranon & Rungrojsuwan, 2009) or intonation-driven shape variation (Luksaneeyanawin, 1983). For this reason, the F and the R tones represent ideal case studies to study tonal timing, given that both their onsets and initial targets can be more easily identified, like the closure phase of a consonant in articulatory data. Additionally, the f0 raising of the F tone and the f0 lowering of the R tone can be interpreted as movements in their own right, given that they correspond to peaks in cricothyroid and strap muscles, e.g., thyrohyoid activity, respectively (Erickson, 1976, 2011, 2013).

Additionally, Thai tones, unlike Mandarin tones, do not have obligatory – or even common – voice quality cues that would need to be tracked together with f0, as these are also an important expression of tonal contrasts in some languages (Brunelle et al., 2010; Chen & Gussenhoven, 2015; Kuang, 2017). In Thai, the only voice quality cue we are aware of is a moderate amount of creak at the lowest value of the Low tone; this cue is not, however, consistent across speakers (Thepboriruk, 2009).

Thai tones are an interesting case study not only because of their contour shapes, but also because they have been analyzed both in Autosegmental-Metrical phonology(Gandour, 1976; Morén & Zsiga, 2006; Zsiga & Nitisaroj, 2007) and in gestural models of phonology (Burroni, 2023a; Gao, 2008; Karlin, 2018).

In Autosegmental-Metrical accounts, all tones are hypothesized to be acoustically anchored to either the syllable or moras. They are represented as floating High and Low tonal primitives associated with a tone-bearing unit (TBU). The M tone is treated as an underspecified tonal target or as an M primitive by scholars who assume such a representation. The L and H tones are represented by a single tonal primitive, H and L, respectively. The contour tones are treated as a

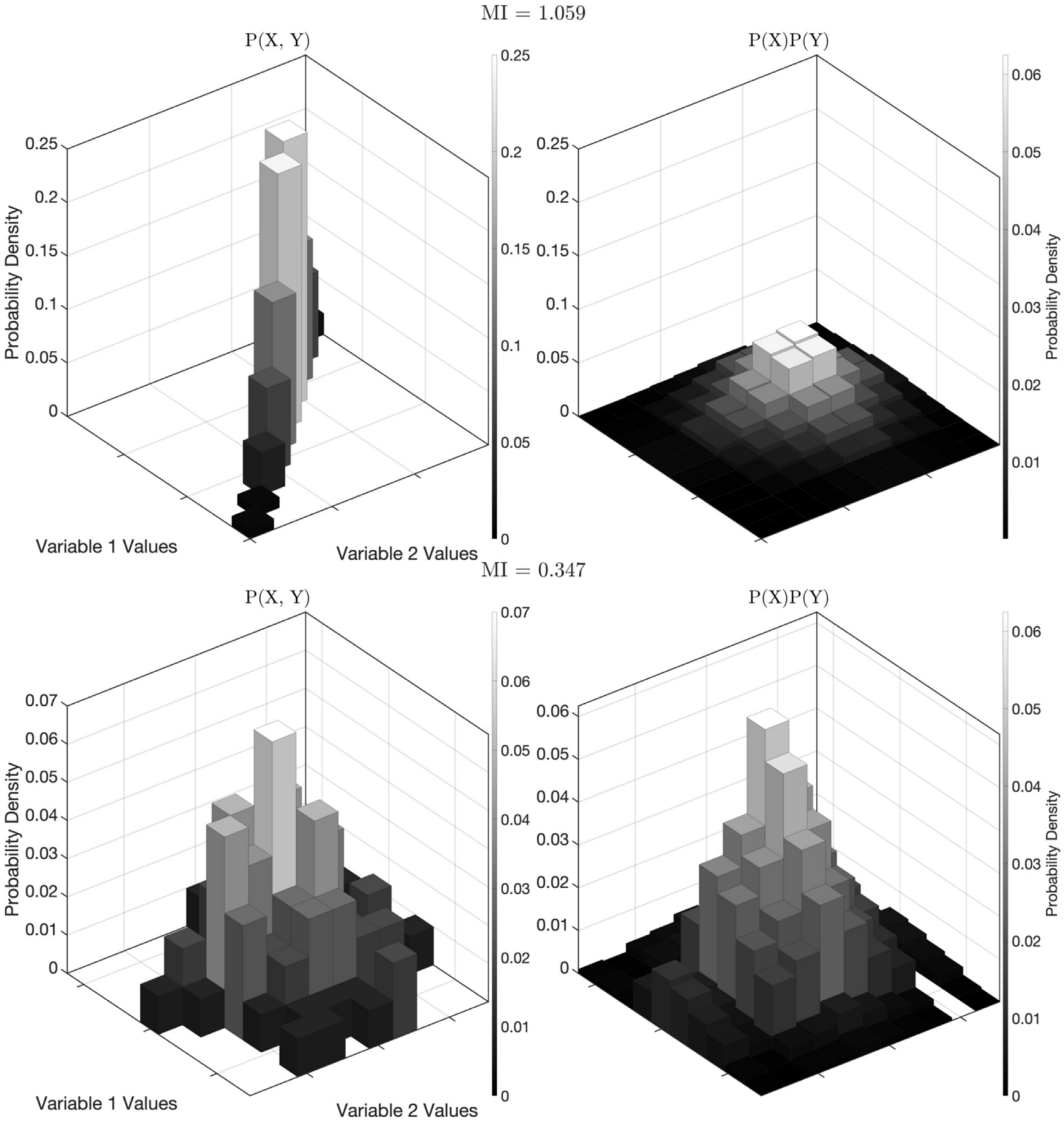

Fig. 3. Top: low MI indicating that two random variables X and Y have an observed joint distribution P(X,Y) that does not look very different from a joint distribution estimated under an independence assumption P(X)P(Y). Bottom: high MI indicating that two random variables X and Y have an observed joint distribution P(X,Y) that strongly deviated from a joint distribution estimated under an independence assumption P(X)P(Y).

sequence of HL (F) and LH (R) primitive tones. These systems propose onset-to-onset timing and/or target-timing to the right of edge of a TBU, either the vowel/mora or the syllable in different accounts (Gandour, 1976; Karlin, 2018; Morén & Zsiga, 2006; Yip, 2002), Fig. 5.

On the basis of the Autosegmental Metrical Phonology accounts above, it seems reasonable to hypothesize that the onsets and/or targets of the Thai tones may be aligned to particular acoustic landmarks, for example the syllable or vocalic onset, in line with the SAH model adopted in Autosegmental-Metrical Phonology and other theories. Crucially, Autosegmental-Metrical Phonology accounts also posit that the H portion of the H tone has its own timing regime relative to the segmental string, i.e., the two tonal primitives are hypothesized to be timed independently (Zsiga & Nitisaroj, 2007).

Turning to gestural models, a gestural account of Thai tones has been proposed by Gao (2008).

Gao proposes treating the Mid tone as a perfectly overlapping H and L f0 gestures; the Low tone as an L f0 gesture; the Falling tone as a sequence of H and L; the High tone as an H tone gesture; and the Rising tone as a sequence of an L and an H f0 gesture, Fig. 6 top row. Despite the merits of attempting to develop an alternative gestural organization of Thai tones,

Table 1
Predictions for landmarking based on different timing models (Top) and predictions for modality under the assumption of articulatory and acoustic modality for tonal timing (Bottom).

	Onset-to-Onset	**Target-to-Target**	**Full Synchronization**
Variability	Var. Ons. < Var. Targ.	Var. Targ. < Var. Ons.	Var. Ons. = Var. Targ.
	Articulatory Modality	**Acoustic Modality**	
Variability	Var. Art. < Var. Ac.	Var. Ac. < Var. Art.	
Stability	No effects on Art. Lag	No effects on Ac. Lag	
Informativity	MI Tone – Art. Ev > MI Tone – Ac. Ev	MI Tone – Ac. Ev > MI Tone – Art. Ev	

Gao's representation of the Thai tones is problematic. For example, the H tone, as hypothesized and synthesized by Gao, is a high-level tone, but this is not how the Thai H tone appears in recent acoustic investigations, as we have also remarked above. The representation proposed by Gao cannot account for the initial level f0 observed for the H tone, followed by an f0 rise and an (optional) slight fall. Other problems hold for the M tone as well, as it is not level either. For these reasons, a modified gestural account of Thai tones has been proposed, Fig. 6 bottom row.

The revised gestural model has also been computationally implemented to assess whether it can generate the Thai tonal contours. Fig. 7 illustrates the close fit between the revised tonal representation and Thai f0 contours, averaged across 20 speakers (cf. Burroni, 2023a for more details).

Regardless of the specific tonal gesture representation, in view of the gestural accounts presented above, it seems reasonable to hypothesize that the onset of the Thai tones may be aligned to specific articulatory landmarks, such as the vocalic gesture onset, in line with the predictions of the coupled oscillator model of AP.

In sum, both acoustic and articulatory timing, as well as different landmarks, have been proposed for tonal timing in Thai. The issue of which landmarks may be coordinated in tonal timing and whether tonal landmarks display more stable timing regimes to acoustic or articulatory events is a question that requires further investigation. Accordingly, in this paper, we offer a first step in this direction by examining which relative timings among events seem to be more stably employed by speakers within an acoustic or an articulatory modality. Additionally, we also compare the more stable landmarks across different modalities to draw inferences about Thai tones specifically and tonal timing in production more generally.

2. Methodology

2.1. Participants

Eight (3 male and 5 female) native speakers of (Central) Thai (age range 26–32, mean = 29.75, standard deviation = 2.18) participated in the experiment. All participants were graduate students at a North American university at the time of the recordings. They did not disclose any speech or hearing impairments. All speakers were screened for nativeness in Thai prior to data collection by a native speaker trained in phonetics.

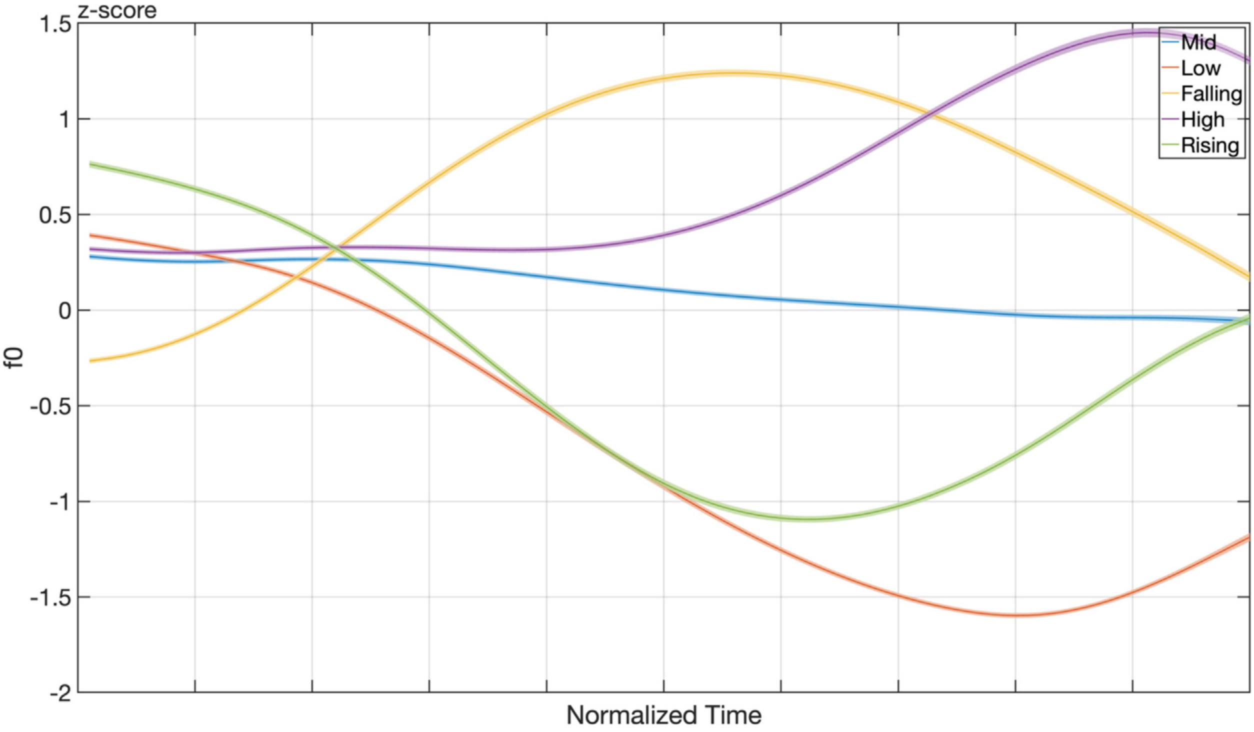

Fig. 4. mean f0 values of the five tones of Thai produced in connected speech over the syllables [miː]. Solid lines represent mean f0, and shaded areas represent two standard errors. Data are averaged across 20 speakers (Falling, Rising) or 24 speakers (Mid, Low, High). Data from Burroni & Kirby (2023).

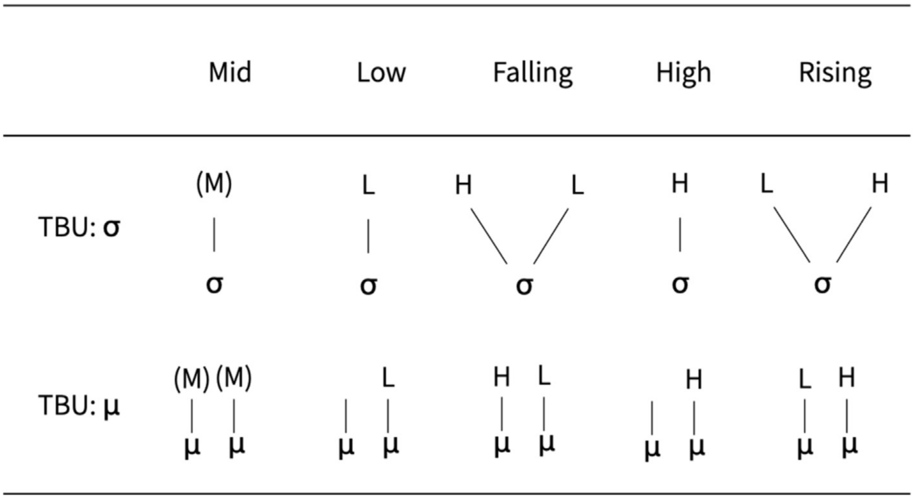

Fig. 5. Phonological Representation of Thai tones in autosegmental-metrical framework. Top row: syllable is the TBU, Bottom row: mora is the TBU (modified after (Morén & Zsiga, 2006).

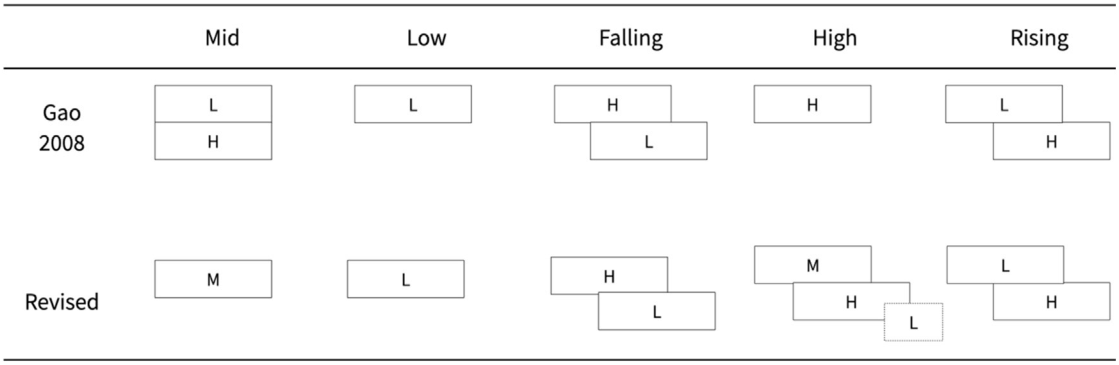

Fig. 6. Gestural representation of Thai tones. Top row: Gao's original proposal. Bottom row: revision proposed in more recent work Burroni (2023a).

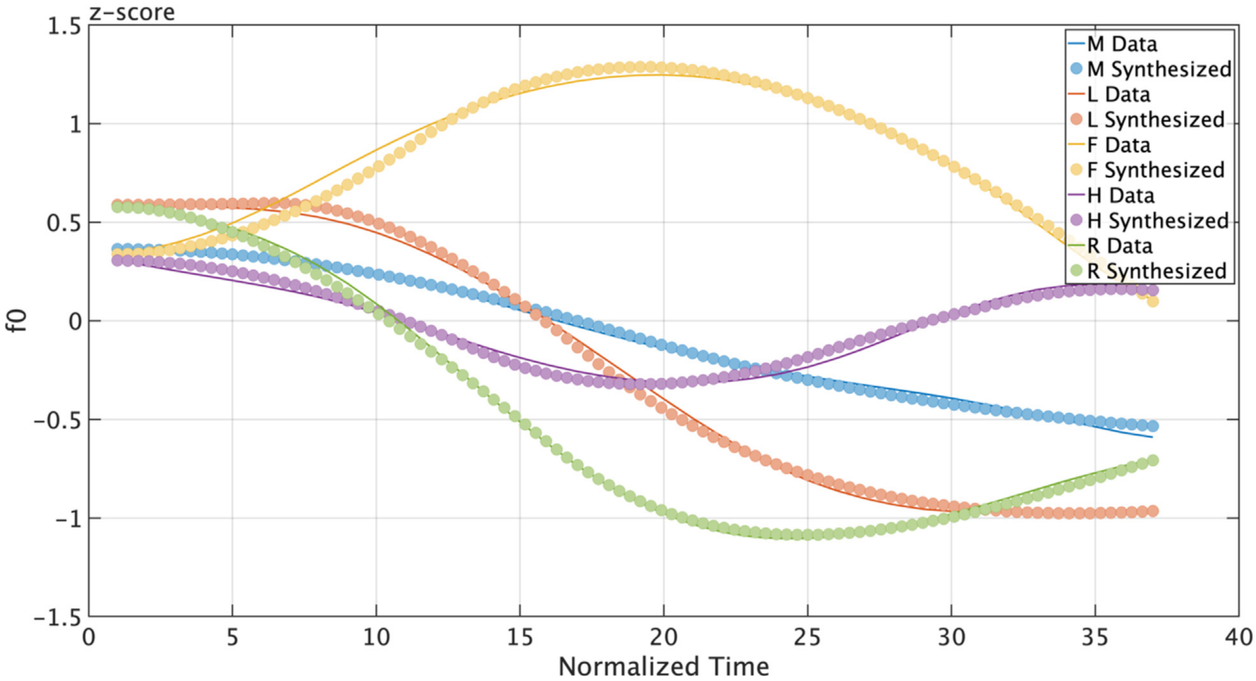

Fig. 7. Empirically observed f0 contours (solid lines) averaged across all speakers and fit based on the analyses and models presented in Burroni (2023a).

Although 8 participants may seem small, it is a significant improvement compared to previous articulatory work on Thai that relied on a single speaker (Karlin & Tilsen, 2014). Additionally, this study was conducted during the COVID-19 pandemic, which restricted our ability to access lab facilities and recruit more participants. Moreover, eight participants is twice the median number of four reported in EMA experiments appearing in work published until 2021 (calculations based on the data reported in Rebernik et al., 2021).

2.2. Experimental materials

Participants produced the Thai F and R tones followed by all five tones of the language (M, L, F, H, R) in combination with five different rate cues (very slow, slow, normal, fast, very fast).

The F and R tones were chosen, as previously pointed out (Section 1.6), for several reasons. First, their trajectory resembles the production of a consonant, a fact that makes it possible to employ similar landmarking procedures from acoustic and articulatory trajectories and to obtain estimates for both tonal onset and initial target. Second, identifying both targets and onsets would not be possible for basically any of the other tonal combinations involving a preceding Mid tone, especially at faster speech rates. Consider, for example, an M−L combination, Fig. 8 left panel, where it is hard to distinguish the target of the initial M tone, presumably a value in the mid to low f0 range, and the onset of the following L tone. On the other hand, in the combinations we have chosen, such as M−R or M−F combinations, Fig. 8 mid and right panels, the onset and target of the second tone can be identified more easily.

Third, the initial portions of the F and R tones are not purely autosegmental H and L phonological targets. They have a clear motor control interpretation, that is, an increase in cricothyroid and strap muscle activity when raising or lowering f0 from the mid-range, respectively (Erickson, 1976, 2011, 2013). Such coding for f0 raising and lowering is reflected not only in muscular activity but also in the neural activity recorded in Mandarin tone production (Lu et al., 2023). These are consideration pointing to the existence of neural commands underlying such pitch lowering and raising activities distinct from the tonal category itself. On this basis, we think that the study of a specific part of a contour tone is justifiable in terms of a speech motor plan with a distinct target, that is encoded in neural populations producing a specific movement pattern.

We introduced rate variation and tonal context manipulation in the experimental paradigm to probe the stability, variability, and informativity of different tonal timing landmarks and modalities. Speech rate requires little comment because, as remarked above (Section 1.5), it is often used to probe tonal timing stability either via direct manipulation or in post-hoc analyses (e.g., Arvaniti et al., 1998; Flemming & Cho, 2017; Torres & Fletcher, 2020; Xu, 1998).

Beyond speech rate we also manipulated tonal context. The manipulation of tonal context was added because Thai tones are known to be sensitive to contextual effects, both anticipatory and preservative (Burroni, 2023a; Burroni & Kirby, 2023; Gandour et al., 1994; Potisuk et al., 1997) and because this particular manipulation has not been tested before, as compared, for instance, to syllable structure (Gao, 2008; Karlin, 2014; Karlin & Tilsen, 2014). In this paper, we focus on manipulating the following tonal context over the preceding tonal context for two reasons. The first consideration was merely a practical one. Keeping the preceding tonal context constant is vital to obtain precise estimates of tonal onsets. By keeping the preceding tonal context fixed to an M tone, we could ensure that we could reliably track the F and R tone onsets. If we had manipulated the preceding tonal context, the onset of target F/R tones would be biased by the preceding content resulting in less robust estimates of gestural onset.

For instance, tones with a low final target, L and F, would make it impossible to identify the onset of an R tone, Fig. 9 top panels. Similarly, tones with a high final target would make it impossible to identify the onset of an F tone, Fig. 9 bottom panels.

Given the difficulty in isolating tonal onsets once the preceding context is manipulated, we decided to manipulate the following tonal context, which has been shown in previous

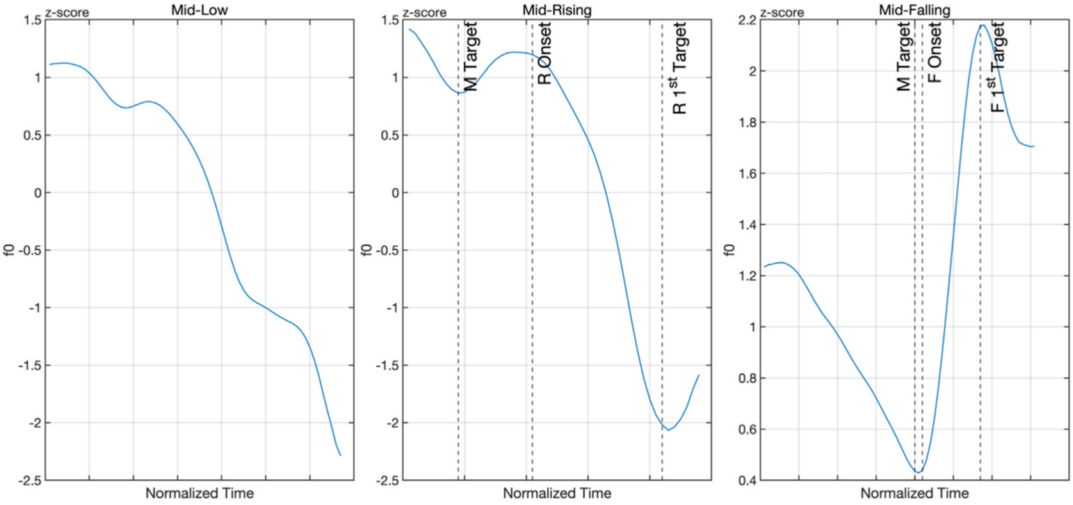

Fig. 8. Left: Example of a M−L f0 contour produced at relatively fast speaking rate. Mid: example of a M−R f0 contour produced at a relatively fast speaking rate. Right: example of a M−F f0 contour produced at a relatively fast speaking rate. Raw unsmoothed trajectories from single tokens obtained from data in Burroni & Kirby (2023).

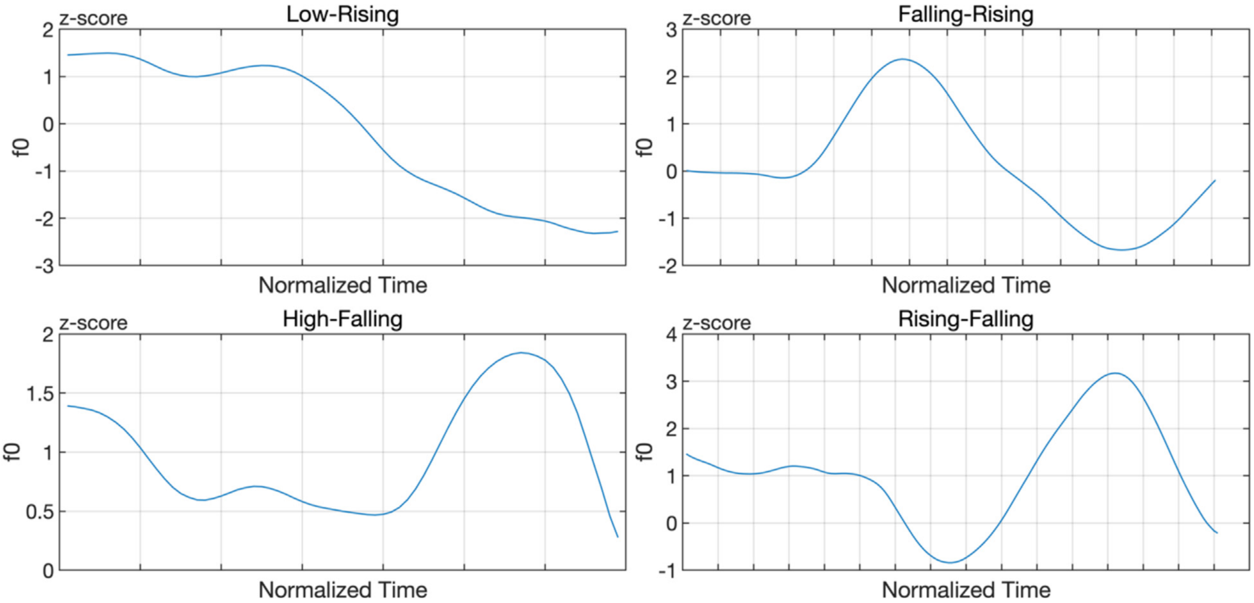

Fig. 9. Top: examples of tonal combinations where the onset of an R tone is impossible to distinguish from the final target of an L (top left) or an F (top right). Bottom: examples of tonal combinations where the onset of an F tone is impossible to distinguish from the final target of an H (bottom left) or an R (bottom right).). Raw unsmoothed trajectories from single tokens obtained from data in Burroni & Kirby (2023).

acoustic work to affect tonal targets and even regions close to the onset (Burroni, 2023a). An additional advantage of this choice is that we can still track the tonal target of the f0 raising for F tones and the target of the f0 lowering for R tones, because the context that determines the relevant tonal landmark—the rising portion of R tones (which are LH contours) and the falling portion of F tones (which are HL contours)—remains present regardless of the following tone.

A potential downside to our choice is that we may not be able to robustly estimate the target of the second components of these contour tones, due to the variation in the following tone of the final f0 lowering, i.e., the target of the falling portion of the F tone and the target of the rising portion of the R tone. This is not a problem that could have been solved simply by keeping the following tonal context fixed, since if we had kept the following constant as an M tone, we still would not be able to isolate the contributions of F/R and M, as both tones require a movement towards the mid of the f0 range, Fig. 10.

This is, of course, a well-known problem in speech production work, where a similar case is the observed kinematic trajectory for a consonant release. These trajectories cannot be attributed entirely to the consonant, as they are also influenced by the vocalic gesture following the consonant; thus, they are often considered a blend of the two (Nam, 2007b). A similar situation is likely to hold for the final tonal target cases just discussed.

For the reasons outlined above, we decided (i) to use an F/R tone set as the main target words, given that we can reliably isolate an onset and initial target and have clear theoretical and empirical importance attached to both events; (ii) to manipulate the following tonal contexts using the full set of Thai tones to introduce variation. Other manipulations would influence onset tracking (preceding context) or would still not have allowed us to track the end portion of tones (keeping the following context fixed), thus presenting no advantage over the manipulations we chose.

We refer to the first set of tones as tone 1 (T1 F/R) and the second set as tone 2 (T2 M, L, F, H, R). The disyllabic combinations were embedded in a carrier sentence with a fixed number of words and syllables; all carrier words are M−toned. Participants were told that the F/R M/L/F/H/R combinations represent a nonce noun-noun compound. Each unique target sentence is an M F/R_M/_L/_F/_H/_R M M string that can roughly be translated as “I look at a [mîː/mǐː] (with fur? pattern) [māː/màː/mâː/máː/mǎː] on a star/ball/box,” in Thai orthography ดู (หมี้|หมี) (มา|หม่า|หม้า|ม้า|หมา) บน (ดาว|บอล|ลัง), in IPA [dūː (mîː|mǐː) (māː|màː|mâː|máː|mǎː) bōn (bɔn|dāːw|lāŋ)], Table 2.

2.3. Experimental procedure

During the experiment participants sat in front of a computer monitor. A custom MATLAB GUI was used to present the stimuli and collect synchronized acoustic and articulatory data. Audio was collected with a sampling frequency of 44.1 kHz and 16 bits per sample using a shotgun microphone positioned around 1.25 m away from the participant. Articulatory data were collected at a sampling frequency of 400 Hz using an NDI Wave electromagnetic articulometer (EMA). To collect articulatory data, sensors were adhered midsagittally on the lower and upper lip (LL, UL) on the vermilion border. One sensor was placed on the lower right incisor to capture jaw movement (JAW). Two sensors were placed on the tongue: tongue tip (TT) to measure movement, and tongue body (TB) approximately 6–7 cm posterior to the TT sensor. Reference sensors were positioned on the nasion and left and right mastoid processes.

In the training phase, participants saw all ten unique target disyllabic tonal combinations appearing in the carrier sentence. The training helped familiarize participants with the spelling of words, the task, and the intended tonal contours. Furthermore, each of the 10 unique stimuli was combined with 5 unique speech rate instructions: very slow, slow, normal, fast,

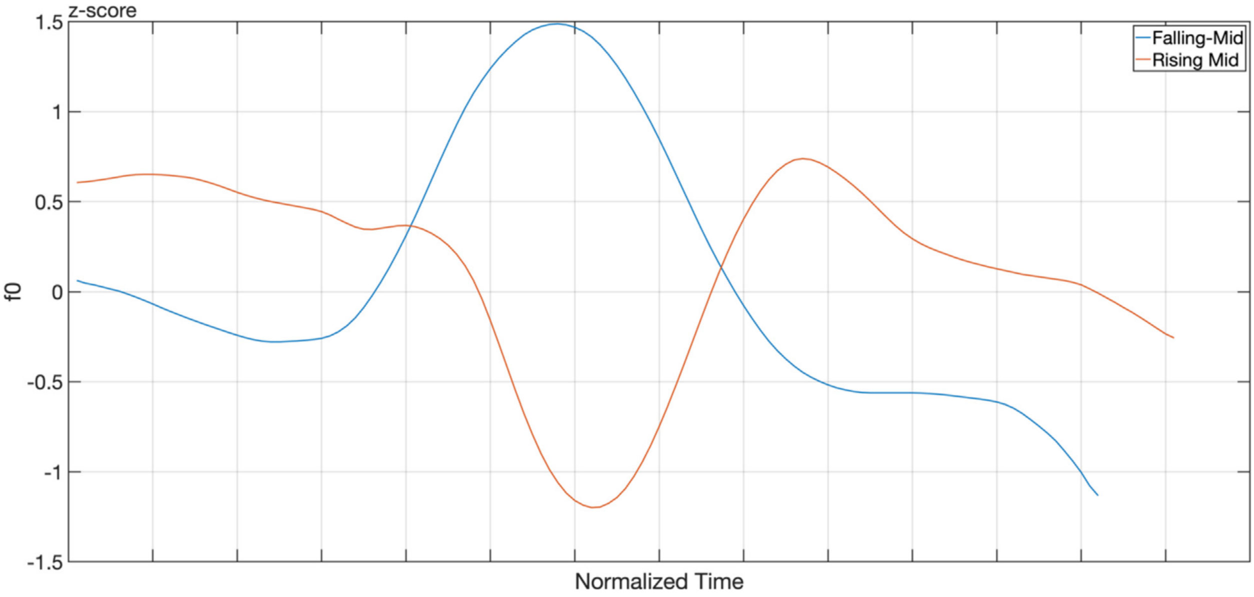

Fig. 10. Example of F-M and R-M f0 trajectories. Raw unsmoothed trajectories from single tokens obtained from data in Burroni & Kirby (2023).

Table 2
Experimental items.

w1	w2 (target 1)	w3 (target 2)	w4	w5
ดู [dūː]	หมี่ [mîː] หมี [mǐː]	มา [māː] หม่า [màː] หม้า [mâː] ม้า [máː] หมา [mǎː]	บน [bōn]	ดาว [dāːw] ดิน [dīn] ลัง [lāŋ]

Note that the target disyllabic combinations were chosen to consist entirely of voiced sonorant segments to facilitate the extraction of f0 contours. The last word of each sentence was never the same across two consecutive trials to act as a distractor and prevent participants from focusing solely on the disyllabic tonal combinations (w2 and w3).

and very fast (in Thai ช้ามาก, ช้า, ปกติ, เร็ว, เร็วมาก). In the training phase, each rate manipulation was repeated twice in the order just described. Note that the purpose of the rate instruction was not to elicit specific rates but simply to induce variation, and we used produced duration of the utterance as a predictor in our statistical models, not the categorical rate cues.

Target disyllabic tonal combinations were cued in Thai orthography with the text in black color. Above the carrier sentence, a speech rate indication appeared in the form of Thai orthography. The text appeared in five linearly interpolated color steps ranging from blue (very slow), light blue (slow), purple (normal), light red (fast), and red (very fast) to map rates to a color continuum. Note the color scheme did not interfere with randomization. After 1.5 s from the appearance of the target phrase and rate cue, a green bar appeared below the text to signal participants that they could start speaking.

To check that the rate cue induced variation in rate, we estimated the effect of the rate cue on response duration using linear mixed-effects regression, comparing a model with a fixed effect of rate (i.e., "Dur ~ rate cue + (rate cue|SP).") to one without that term. The rate cue was coded as a numerical variable ranging from 1 (very slow) to 5 (very fast). The addition of the fixed effect for rate cue significantly improved model fit ($\chi^2_{(1)}$ = 14.99, p < 0.0001, Adj. R2 = 0.85). The model estimates suggest that utterance duration decreases by about 500 ms at every step from very slow to very fast, Fig. 11.

In total, we collected 1698 tokens for the Falling tone analysis and 1837 for the Rising tone. A full breakdown of the tokens by subject and tonal conditions is presented in Appendix A. Not all the data could be landmarked and used for the analyses. In our dataset, this was mostly due to errors in intended tone production and changes in articulatory movements at faster speeds which made unequivocally identifying landmarks not possible. For this reason, 215 tokens (~12.6%) were discarded for the Falling tone and 163 tokens (~8.8%) for the Rising tone. Thus,1483 tokens for the Falling tone and 1674 for the Rising tones were left for analysis, for a total of 3157 tokens.

2.4. Data processing and independent variable extraction

Speaker-specific monophone Hidden Markov Models were trained in Kaldi (Povey et al., 2011) and used to perform forced alignment by speaker, followed by manual checking and correction. F0 was extracted using the Sum of Residual Harmonic algorithm (Drugman & Alwan, 2011) in MATLAB with a 52 ms window and a 10 ms overlap. For men, we used a range [60, 200] Hz and, for women, a range [100 400] Hz. The raw f0 trajectories were cleaned of f0 "jumps" greater than 20 Hz between frames. The contours were interpolated with cubic spline interpolation, and smoothed using a moving median followed by a moving average filter.

Missing values in articulatory trajectories were obtained with linear interpolation. The position of articulatory sensors was corrected for head movement. The trajectories were smoothed using a 3rd order low-pass Butterworth filter with a 10 Hz cut-

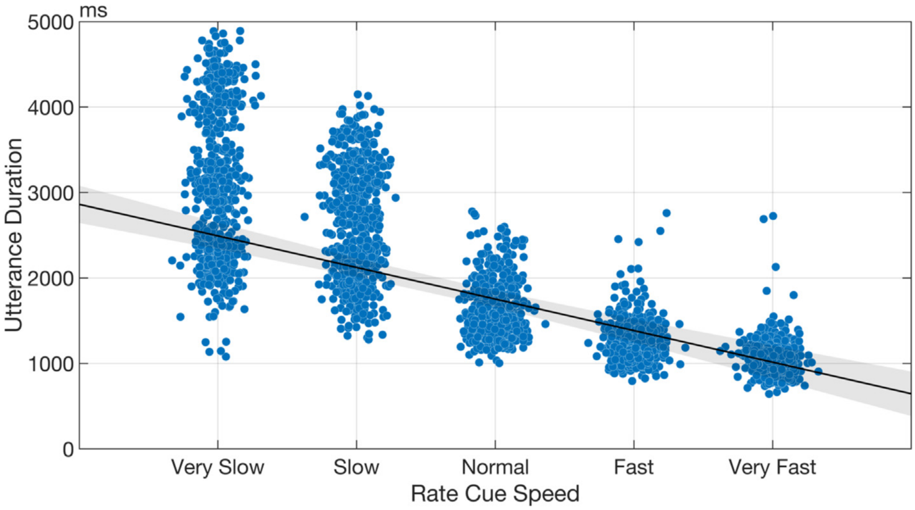

Fig. 11. Estimated effect of rate cue manipulation on utterance duration based on fixed effects. Dashed lines represent mean values, solid lines represent 95% CI. Note that utterance durations' x-values are horizontally jittered for illustration purposes. These effects suggest that the manipulation was successful and speakers modified their rate according to the received indications.

off. In this paper, we focus on the trajectories involved in the production of [m] and [iː] in w2. The [m] gesture closure and release were identified using a lip aperture (LA) time series. LA is defined as the Euclidean distance between the vertical and horizontal components of the LL and UL movements. Landmarks associated with the formation of [iː] were obtained from the horizontal component of the tongue body (TBx) movement. We focus on the horizontal component because the carrier sentence has a transition from [uː] to [iː] that is maximally differentiated on the horizontal plane. We confirmed this property by running a principal component analysis that combines both horizontal and vertical movement of the tongue body. We found that the direction of maximum movement, represented by the 1st PC, explains 91% of the variance on average and correlates almost perfectly with horizontal movement (median r = 0.98). These findings suggests that tongue movement during vowel production in the target word is mostly in the horizontal dimension.

Tones were landmarked based on the f0 trajectory. First, the f0 peak (for Falling tone), f0 minimum (for Rising tone), and positional extrema (for articulatory trajectories) were located. Then, velocity extrema that precede and follow the inflection/midpoint were identified. Positional extrema that precede and follow the first and second velocity extrema, respectively, were then identified. Within a region spanning the first positional extremum and the first velocity extremum, the **onset** of the movement, defined as the first time point at which the movement velocity surpasses 20% of the maximum velocity, was landmarked. We then located the gestural **target** as the last time point after the first velocity extremum and before the inflection point where movement velocity falls below a 20% threshold of maximum velocity. The operation was repeated between the inflection point and second velocity extremum to locate the release, and between the second velocity extremum and second positional extremum to locate the movement offset. The results of this landmarking procedure are illustrated with an example of [mĭː] in Fig. 12.

Our choice of adopting a threshold on peak velocity close to a velocity zero-crossing representing a positional extremum follows recommendations found in the literature for yielding the most stable results in the face of variation in kinematic trajectories' shape. (Bombien et al., 2013; Hoole et al., 1994; Kroos, 1996; Lorenc et al., 2023; Smith et al., 2019), with some work also suggesting that 20% is an ideal threshold (Kroos, 1996; Lorenc et al., 2023). There are four main aspects that we have considered in making this choice.

First, our choice aligns with the majority of papers. We collected a sample of papers published in the Journal of Phonetics over the last ten years to substantiate this claim. Out of thirty papers we identified in which EMA data was analyzed, twenty-two (73.3%) used a threshold on peak velocity, three used velocity zero-crossings (10%), two used acoustic boundaries (6.6%), and one each used acceleration zero-crossings (3.3%), Generalized Additive Mixed Models (3.3%), and peak velocity (3.3%). We also found that among the papers that do not use peak velocity thresholds to segment articulatory events, only two (6.6%) are concerned with segmenting articulatory events, while the others either analyze entire trajectories or measurements derived from the trajectory, e.g., angular distance among EMA sensors.

Second, we found that a 20% threshold on peak velocity offers more robust estimates for the onset of trajectories that are stationary in their initial portion with small fluctuations or wiggles due to motor or tracking noise. This is illustrated in Fig. 13, where the shaded area represents the acoustic boundaries of a bilabial nasal closure [m], the initial consonant of the target word [miː]. Note how the onset estimated from the velocity zero-crossing (top panel) occurs too early with respect to the acoustic boundaries compared to the onset estimated using 20% of peak velocity, occurring almost 100 ms earlier. This is a general issue with detecting gestural landmarks based on zero-crossings. Sometimes there is no zero-crossing to detect gestural onset, even though some values very close to a zero-crossing (mid panel dashed line) exist, they are not picked up algorithmically, Fig. 13 mid panel. Note also that there are no obvious peaks or minima in the acceleration signal that could be used to landmark gestures, Fig. 13 bottom panel.

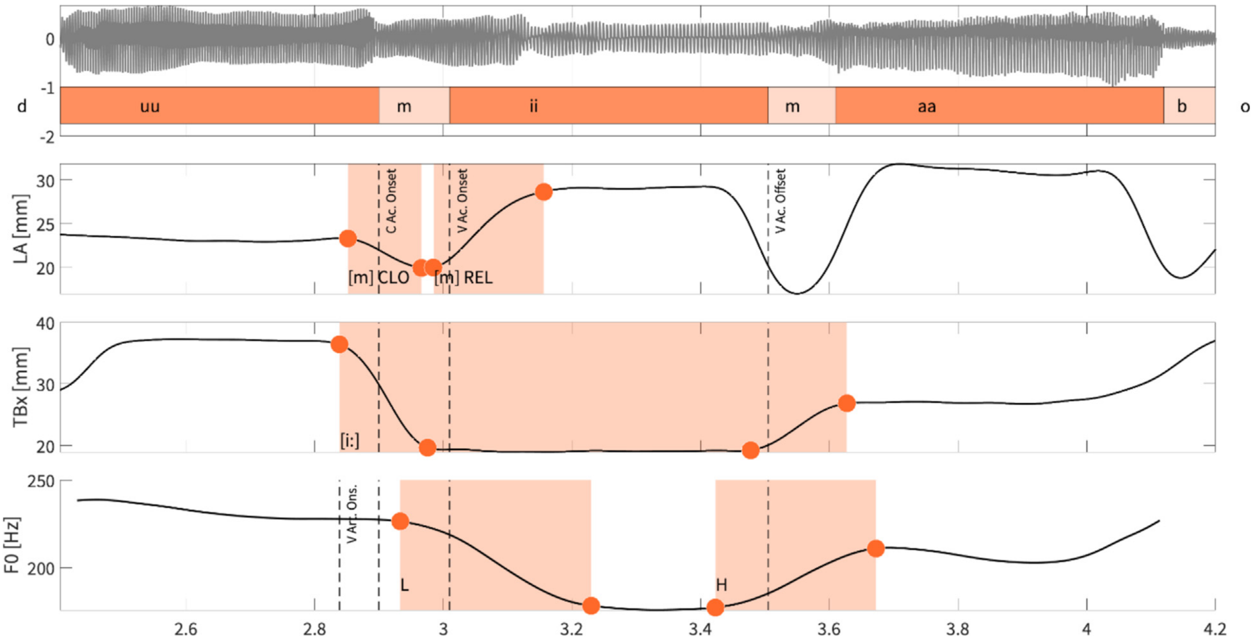

Fig. 12. Top: Example of waveform and segmental boundaries. Top-Mid: Lip Aperture (LA) time series used to identify onset and offsets (marked by dots) of closure (CLO) and release (REL) phases of the bilabial closure gesture [mi:]. Bottom-Mid: Tongue Body horizontal movement (TBx) time series used to identify onset, target, release, and offset (marked by dots) of the high front vocalic gesture. Bottom: f0 acoustic trajectories used to identify onset and target (marked by dots) of Low (L) and High (H) tonal gestures.

We also wish to note that to employ velocity zero-crossing trajectories need to be smoothed. In some studies, trajectories are smoothed heavily via local smoothers (Svensson Lundmark et al., 2021). Unfortunately, local smoothers are well-known to introduce delays in time series. Alternatively, PCA is often used to obtain a principal component of movements followed again by smoothing (e.g., Elie et al., 2023). Beyond the necessity for smoothing, this second technique has advantages, as PCA can rely on more than one dimension. However, it can also introduce substantial uncertainty in the relationship between individual articulator position in one dimension and the 1st PC trajectory.

Third, 10–20% thresholds on peak velocity are not specific to kinematic segmentation in speech research; they are also used in other domains of human movements, including electromyographic signals. This speaks to the robustness of this method and situates speech production research within the wider context of human movement (cf. Kuberski & Gafos, 2023 for a recent discussion of this point).

Fourth, our choice was motivated by a desire to use an identical algorithm for the segmentation of articulatory gestures and tone gestures. In so doing, we aimed for a compromise between previous work on tone that used even more extreme thresholding, (e.g., 30% in Yi, 2017), and previous work that

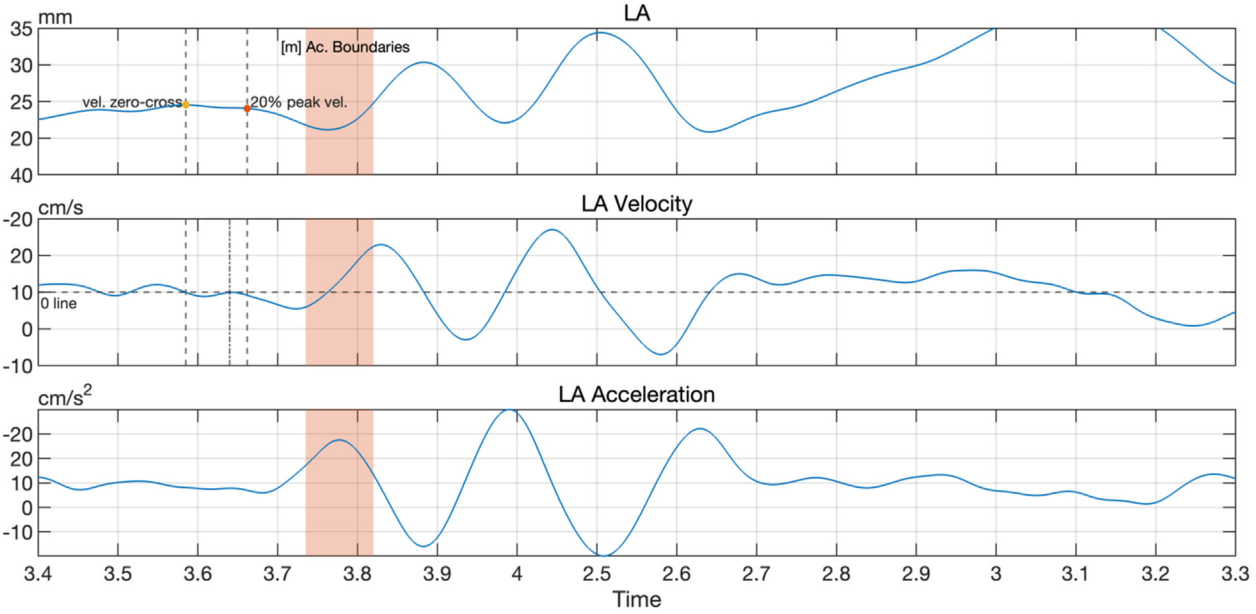

Fig. 13. Top: Lip Aperture (LA) values (top), velocity (mid) and acceleration (bottom) panel, together with onset of a bilabial closure gesture [m] estimated based on velocity zero-crossing and 20% peak velocity. Shaded region indicates the acoustic boundaries of the [m] gesture.

conceded being unable to use f0 velocity due to its noisy nature and had to resort to f0 positional extrema as proxies (Gao, 2008). Note specifically that positional extrema for f0 are not particularly meaningful in our data because the target F and R tones produced can have long plateau phases, Fig. 9, where the positional extremum position can fluctuate significantly.

From the landmarking, various lags between acoustic and articulatory landmarks were extracted. All acoustic and articulatory landmarks were inspected and corrected as needed by a research assistant who was a native speaker of Thai. Acoustic landmarks were obtained from forced alignment boundaries, which was also inspected and corrected as needed by the same research assistant.

To assess evidence for all logically possible combinations of landmarks considered here, we derived possible pairwise combinations of eight articulatory lags and five acoustic lags based on the landmarks described in Fig. 12. Our aim was to explore as many possible landmarks as we could, within the limits of coordination patterns that seemed reasonable. Our articulatory lags basically exhaust all possible combinations of articulatory gestures onsets, that is initiation based on landmarking, and targets, achievement of a target state based on landmarking, and initial tonal onsets or targets, i.e. the onset of f0 raising/lowering for the F/R tones, respectively. The lags are illustrated in Table 3 below:

This choice of lags allows us to explore a variety of possible landmark alignments and specifically to test whether tones are more closely coordinated with consonantal or vocalic gestures and whether that coordination is primarily driven to start or achieve a target at the same time, or both.

Additionally, we also examined all possible combinations of tonal onset and target with acoustic events of interest, such as the consonant/syllable acoustic onset, the vocalic acoustic onset (also representing the consonant offset), and the vocalic/syllable acoustic offset. Crossing these events gave us six lags of interest. From these six lags, we excluded the tonal onset and vocalic/syllable acoustic offset lag as we are not aware of models that posit a coordination of this type. The five acoustic lags are presented in Table 4 below:

The rationale for the onset lags requires little explanation as tones could be timed to syllable or vocalic acoustic onsets. Our reasoning for the target lags is that the target of the first portion of a F and R tone could also have its own acoustic timing pattern near to the acoustic boundaries of some events in the syllable (*cf*. Zsiga & Nitisaroj, 2007 and the discussion in Section 1.6).

After examining the landmarks indicated in the table above, we identified the two landmarks that are most stable in terms of their variability and stability to external perturbations within each modality, acoustic or articulatory. We then compared them with each other to assess whether speakers rely more on acoustic or articulatory modalities for tonal timing.

2.5. Statistical analyses

To test the hypotheses discussed in Section 1.5 regarding landmark modality and landmark types, we conducted variability and stability analyses within each modality. All analyses were conducted separately for Falling and Rising tones as they represent distinct data-generating processes.

Given the large number of lags explored and the fact that their exact distribution is unknown, the variability analysis based on bootstrapping of a robust measure, the median absolute deviation. We also point out that qualitatively identical results are obtained using a more commonly used measure, like the standard deviation (see Supplementary Material).

Stability was assessed by fitting a maximal mixed effects model to determine whether variation in the lag values is accounted for by the perturbing factors of rate and following tonal context. The model also included random intercepts by subject and random slopes for rate and tonal context by subject to accommodate between speaker variation. Note that the random effects intercepts and slopes were uncorrelated since models with correlated random effects had convergence issues. Speech rate was estimated as z-scored carrier sentence duration from which we subtracted the durations of the target word and of the following word, as these contain the target lags and are also used to manipulate tonal context. In doing so, we follow the recommendation of excluding target intervals, as including them may artificially inflate rate-target correlations and bias coefficient estimates (Tilsen & Tiede, 2023). Tonal context was treated as a categorical variable with baseline _M. The model coefficients were then inspected for overlap with zero to assess whether these factors influence a lag and, thus, to ascertain whether a particular lag is systematically influenced by our two experimental perturbations.

To compare stability and variability of lag between modalities, we conducted four parametric and non-parametric analyses on two less variable and more stable acoustic lags (syllable acoustic boundary to tonal onset) and articulatory lags (vowel articulatory onset to tonal onset) comparing their variability. All analyses were again conducted separately for F and R tones as they represent distinct data-generating processes. Variability was tested with a wider set of methods given that differences may be intrinsic to difference modalities and to data analysis procedures, as we have discussed in Section 1.5.

First, F-tests for variance were used to compare the variance of articulatory and acoustic lags to each other (e.g., Mücke et al., 2009). Second, since the lag distributions have

Table 3
Tone lag with articulatory events.

	Consonant Closure Gesture Onset (C Ons.)	Vocalic Gesture Onset (V Ons)	Consonant Closure Target (C Targ.)	Vocalic Gesture Target (V Targ.)
Tone Onset (T Ons)	T Ons. – C Ons.	T Ons. – V Ons.	T Ons. – C Targ.	T Ons. – V Targ.
Tone Target (T Targ.)	T Targ. – C Ons.	T Targ. – V Ons.	T Targ. – C Targ.	T Targ. – V Targ.

Table 4
Tone lag with acoustic events.

	Consonant/Syllable Acoustic Onset (C Ac. Ons.)	Vocalic Acoustic Onset (V Ac. Ons.)	Vocalic Acoustic Offset (V Ac. Off.)
Tone Onset (T Ons)	T Ons. – C Ac. Ons.	T Ons. – V Ac. Ons.	_
Tone Target (T Targ.)	T Targ. – C Ac. Ons.	T Targ. – V Ac. Ons.	T Targ. – V Ac. Off.

high kurtosis, we also conducted non-parametric tests for variance. Specifically, we conducted a bootstrapped version of the Brown-Forsythe test, which is claimed to be the most reliable test in the presence of deviations from normality assumptions and leptokurtic distributions (Lim & Loh, 1996; Wang et al., 2017). Our implementation follows the description of the bootstrapping procedure presented by Lim and Loh (1996), which, in turn, is a modification of that of Boos and Brownie (1989); our implementation is limited to two samples rather than multiple ones. The details of the test are described in Appendix B.

Third, to analyze the predictability and variability of the lag distributions, we calculated the entropy for both acoustic and articulatory lags. Entropy, in this context, is a measure of uncertainty or unpredictability within the distribution of the lags. First, we discretized the continuous lag data into discrete bins. Specifically, we divided the range of both acoustic and articulatory lags into 10 ms intervals. This binning process allows us to count the frequency of lag occurrences within each bin, which is necessary for the entropy calculation. Once the lag data were discretized into 10 ms bins, we calculated the probability distribution of the lags. Let $p(x)$ represent the probability of a lag falling into bin x. This probability is computed as the ratio of the number of lags in bin x to the total number of lags. Once we had obtained the probability distribution $p(x)$, we used the standard formula for entropy, which is given by:

$$H(X) = -\sum_{x=1}^{X} p(x)\log(p(x)) \tag{1}$$

In this formula, $H(X)$ is the entropy of the (discrete) random variable X, $p(x)$ is the probability of the lag occurring in the x-th bin, X is the total number of bins. The logarithm used is typically the natural logarithm, although other bases can be used depending on the context.

The summation is carried out over all bins, where each term $p(x)\log(p(x))$ represents the contribution of bin x to the total entropy. The negative sign ensures that the entropy value is positive since probabilities $p(x)$ are between 0 and 1, making $p(x)\log(p(x))$ negative. By summing these contributions, we obtain the entropy H, which quantifies the overall uncertainty in the lag distribution. A higher entropy value indicates greater unpredictability and variability in the lag timings, while a lower entropy value suggests more predictability and less variability. To compare the two lags we took their entropy difference to determine which lag has a more predictable distribution.

Fourth, we also conducted analyses of variability by subject to test whether acoustic and articulatory lags are more stable at the individual speaker level. To do so, we applied a jackknife procedure to the measure of variability that is least sensitive to outliers (Maronna et al., 2019), namely the Interquartile Range (IQR). This was necessary because the number of observations per speaker is much smaller, and a few outliers could have a disproportionately large impact compared to when pooling all speakers together. For every speaker, we calculated the IQR of acoustic and articulatory lags $n - 1$ times, where n is the number of samples, by systematically leaving out one sample and estimating the summary statistic of interest. This jackknife procedure allowed us to obtain 95% confidence intervals (CI) for both acoustic and articulatory lag IQR. By checking whether these confidence intervals overlap, we were able to determine whether the variability of acoustic and articulatory lags is significantly different from each other.

Stability was assessed, as for the tonal timing landmarks, by checking whether adding fixed effects for rate and following tonal context to a model containing random intercepts by subject and random slopes for rate and tonal contexts would improve model fit. Note that the random effects were uncorrelated since models with correlated random effects had convergence issues. The models for each dependent variable are reported in Appendix C and were based on stepwise fitting starting from the model described for landmarks.

Finally, as we have anticipated in section 1.5, to compare across modalities we supplemented traditional variability and stability analyses with an informativity analysis based on mutual information. In doing so, we took inspiration from previous work on coarticulation (Chen et al., 2015; Iskarous et al., 2013) and extended the analysis from spatial mutual information to temporal mutual information. Specifically, we calculated the mutual information (MI) of the tonal onset and the articulatory vowel onset timing versus the MI of the tonal onset timing and acoustic consonantal/syllable onset timing. The MI of two variables is defined as:

$$MI(x,y) = \sum_{x=1}^{X}\sum_{y=1}^{Y} p(x,y)\log\frac{p(x,y)}{p(x)(y)} \tag{2}$$

In Eq. (2), $p(x,y)$ represents the joint probability of the random variables x and y and $p(x,y)$ represents their joint conditional probability, while $p(x)(y)$ represent their joint probability obtained from the product of their marginals under an independence assumption. Thus, MI is a measure of how much information the two variables share or how much one can be "predicted" from the other. MI is a complementary way to assess which variables are more strongly dependent on each other in an information-theoretic manner, which has the advantage of being less sensitive to distributional assumptions.

Following previous work (Chen et al., 2015; Iskarous et al., 2013), probability estimations were computed using marginal and joint histograms, with counts smoothed using Jeffrey-Perks law by adding a count of 0.5 to all bins prior to probability computation. MI is known to be strongly affected by the number of bins used to obtain probability estimates. Accordingly, the results reported in this paper use both a coarse number of bins (20) and a finer number of bins (100) to demonstrate that MI patterns do not qualitatively depend on the binning procedure.

A way in which we departed from previous work is that we did not estimate MI once from the raw data. Instead, we relied on bootstrapping to obtain distributions of MI from subsamples of the data. Specifically, for each acoustic or articulatory event timing, and separately for the Falling and Rising tones, we drew k samples, where k is the sample size for that event, with replacement. We repeated the procedure 100 times and obtained 95% confidence intervals (CI) for the average MI between events.

There are two reasons behind this choice. First, this procedure aligns with the largely non-parametric approach adopted in this paper, which, in turn, takes into account the fact that properties of lag distributions at different scales (e.g., across and within subjects) are not known. Second, we wanted to ensure that the inferences we make regarding MI also hold across different subsamples of our entire dataset, as there are many sources of variability, e.g., different subjects, temporal microstructures in experiments, and others.

3. Results

Our results are organized around the two main research questions of this paper: tonal timing landmark types and tonal timing modality (*cf*. Section 1.5). First, we present an investigation of tonal timing landmarks based on their patterns of variability and stability in response to rate and tonal context manipulations. After identifying the landmarks that are most stable within each modality, we then compare these landmarks to address the question of tonal timing across modalities, determining which modality is more likely employed by speakers.

3.1. Landmarks for tonal timing

3.1.1. Articulatory landmarks variability and stability

For the F tone, we found that among the articulatory lags, the one with lowest variability, in the form of, measured by Median Absolute Deviation (MAD) is the lag between the tone onset and the vocalic gesture onset (T Ons. – V Ons., 95% CI [24 26] ms), Fig. 14. A slightly higher MAD is observed for tone onset and the consonant gesture target (T Ons. – C Targ., 95% CI [29 31] ms), tone onset and consonant gesture onset (T Ons. – C Ons, 95% CI [33 36] ms), and the tone onset and the vowel gesture target (T Ons. – V Targ., 95% CI [34 38] ms), Fig. 14.

The lags between tonal targets and articulatory events are more variable than those of tonal onsets and articulatory events. From least to most stable, we observe tone target and consonantal gesture target (T Targ. – C Ons., 95% CI [43 47] ms), tone target and vocalic gesture target (T Targ. – V Targ., 95% CI [49 53] ms), tone target and vocalic gesture onset (T Targ. – V Ons., 95% CI [60 65] ms), and tone target and consonantal gesture onset (T Targ. – C Ons., 95% CI [60 66] ms)) Fig. 14.

For the R tone we also found that, among the articulatory lags, the one with lowest MAD is the lag between the tone onset and the vocalic gesture onset (T Ons. – V Ons., 95% CI [40 45]), Fig. 15. A slightly higher MAD is observed for tone onset and the consonant gesture target (T Ons. – C Targ., 95% CI [48 56] ms), tone onset and consonant gesture onset (T Ons. – C Ons., 95% CI [49 55] ms) and tone onset and vowel gesture target (T Ons. – V Targ. 95% CI [54 65] ms), Fig. 15.

The lag between the tonal targets and articulatory events are again more variable than those of tonal onsets and articulatory events. From least to most stable, we observe tone target and vocalic gesture target (T Targ. – V Targ., 95% CI [59 63] ms), tone target and consonant gesture target (T Targ. – C Targ., 95% CI [60 65] ms), tone target and vocalic gesture onset (T Targ. – V Ons., 95% CI [77 85] ms), and tone target and consonantal gesture onset (T Targ. – C Ons., 95% CI [78 85] ms), Fig. 15.

For the F tone, we found that all articulatory lags are significantly affected by utterance duration (i.e., slower rates), except for the lag between tone onset and vocalic gesture onset, the same lag that was also found to be the least variable, Table 5. Interestingly, as utterance duration increases, the tone onset gets closer to the consonantal gesture onset by $-$ 20 ms per one z-score unit increase (~900 ms) in utterance duration, Table 5.

To understand why this effect occurs, we entertained the hypothesis that the tone may be initiated earlier because the vocalic gesture is initiated earlier and the two are coordinated to each other. Evidence in favor of this inference comes from running an identical linear mixed model on the lag between the consonantal and vocalic gesture lag. The vocalic gesture onset and the tone onset both get closer to the consonantal gesture onset as speakers speak more slowly (C Ons – V. Ons. Intercept = 14 ms, Duration = $-$ 13 ms). In other words, at slower rates, all gestures are starting more synchronously.

The opposite effect is observed for the lag between tone onset and consonant and vocalic gesture target. The tone lags more and more behind as speakers speak more slowly, as indicated by all the negative coefficients in Table 5. Finally, as speakers speak more slowly, the tonal target is delayed compared to the consonantal and vocalic gesture onsets and targets, indicating that the tonal targets are reached long after both the articulatory gesture onsets and targets. Such lags even increase with longer durations, as shown by all the positive coefficients in Table 5, indicating greater and greater distances at slower rates. This suggests that the various gestural landmarks and the tonal targets are "sliding" more and more apart.

The effects of following tonal contexts are rather limited; very small effects are observed for the tonal onset and vocalic gesture onset ($-$9_H) and tonal onset and consonantal gesture target ($-$7_H). These findings indicate that these lags are very slightly shortened in F–H combinations, Table 5. However, the effects are modest in size, and we are not sure whether any interpretation of them is of theoretical significance.

For the R tone, we found nearly identical patterns. All articulatory lags are significantly affected by longer utterance duration, i.e., slower rates, except for the tone onset and vocalic gesture onset lag, Table 6. Numerical values indicate the mean coefficient estimated from mixed effect linear regression, ns indicates non-significant effects.

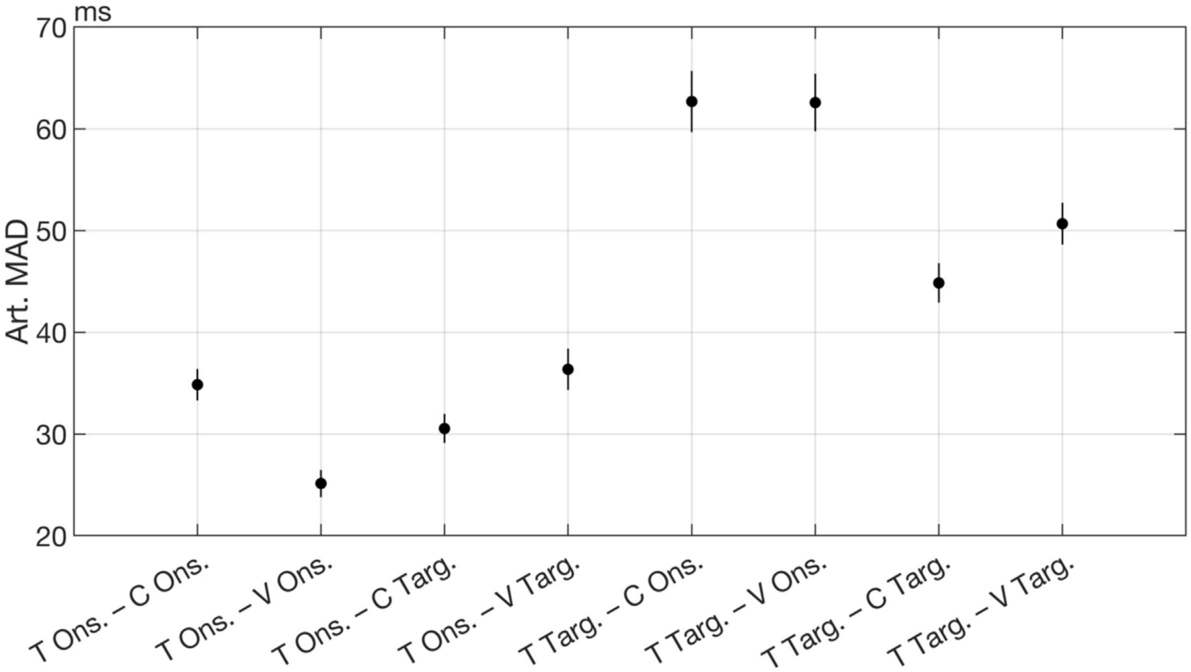

Fig. 14. Bootstrapped median absolute deviation between tonal onset or tonal targets and various articulatory events for Falling tone. The lags are ordered relative to tonal onset and targets (first four and last four lags) followed by the events that occurs earliest in time, namely consonantal gesture onset, vocalic gesture onset, consonantal gesture target, and vocalic gesture target.

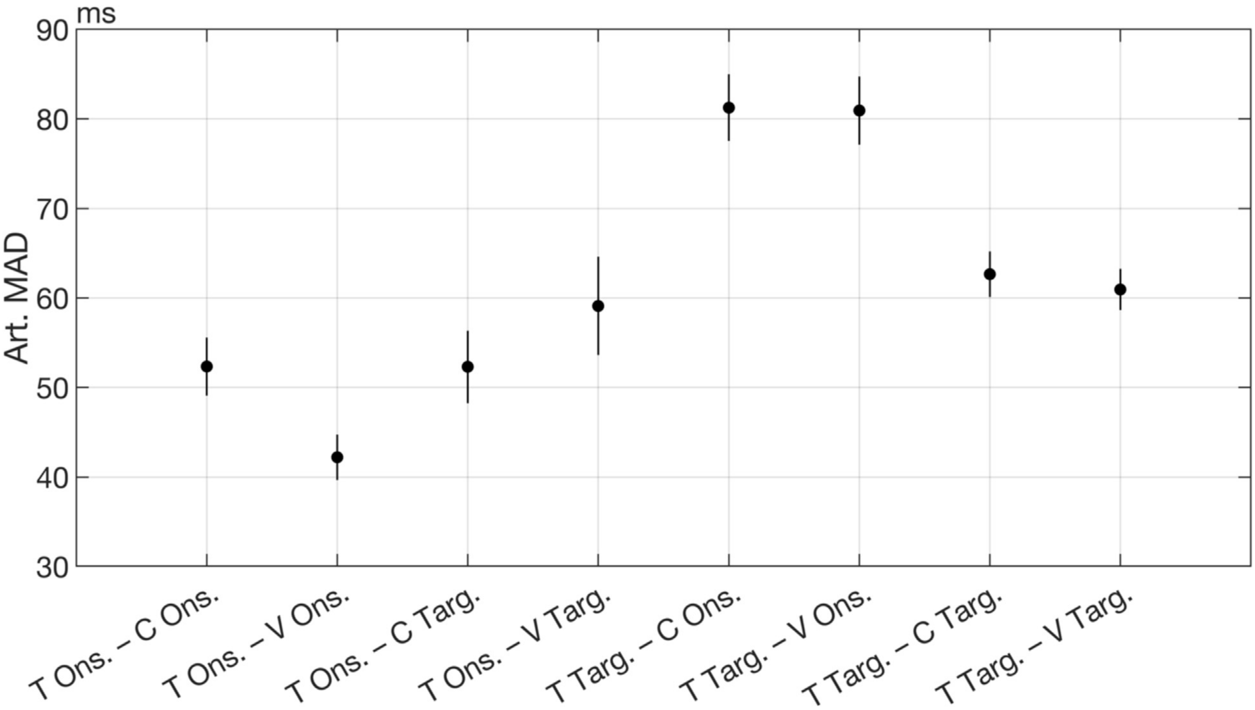

Fig. 15. Bootstrapped median absolute deviation between tonal onset or tonal targets and various articulatory events for Rising tone.

Interestingly, as utterance duration increases, the tonal onset gets again closer to the consonantal and vocalic gesture onset. This pattern reflects the fact that all gestures are starting more synchronously, as indicated by the negative coefficient for this lag, Table 6. The effect is again observed for the vocalic gesture as well, which is realized closer to the consonantal gesture at slower rates (C Ons – V. Ons. Intercept = 19 ms, Duration = – 10 ms), just like for the Falling tone.

The opposite effect is observed for the lag between tonal onset and consonant and vocalic gesture target: the tone lags more and more behind as speakers speak more slowly, just like for the F tone, as indicated by the negative coefficients

Table 5
Values for different lags between Falling tone onset and articulatory events estimated from mixed effect modeling. Intercept indicates average value in_M context, each column represents the effect of increasing duration or changing following tonal context on the value of the lag.

	Intercept	Duration	_L	_F	_H	_R
T Ons. – C Ons.	71	−20	ns	ns	ns	ns
T Ons. – V Ons.	57	ns	ns	ns	−9	ns
T Ons. – C Targ.	−27	−34	ns	ns	−7	ns
T Ons. – V Targ.	−96	−40	ns	ns	ns	ns
T Targ. – C Ons.	276	58	ns	ns	ns	ns
T Targ. – V Ons.	262	71	ns	ns	ns	ns
T Targ. – C Targ.	178	44	ns	ns	ns	ns
T Targ. – V Targ.	110	38	ns	ns	ns	ns

for already negative lags, Table 6. Finally, as speakers speak more slowly, the tonal target is delayed compared to the consonantal and vocalic gesture onsets and targets, as indicated by the positive coefficients for last four positive lags in Table 6. These coefficients indicate that the tonal targets are reached long after the articulatory gestures onsets and targets and that such lags even increase with longer durations. The effects of following tonal contexts are slightly more pronounced for the R tone compared to the F tone.

No effects are observed on lags involving the tonal onset, however, subtle increases in lag duration are observed in_F and_H for all lags that include the tonal target, with coefficients ranging between 12 and 20 ms increases, Table 6. Likely, these effects are motivated by more extreme, i.e., lower tonal target before tones that involve a high target, a so-called pre-high lowering effect that we have observed in our previous acoustic work (Burroni, 2023a).

To take stock of the patterns observed in the articulatory modality, we found that the least variable and most stable articulatory lag is the lag between the tone onset and the vocalic gesture onset. Unlike this lag, other lags between tonal onsets and articulatory gestures are more variable and can change their profiles under the influence of rate, becoming more synchronous in some cases (e.g., the tone onset and consonantal gesture onset lag) or more asynchronous in others (e.g., consonantal/vocalic gesture target and tonal target lags).

3.1.2. Acoustic landmarks variability and stability

For the F tone, we found that the acoustic lag with lowest variability is the lag between the tone onset and the consonantal/syllable acoustic onset (T Ons. – C Ac. Ons., 95% CI [27 30] ms), Fig. 16.

Slightly higher MADs are observed for the tone onset and vowel acoustic onset (T Targ. – V Ac. Ons, 95% CI [36 40] ms), the tone target and the vowel acoustic onset (T Ons. – V Ac. Ons, 95% CI [36 39] ms) and the tone target and the consonant/syllable acoustic onset (T Targ. – C Ac. Ons., 95% CI [50 54] ms). The lag between the tone target and the vowel/syllable offset is much more variable (T Targ. – V Ac. Off., 95% CI [138 148] ms), Fig. 16.

For the R tone, similar patterns emerged. We found that the acoustic lag with lowest variability is the lag between the tone onset and the consonantal/syllable acoustic onset (95% CI [44 51] ms), Fig. 17.

Slightly higher MADs are observed for vowel acoustic onset and the tone target (95% CI [51 55] ms), the tone onset and vowel acoustic onset (95% CI [51 60] ms), and the tone target and the consonant/syllable acoustic onset (95% CI [66 73] ms). The lag between the tone target and the vowel offset is much more variable 95% CI [108 120] ms), Fig. 17.

Turning to variability, for the F tone, we found that all acoustic lags are significantly affected by utterance duration (i.e., slower rates). However, the lag that is closest to a 0 value, indicating temporal coincidence and that is least affected by rate is the lag between tone onset and syllable acoustic onset. This same lag that was also found to be the least variable, Table 7. The lag between the tone onset and the vowel acoustic onset is negative and becomes more negative at longer durations, indicating that the tonal onset leads the vowel acoustic onset and that it leads more and more. This is reflected in the negative coefficient estimated for the duration effect, Table 7.

On the other hand, the tonal target is quite far in time from both the consonantal acoustic onset (213 ms) and the vocalic acoustic onset (116 ms); while the tonal target also lags behind the vowel/syllable acoustic offset (−156 ms). The distance between all these events gets larger and larger as speech rate slows down, as shown by coefficient signs in Table 7.

Turning to variability, for the R tone, we found that all acoustic lags are significantly affected by utterance duration (i.e. slower rates). Like for the Falling tone, the lag that is both closest to temporal coincidence and that is least affected by rate is the lag between tone onset and syllable acoustic onset. This same lag was also found to be the least variable, Table 8. The lag between the tone onset and the vowel acoustic onset

Table 6
Values for different lags between Rising tone onset and articulatory events estimated from mixed effect modeling.

	Intercept	Duration	_L	_F	_H	_R
T Ons. – C Ons.	69	−19	ns	ns	ns	ns
T Ons. – V Ons.	51	ns	ns	ns	ns	ns
T Ons. – C Targ.	−31	−35	ns	ns	ns	ns
T Ons. – V Targ.	−94	−38	ns	ns	ns	ns
T Targ. – C Ons.	300	85	ns	13	15	ns
T Targ. – V Ons.	282	97	ns	17	13	ns
T Targ. – C Targ.	200	68	ns	20	12	ns
T Targ. – V Targ.	137	65	ns	18	13	ns

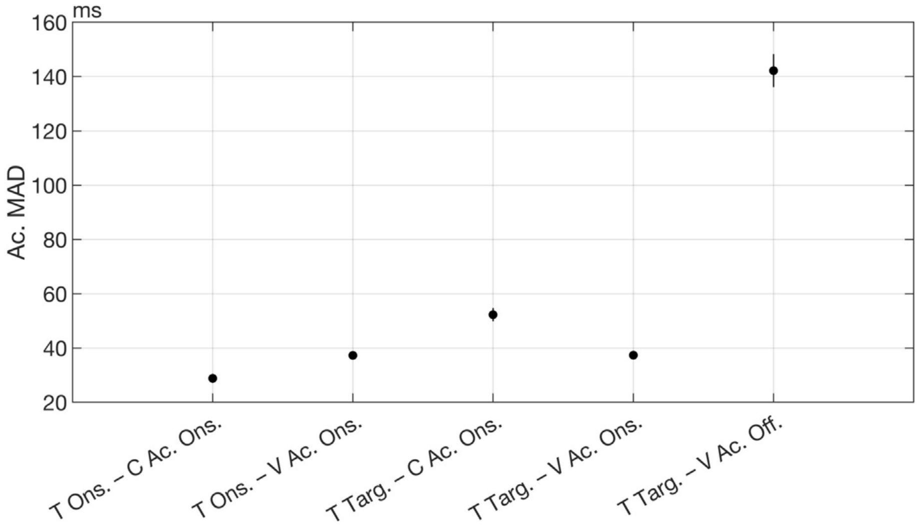

Fig. 16. Bootstrapped median absolute deviation between tonal onset (T Ons.) or tonal targets (T Targ.) and various articulatory events for Falling tone. The lags are ordered relative to tonal onset and targets (first four and last four lags) followed by the events that occurs earliest in time, namely consonantal acoustic onset (C Ac. Ons.), vocalic acoustic onset (V Ac. Ons.), and vocalic acoustic offset (V Ac. Off.).

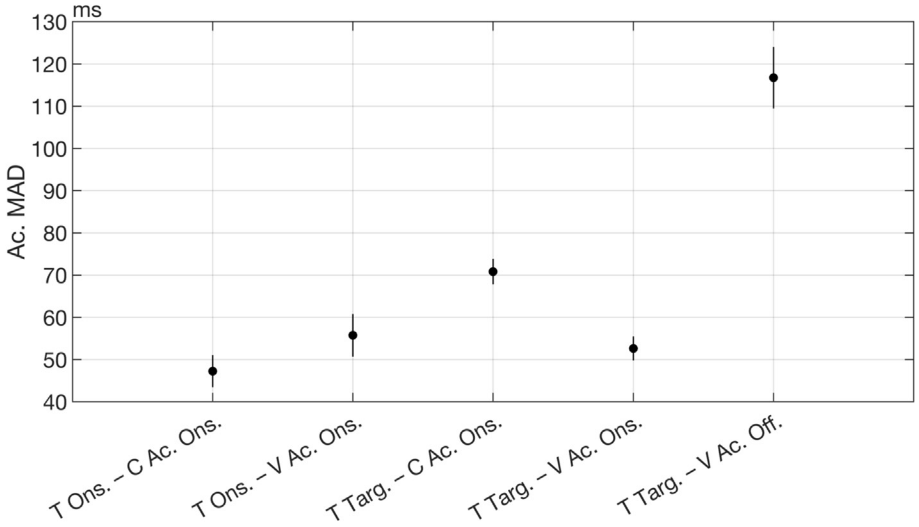

Fig. 17. Bootstrapped median absolute deviation between tonal onset (T Ons.) or tonal targets (T Targ.) and various articulatory events for Rising tone. The lags are ordered relative to tonal onset and targets (first four and last four lags) followed by the events that occurs earliest in time, namely consonantal acoustic onset (C Ac. Ons.), vocalic acoustic onset (V Ac. Ons.), and vocalic acoustic offset (V Ac. Off.).

Table 7
Values for different lags between Falling tone onset and acoustic events estimated from mixed effect modeling.

	Intercept	Duration	_L	_F	_H	_R
T Ons. – C Ac. Ons.	0	−24	ns	ns	ns	ns
T Ons. – V Ac. Ons.	−90	−46	ns	ns	ns	ns
T Targ. – C Ac. Ons.	213	54	ns	ns	ns	ns
T Targ. – V Ac. Ons.	116	32	ns	ns	ns	ns
T Targ. – V Ac. Off.	−156	−160	ns	ns	ns	ns

Table 8
Values for different lags between Falling tone onset and acoustic events estimated from mixed effect modeling.

	Intercept	Duration	_L	_F	_H	_R
T Ons. – C Ons.	0	−25	ns	ns	ns	ns
T Ons. – V Ons.	−91	−44	ns	ns	ns	ns
T Ons. – C Targ.	233	78	ns	19	16	ns
T Ons. – V Targ.	140	59	ns	15	13	−11
T Targ. – C Ons.	−124	−133	ns	19	20	ns

is negative and becomes more negative at longer duration, indicating that the tonal onset leads and leads more and more at slower rates, as indicated by the negative coefficient of the duration effect, Table 8.

On the other hand, the tonal target is quite far in time from both the consonantal (233 ms) and vocalic acoustic onset (140 ms); while the tonal target also lags behind the vowel/syllable acoustic offset (-124 ms). The distance between all these events gets larger and larger as speech rate slows down, as shown by coefficient signs matching the positive and negative lags in Table 8.

To take stock of the patterns observed in the acoustic modality, we found that the least variable and most stable acoustic lag in the acoustic modality is the lag between the tonal gesture onset and the syllable/consonantal acoustic onset. Unlike this lag, other lags between tonal onsets and acoustic boundaries are more variable and change their profile more under the influence of rate, in ways that suggest less synchrony.

In sum, in this section, we have investigated various possible landmark configurations that can be used to for tonal timing. Investigating separately within each modality, we found that the most stable articulatory lag is the lag between tonal onset and vocalic gesture onset in articulation and the lag between the tonal onset and the acoustic syllable onset in acoustics. This is in line with the predictions of AP and the SAH, respectively. We now turn to comparing these two lags against each other to investigate the question of whether speakers are more likely to be relying on articulatory or acoustic timing in tone production. Henceforth we refer to them as articulatory and acoustic lag for shorthand, respectively.

3.2. Tonal timing modality

For tonal timing modality, we present the results of our analyses organized along the lines of variability, stability to rate and tonal context manipulations, additionally, given that modality-intrinsic differences may exist, as we have pointed out (Section 1.5), we supplement such analyses with an informativity analysis based on mutual information.

3.2.1. Variability: F-tests, bootstrapped brown-forsythe, entropy, jackknife

Separately for F and R tones, F-tests were conducted to assess whether lags between tonal onset and vowel articulatory onset and tonal onset and syllable acoustic onsets have equal variance.

For the Falling tone, we found that the articulatory lag has a lower variance with a ratio to the acoustic lag estimated at around $\sim$ 0.7 ($F_{(1483,1483)}$ = 0.72 95% CI [.65 0.80], p < 1e-4). For the Rising tone, we found that the articulatory lag has a lower variance with a ratio to the acoustic lag estimated at around $\sim$ 0.85 ($F_{(1674,1674)}$ = 0.83 95% CI [.75 0.91]], p < 1e-4).

Since F-tests are sensitive to departures from normality of the underlying data distributions (e.g., Brown & Forsythe, 1974; Lim & Loh, 1996), we assessed whether the articulatory or acoustic lags may be deviating from normality. We observed that the distributions of the two lags, separately by tone, have higher probability values concentrated in their tails than expected under a normal distribution, in other words, they are all leptokurtic, with kurtosis $\kappa > 3$. This is shown by the fact that scaled t-distribution, which is also leptokurtic, provides a better fit to the data, Fig. 18.

Given that the lag distributions depart from normality and, specifically, that they are leptokurtic, we conducted more robust testing for homogeneity of variance that takes this fact into account, the bootstrapped Brown-Forsythe test. The bootstrapped Brown-Forsythe procedure, Appendix B, confirms that articulatory lags have a lower variance compared to the acoustic lags for both the Falling (p = 0.002) and the Rising tone (p = 0.04).

An additional piece of evidence that further supports lower variability of the articulatory timing to vocalic onset comes from the entropy of their distributions. Entropy is a general measure of uncertainty in a distribution. A distribution with lower entropy is more "predictable", in that it can be described using fewer units of information. A distribution with higher entropy, on the other hand, requires more units of information to be described. By taking a difference between the articulatory minus acoustic entropy we can measure whether the distribution of articulatory lags is more predictable ($H_{Art} - H_{Ac} < 0$), equally predictable ($H_{Art} - H_{Ac} = 0$) or less predictable ($H_{Art} - > 0$) than the distribution of acoustic lags. For both the F and the R tones, the difference of articulatory to acoustic lags entropy is lower than zero, Fig. 19, indicating that articulatory lags are more predictable.

An important consideration tied to the lower variability emerging from the analyses presented above is that this lower variability is not accompanied by temporal coincidence. As is evident from the distribution in Fig. 19, the acoustic lags gravitate closer to a central tendency value of 0, while the articulatory lags are consistently positive. Bootstrapping measures of central tendency for the median confirms this observation, Fig. 20. Bootstrapped 95% CI for the median of the acoustic lags are [7 12] ms for the F tone and [8 12] ms for the R tone. On the other hand, for the Articulatory lags, 95% CI for the median are [49 53] ms for the Falling tone and [50 53] ms for the Rising tone.

In other words, the tonal and acoustic syllable onsets tend to be temporally closer, but with more variability than the articulatory onsets. On the other hand, the lag between tone onset and vocalic gesture onset is more stable in terms of its values, but these values are overwhelmingly positive. Human movement lags tend to display a correlation between their mean value and their variability: the longer the interval, the greater its variability, at least for relatively long movements (Schöner, 2002; Shaw et al., 2011; Wing & Kristofferson, 1973). Thus, the fact that the longer lag between the articulatory tone onset and the vocalic gesture onset shows lower variability, despite its longer duration, can be seen as further evidence that this coordination is stable and controlled by speakers, albeit not necessarily in terms of temporal coincidence. Alternatively, it is also known that the relationship between mean and standard deviation may not be observed for relatively short movement intervals (Wing & Kristofferson, 1973).

A final issue that we need to address is whether the group pattern of less variable articulatory lags for tones are observed also at the individual level. To tackle this question, we used a jackknife procedure to get (a distribution of) estimates of vari-

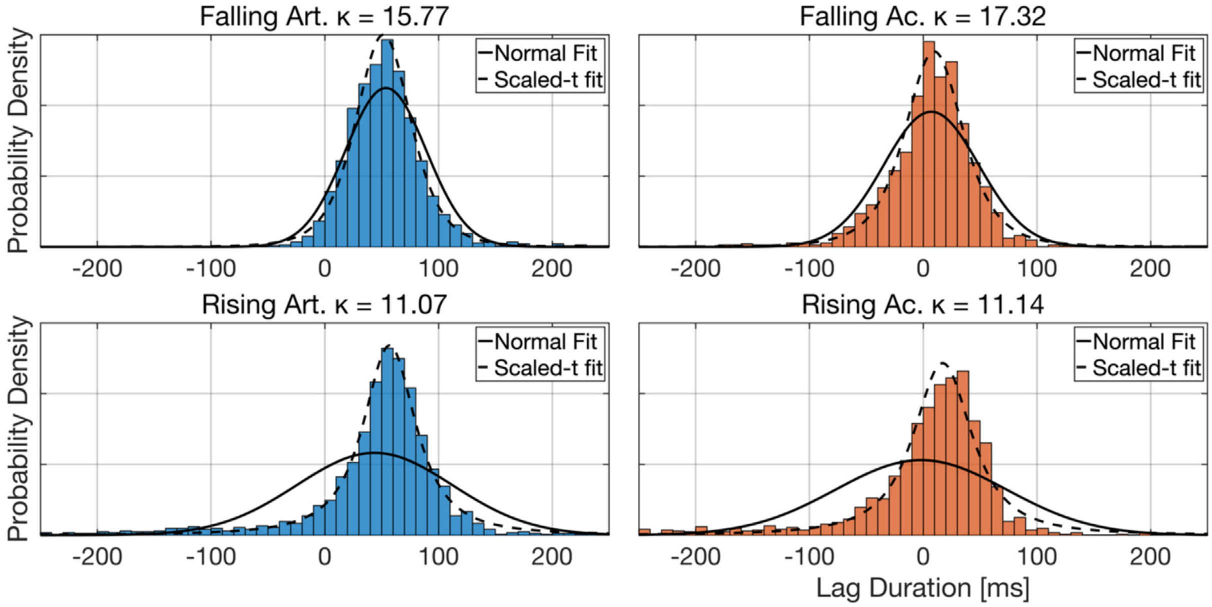

Fig. 18. Histograms of articulatory (right) and acoustic lags (left) separately by tone (top: Falling, bottom: Rising) with superimposed normal and scaled-t probability density function fits. Note that all distributions are leptokurtic as indicated by their kurtosis value κ > 3.

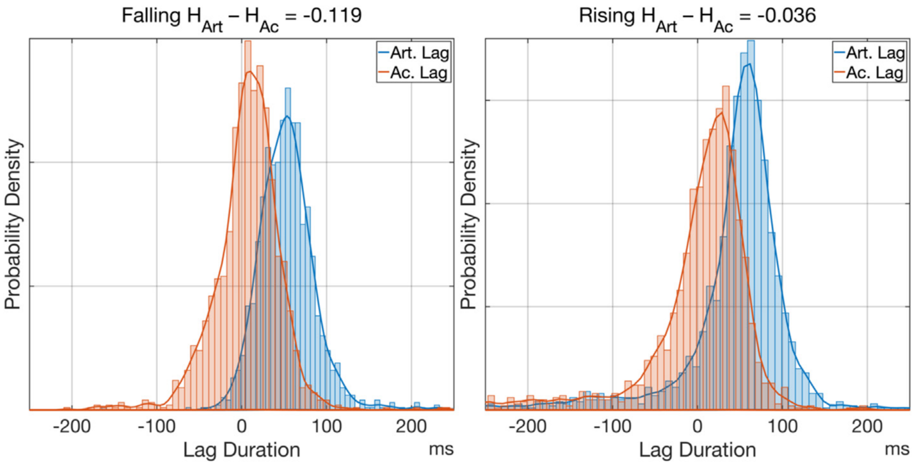

Fig. 19. Histograms of acoustic and articulatory lag durations together with entropy ratios for Falling tone (left) and Rising tone (right).

ability, the interquartile ranges (IQRs), of articulatory and acoustic lags by participant and tone. We used a jackknife procedure applied to IQR because it is less sensitive to outliers than other measure like the MAD (Maronna et al., 2019). For the Falling tone, we observed that 6/8 speakers (SP2, SP3, SP4, SP6, SP7, SP8) display less variable lags between tone onset and vocalic gesture onset, Fig. 21 top. For the Rising tone, 7/8 speakers display significantly a lower IQR for the same articulatory lag (SP1, SP2, SP3, SP5, SP6, SP7, SP8), Fig. 21 bottom.

Grouping the data by individual speakers drastically reduces the number of observations and so can exaggerate the influence that a small number of outliers can play on variability patterns. Thus, caution should be exercised in interpreting the patterns at the individual level. Additionally, we should note that for many speakers in Fig. 21 the differences in variability between acoustic syllable timing and articulatory timing to the vocalic gesture onset are rather subtle.

Keeping in mind these limitations of individual-level analyses, the fact that 5/8 (SP2, SP3, SP6, SP7, SP8) speakers display less variable articulatory timing to the vocalic onset for both tones lends further support to the notion that the lag between the tonal onset and the vocalic gesture onset seems more stable than the lag between tonal onset and acoustic syllable boundary. The fact that all speakers display more stable articulatory lags for at least one tone and that none among the eight speakers displays more stable acoustic lags across both the F and R tones provides further evidence for the hypothesis that landmark timing is controlled in the articulatory domain.

3.2.2. Stability: Rate and surrounding tonal context effects

Turning to stability, for the F tone articulatory lag, loglikelihood ratio tests showed that a model with a term for following tone is preferable to one without ($\chi_{(2)}$ = 10.78, p = 0.03) The difference is driven by slightly shorter lags in the_H context where the lag is −8.5 ms shorter (95% CI [-15–2] ms, t

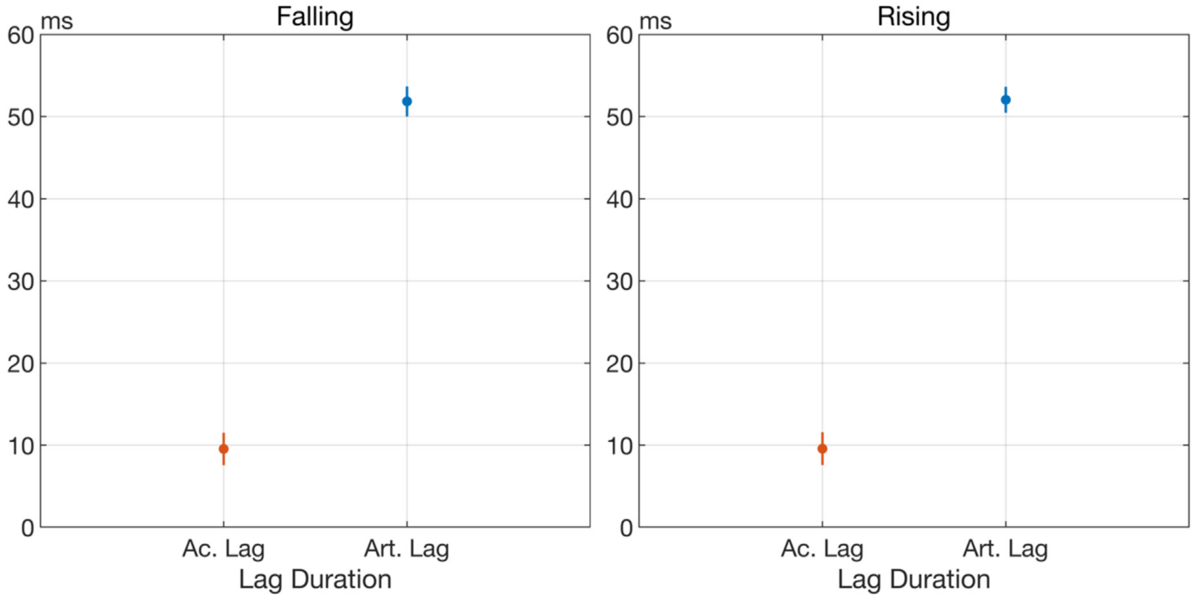

Fig. 20. Bootstrapped 95% CI for median of acoustic and articulatory lags.

(1478) = 2.66, p = 0.007), as we had already pointed out in Section 3.1.1. No other tonal context has a significant effect, nor does speech rate whose inclusion is not necessary in the best-fitting model. In contrast, for the F tone acoustic lags, log-likelihood ratio tests showed that an alternative model is preferable to the null model ($\chi_{(2)}$ = 7.84, p = 0.005) when a term for rate is included. Specifically, for one z-score unit increase in utterance duration, the acoustic lag duration decreases of −30 ms (95% CI [-48–13] ms, t(1478) = -3.52, p = 0.0004). No effect of following tonal context is observed and the term is not included in the final model.

For the R tone articulatory lags, a model with terms for following tone and rate is not preferable to model without those terms ($\chi_{(2)}$ = 0.82, p = 0.66). In contrast, for the R tone acoustics lags, we found via loglikelihood ratio testing that a model with rate is preferable to one without ($\chi_{(2)}$ = 12.20, p = 0.0004). Specifically, for one z-score unit increase in carrier duration the acoustic lag duration decreases of −54 ms (95% CI [−48–13] ms, t(1671) = −5.15, p < 1e-5).

To summarize, the articulatory lags are not affected by rate. For the F tone only, we observed a small effect of following tonal context in the form of a −8.5 ms reduction in the_H context. On the other hand, the acoustic lags are robustly affected by rate with decreases in their duration at longer rates with effects estimated at −30 ms and −54 ms for the F and R tones respectively, Fig. 22. This provides further evidence in support of the hypothesis that temporal control is accomplished in the articulatory domain.

3.2.3. Informativity: Data generation and mutual information

Turning to informativity, we aimed to assess whether the timing of the vocalic gesture onset is a better predictor than the acoustic syllable onset of the timing of the tonal onset. Bootstrapped confidence intervals of mutual information (MI) show that the MI between the tonal onset and the articulatory vocalic gesture onset is always higher than with the acoustic consonant/syllable onset, no matter the number of bins or the tone involved, Fig. 23.

Using 20 bins, for the F tone, bootstrapped 95% CI are [.21 29] and [.168 0.174] for the tonal onset and the vocalic gesture onset and for the tone onset and acoustic syllable onset respectively. For the R tone, bootstrapped 95% CI are [.20 0.27] and [.169 0.175] for the tonal onset and the vocalic gesture onset and for the tone onset and acoustic syllable onset, respectively.

Using 100 bins, for the F tone, bootstrapped 95% CI are [1.63 1.67] and [1.55 1.59] for the tonal onset and the vowel gesture onset and for the tone onset and acoustic syllable onset respectively. For the R tone, bootstrapped 95% CI are [1.61 1.65] and [1.57 1.6] for the tonal onset and the vowel gesture onset and for the tone onset and acoustic syllable onset, respectively.

To summarize, the tonal onset shares more information, and it is better predicted by, the articulatory vocalic gesture onset. This is irrespective of the tone involved and the number of bins used. MI analyses indicate, thus, that a stronger dependency holds between tonal onset timing and an articulatory event onset timing than between the tonal onset timing and an acoustic event timing, the syllable acoustic onset. Higher MI between movements is known to reflect coordination and temporal coupling, even at the neural level (Gupta & Bahmer, 2019), which underlies the production of said movements. The informativity analysis thus provides additional evidence that timing control privileges articulatory events rather than acoustic ones.

4. Discussion

Two research questions were investigated in this paper: (i) which types of landmarks are employed in control of tonal timing and (ii) which modality – articulatory or acoustic – are those landmarks defined in. These two questions relate to hypotheses associated with different phonological conceptions of tone production, developed in Autosegmental-Metrical Phonology/SAH and Articulatory Phonology. The results ultimately lend plausibility to the Articulatory Phonology perspective, in which

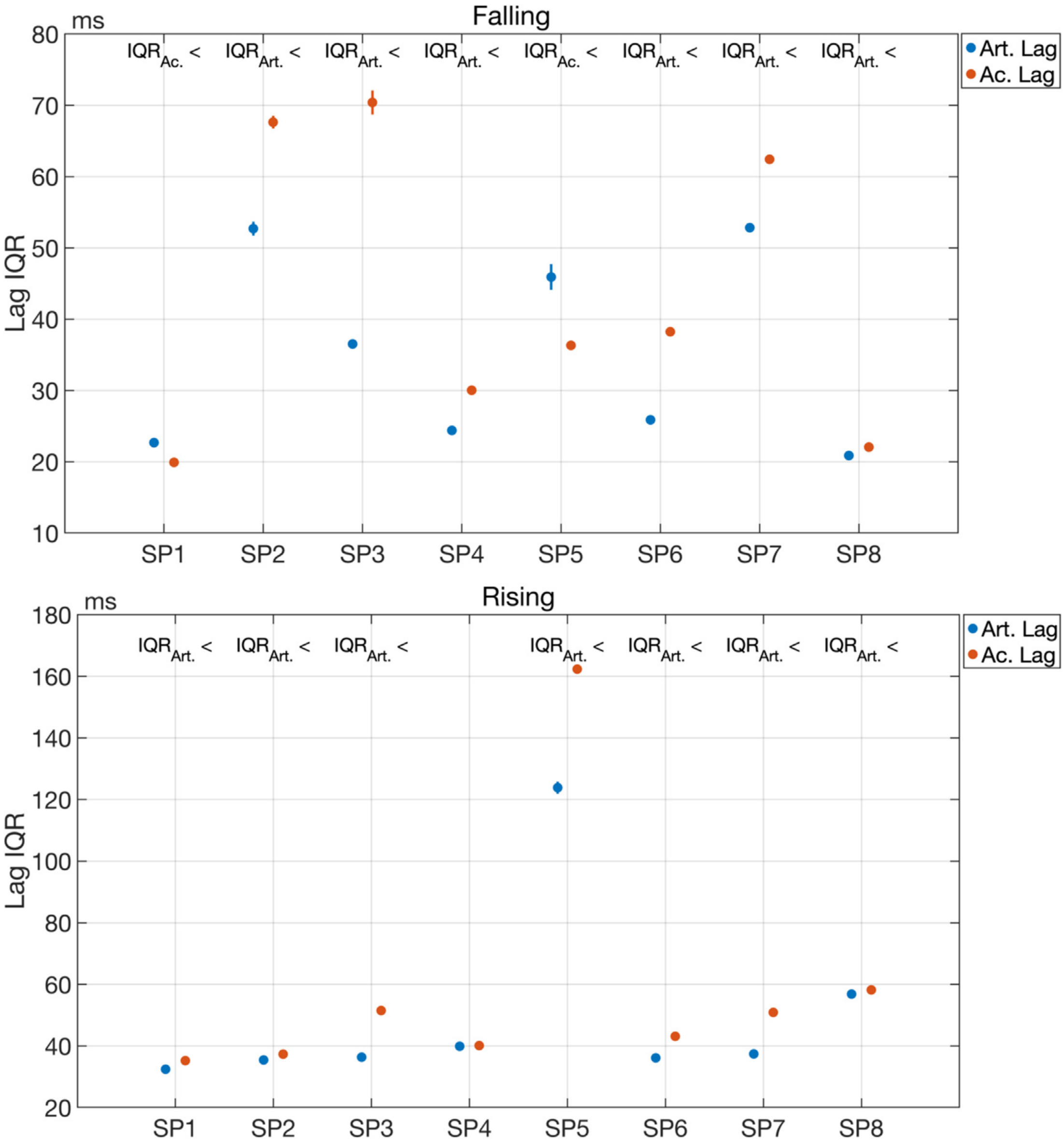

Fig. 21. By speaker 99% CI of IQR estimated using a jackknife procedure for Falling Tones (top) and Rising tones (bottom). IQR_{Art} < and IQR_{Ac} < indicate less variability in acoustic and articulatory lags respectively. The absence of text indicates no significant differences in the variability of acoustic and articulatory lags. Dots represent mean value and error bars 2 standard deviations above and below the mean for the median absolute deviation for each speaker.

speakers control the timing of tone and gestural onset events in the articulatory domain over the alternatively models where tones are implemented as tonal targets aligned to the segmental string.

We now discuss how our findings lead to this conclusion in more detail.

4.1. Tonal timing landmarks: Tonal onsets, tonal targets, and their implications for theories of speech timing

For tonal timing landmarks, we examined variability and stability measures separately in each modality to infer whether speakers use onset-to-onset timing, target-to-target timing, or both. For articulation, we found that, for both the F and R tones, the least variable lag is between the tonal gesture onset, estimated from the f0 trajectory as detailed in Section 2.3, and the vocalic gesture onset, for articulation, or the consonantal/syllable acoustic onset, for acoustics.

In articulation, the lag between the tone onset and the vocalic gesture onset is also the most stable in the face of rate and tonal context variation, except for a modest decrease in size in F-H sequences. Other articulatory landmarks, on the contrary, not only are more variable, but they are also systematically affected by rate, especially those that involve gestural targets, either tonal or consonantal, or vocalic. Similar patterns of a higher stability and lower variability of onset-to-onset timing were also observed within the acoustic modality, providing evidence for the widely held notion that tones may start synchronized to an acoustic syllable. However, we note that the lag with the acoustic syllable onset is indeed less variable than other landmarks, but just like other landmarks, it is also influenced by rate and therefore relatively unstable. We return to

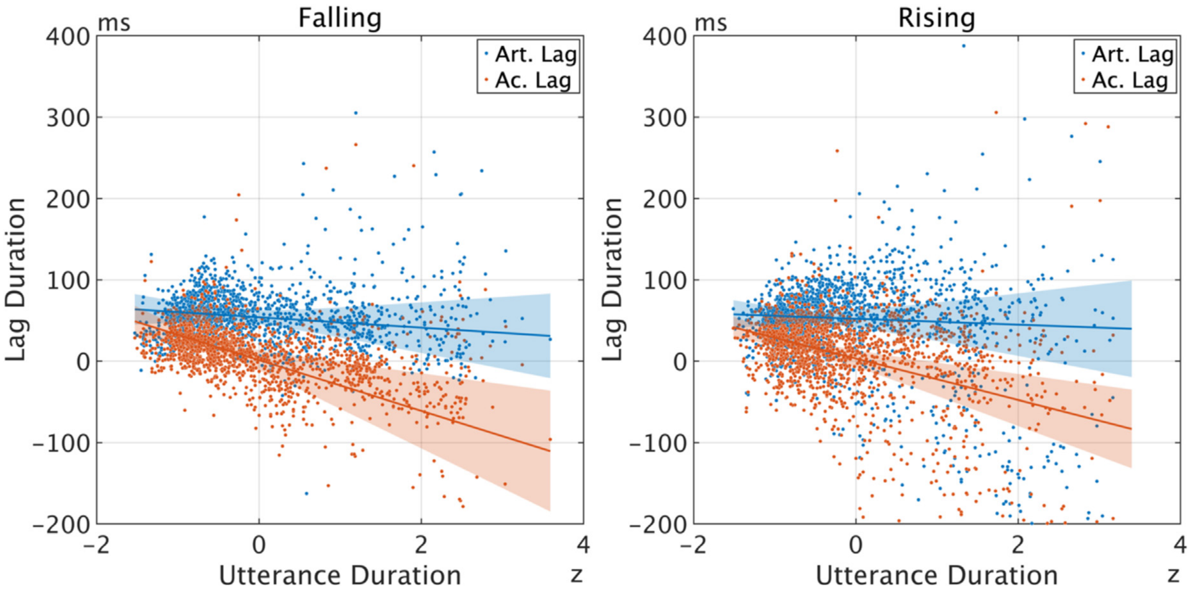

Fig. 22. Effects of utterance duration on the acoustic and articulatory lags for the F tone (left) and the R tone (right), as estimated from marginal predictions, i.e., using fixed effects only, from the mixed effect models described in text. Circles represent actual datapoints, solid lines represent mean and shaded regions 95% confidence intervals for mixed effect model predictions.

this question when discussing and comparing acoustic and articulatory modality and present explicit modeling of this phenomenon, Section 4.2.

If we combine of the variability and stability findings, our data suggest that articulatory timing in a lexical tone language involves a stable ***coordination of tone and vocalic gesture onsets***, while other timing relationships seem to be weaker. This finding has two theoretical implications. First, the primacy of onset-to-onset coordination is compatible with the hypothesis, put forth in the framework of Articulatory Phonology, that there exists an in-phase coupling between vowels and tones (Burroni, 2023a; Gao, 2008; Geissler, 2021; Hu, 2016; Karlin, 2014; Mücke et al., 2012, 2019; Shaw & Chen, 2019; Svensson Lundmark et al., 2021; Yi, 2017; Zhang et al., 2019); and also with recent versions of the PENTA model that also rely on the notion of a fully synchronized articulatory syllable, where vowels, tones, and consonants start at the same time (Xu, 2020; Xu et al., 2022). We note, however, that full synchronization of consonants and vowels is observed only at slower rates in our dataset, Section 3.1.1.

Second, compared to onset-to-onset timing, target-to-target timing seems to be both more variable and to be systematically affected by rate, no matter the exact landmarks involved. We hasten to caution that determining tonal targets is a complex issue both conceptually and practically, as we have discussed in Section 2.2. However, to the extent that the tonal targets we landmarked are representative of a target state for the vocal tract, they do not seem to be strictly timed to other events in ways that suggest speech production coordination or phonological control. This is because the tonal targets get more asynchronous with other acoustic and articulatory events as speakers speak more slowly, Section 3.1.1. This pattern is expected only if the tonal target landmark is not coordinated with other acoustic landmarks, as discussed in Section 1.5.

The lack of stable target-to-target timing is potentially problematic for speech production frameworks that invoke such a notion, like the Timing-Extrinsic three-component model (XT/3C) (Turk & Shattuck-Hufnagel, 2020a, 2020b). Models like XT/3C hypothesize that movements are controlled to achieve their targets synchronously (Elie et al., 2023; Turk & Shattuck-Hufnagel, 2020a, 2020b) in line with the approach to motor control presented in Tau theory (Lee, 2011). Similarly, the SAH also holds that tonal targets are achieved together with some other acoustic event, like a syllable offset. At first sight, the fact that we did not find evidence for target-to-target timing could be attributed to other confounds. For example, this framework relies on a different notion of articulatory gestures, where segmentation is implemented using positional extrema defined by zero-crossings in the velocity signal. As discussed in Section 2.3, such a landmarking approach was not feasible with our data. We also cautioned about the potential effects on temporal dynamics caused by the smoothing required by this approach. We would also add that, although different landmarking strategies may explain away differences in variability, we find it unlikely that different landmarking strategies would be able to explain highly systematic effects of rate perturbations which also suggests a lack of coordination.

In this respect, our data provides a highly controlled environment where we can assess the merits of onset-to-onset and target-to-target coordination, something that has rarely been presented in the literature (*cf.* Tilsen, 2022a). From this point of view, tonal timing in Thai could be interpreted as a case of empirical evidence for the ideas expressed by Tilsen (2022). Tilsen (2022) has argued that theories of direct control of the timing of target achievement in speech suffer from conceptual issues. Additionally, Tilsen (2022) also points out that empirical studies claiming to provide evidence for target-to-target timing do not, in fact, provide compelling evidence, as the measurement are often too different to be comparable. Tilsen (2022) also cites cases like the gestural reduction of *perfect memory* (Browman & Goldstein, 1990) to illustrate that speech targets can be reduced or lost altogether. The less vari-

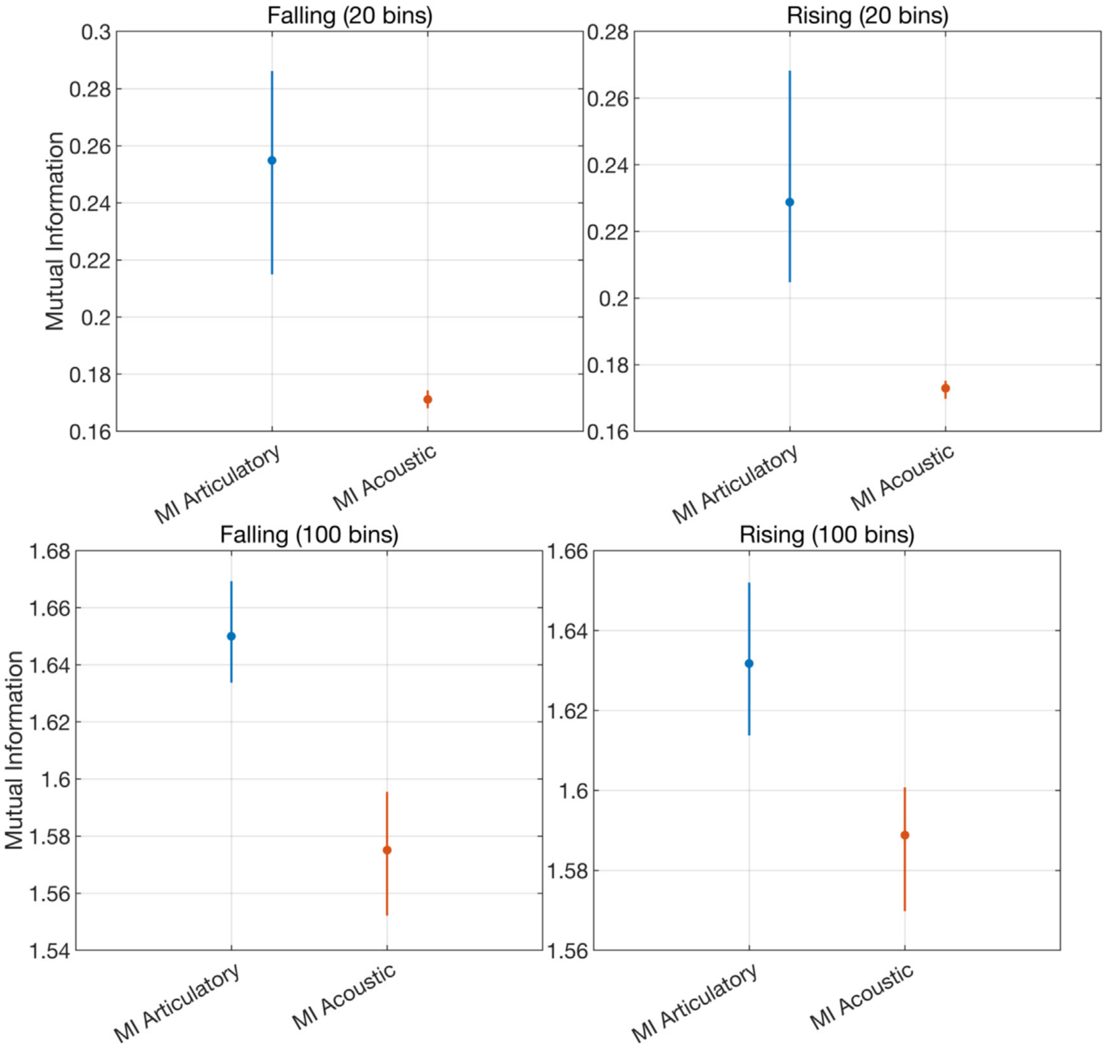

Fig. 23. Bootstrapped 95% confidence interval using 20 histogram bins (top) vs 100 histogram bins (bottom) for tone onset and vowel articulatory onset (MI articulatory) and tone onset and acoustic syllable/consonant onset (MI acoustic) confidence interval are generated from 100 bootstraps drawn with resampling from the original timing values.

able and more stable onset-to-onset timing in Thai tonal production can be added as another, perhaps more direct, piece of evidence supporting the notion of onset-to-onset over target-to-target timing. Much more empirical work is, however, needed before to assess the merits of different timing regimes in speech.

We also wish to emphasize that theoretical differences should not be exaggerated. Even a framework like the XT/3C model cannot really dispose of onset coordination. Despite a theoretical emphasis on target-to-target timing, Tau theory generally requires movement duration to be known for implementation. Thus, both an onset and a target are required to implement the closure gap model (Elie & Turk, 2023). Similarly, the SAH makes use of both onsets and targets, even though the emphasis is often placed on tonal targets. Conversely, a framework like AP consider gestural evolution to be intrinsic to gestures (Browman & Goldstein, 1990; Fowler, 1980; Saltzman & Munhall, 1989) and emphasizes onset-to-onset coordination and timing (Browman & Goldstein, 2000; Burroni, 2022, 2023b; Nam, 2007a; Nam & Saltzman, 2003; Saltzman & Byrd, 2000; Tilsen, 2017, 2018). In doing so, however, AP still needs additional mechanisms to ensure that gestures are suppressed correctly. In part, this is achieved via intergestural timing: new gestures “take over” tract variables from previous gestures and drive their suppression.

However, intergestural timing may not be enough to capture the highly regular temporal relationships across gesture of different tract variable, such as the timing of tones and supralaryngeal gestures associated with host syllables/words. For these reasons, a model may be required where not only tones are initiated together with articulatory gesture, like vowels, but where tones are also suppressed together with the supralaryngeal gestures they are timed to. In this respect, an extension to the framework of Articulatory Phonology, like Selection/Coordination theory (Burroni & Tilsen, 2022; Tilsen, 2016, 2018, 2022), may offer the right tools to implement onset-to-onset timing directly, while also indirectly capturing quasi-regular tone-vowel or tone-syllable timing patterns. This is realized thanks to the introduction of sets of gesture being selected and deselected together. Similar solutions have also been proposed in other frameworks, like the PENTA model (Xu et al., 2022), where tones, consonants and vowels start together and they are terminated or truncated as a new articulatory syllable starts (Xu, 2020). Rigorous computational modelling will need to be conducting to test the capability of these models to generate empirically observed patterns of timing between articulatory gestures and tones.

To conclude our discussion of timing and moving beyond specific models, there are some noteworthy broader implications of our results. The asymmetry between tonal onset and target timing that we have uncovered suggests that a conception of tone as an f0 contour time-locked to the acoustic boundaries, which underlies much phonological and acoustic work, misses important aspects of speech production. That tones may not be produced by speakers so that they are synchronized to acoustic boundaries, as predicted by segmental anchoring, was already suggested by well-known phenomena that disrupt this synchronization. These are phenomena like tonal coarticulation (Burroni, 2023a; Gandour et al., 1994; Potisuk et al., 1997; Xu & Liu, 2006) and peak delay (Xu, 2001), which can even lead to spillover of tonal contours over following syllables (Burroni, 2023a; Rose, 2014). The timing regimes we have investigated show that, at least in Thai, speakers prioritize onset alignment over target alignment, suggesting that initiation is directly controlled, but target achievement (and gestural suppression) may not be. This lack of direct control, coupled with the fact that tonal timing seems to reference articulatory events, could actually be one of the reasons why tones and acoustic boundaries line-up imperfectly. Namely, because the apparent acoustic regularities are the result of more regular underlying articulatory timing patterns that are indirectly present in the acoustic signal. It is exactly this disparity between acoustics and articulation that makes the acoustic patterns only quasi-regular, as some have hypothesized (Gao, 2008; Ladd, 2006). We now turn to the evidence that suggest greater articulatory than acoustic regularities in tonal timing.

4.2. Tonal timing modality: An articulatory model of Thai tonal timing and its implications for models of speech production

The results from our analyses of variability, stability, and informativity converge to support the hypothesis that Thai speakers more regularly time their production of lexical tones onsets to an articulatory event, the vocalic gesture onset, rather than an acoustic event, the syllable acoustic onset.

Based on analyses of variability, we found that the lag for which we observed the lowest variance is the lag between the tonal onset and the onset of the vocalic gesture, compared to the lag between the tonal onset and the acoustic syllable onset. This finding suggests that Thai speakers are more likely to be directly controlling lags between tonal onsets and vowel articulation rather than the syllable boundaries. Based on this finding, we would argue that regularities in acoustics stem from regularities in articulation. This is clear when we consider that lower variability of articulatory lags holds not only at the population, but also at the individual level, *cf*. the difference in spread, illustrated in Fig. 24, and the by-speaker bootstrapping analyses presented in 3.2.1.

This result is again compatible with the notion that the tone and vocalic gestures onsets are in-phase coupled, as put forth in Articulatory Phonology (Burroni, 2023a; Gao, 2008; Geissler, 2021; Hu, 2016; Karlin, 2014; Mücke et al., 2012, 2019; Shaw & Chen, 2019; Svensson Lundmark et al., 2021; Yi, 2017; Zhang et al., 2019); but also recent versions of the PENTA model that also rely on the notion of a fully synchronized articulatory syllable (Xu, 2020; Xu et al., 2022). The assumption underlying much gestural work, namely that "articulatory timing" is directly controlled, unlike "acoustic timing", now has a firmer empirical ground, at least for the case at hand, since we directly compared the two.

Regarding analyses of stability, linear mixed modelling of the effects of rate and tonal context showed that the tonal onset and vowel articulatory onset lag is mostly unaffected by tonal context and rate. The only exception is a small negative effect of the_H context for the F tone, estimated at −8 ms. However, with such a small size, this is not likely to be a robust or meaningful effect. On the other hand, the timing of the tonal onset with the acoustic syllable onset is robustly affected by duration of the utterance. More specifically, the acoustic lag value *negatively* correlates with utterance duration. In other words, the tone onset is closer and closer to the acoustic onset of the syllable, and it may eventually come to precede it, as utterance duration increases. To understand why this is the case and what the implication of these findings are, we first schematically describe this pattern in Fig. 25 below. We then proceed to a numerical simulation illustrating that the negative correlation between rate and tone onset / syllable acoustic onset lag is, in fact, predicted by a model specifying only underlying articulatory timing.

The top vs. bottom panels of Fig. 25 schematize changes in articulatory gestures profiles as duration of the relevant gestures increase, that is, when utterance duration increases and speech rate decreases. Comparing the top to the bottom panel, notice that articulatory gesture durations are longer, and when gestures are longer the constriction movement duration increases (a fact that is well known and empirically confirmed for Thai in Burroni, 2023a; as well as hypothesized in the pi-gesture model of Byrd & Saltzman, 2003). When this closure phase duration increases, the time to target attainment is delayed and affects the acoustic duration of the consonant closure. Recall also that we found that the tone-to-vowel articulatory onset lag is not affected by changes in rate (articulatory lag in Fig. 25 below). Piecing these two findings together, we can establish why the acoustic lag between the tonal onset and the acoustic onset of the consonant is negatively affected by rate. As the target attainment of a consonantal gesture gets delayed in time, together with its acoustic output, the tone onset remains time-locked to the vocalic gesture onset. If the C–V lag also shortens, as we have demonstrated in 3.1.1, the outcome of this process is a shortening of the acoustic lag in faster vs. slower rates, the acoustic lag in top vs. bottom panels, Fig. 25.

We stress again that this puzzling acoustic pattern is a consequence of the delay of the acoustic consonantal output compared to the vocalic gesture onset to which the tone is positively and stably time locked. In fact, we can demonstrate that all patterns follow from articulatory specifications and their mapping to acoustic outputs, without the need to encode acoustic timing patterns. We now proceed to a demonstration of these facts based on our data.

In the model below, we present stochastic simulations of local articulatory timing based on a set of assumed articulatory relationship and lag values estimated from data, following

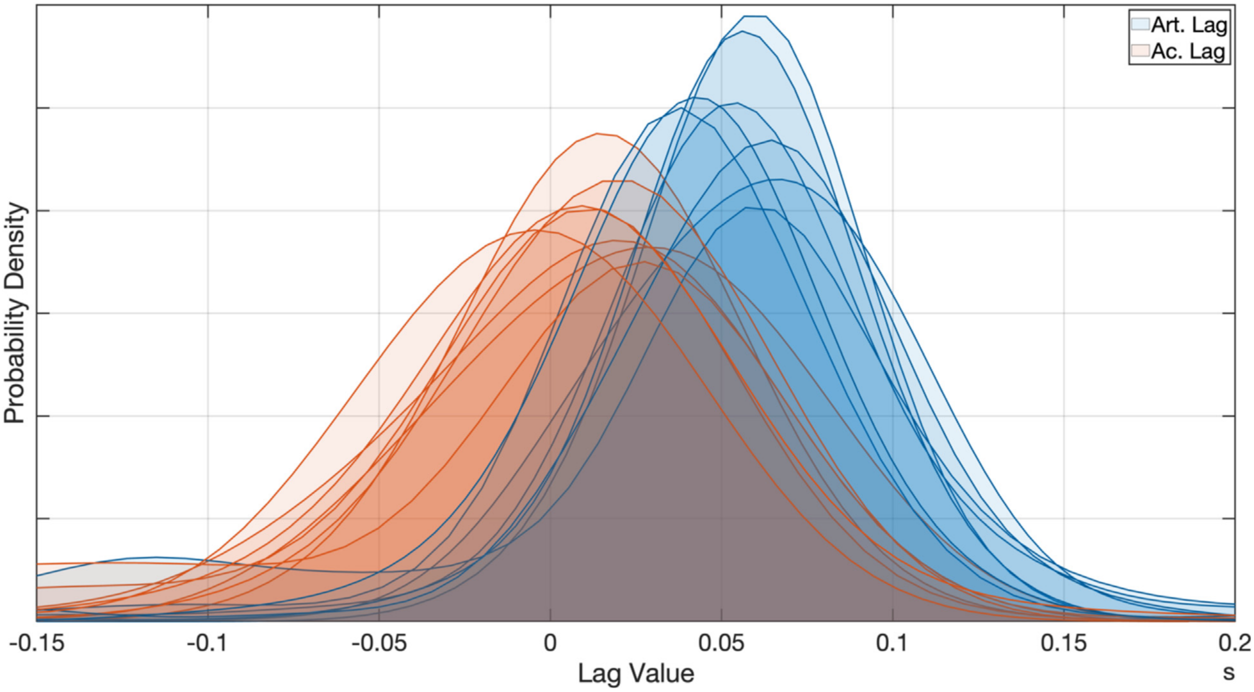

Fig. 24. By-speaker gaussian kernel density estimates for the articulatory and acoustic lag.

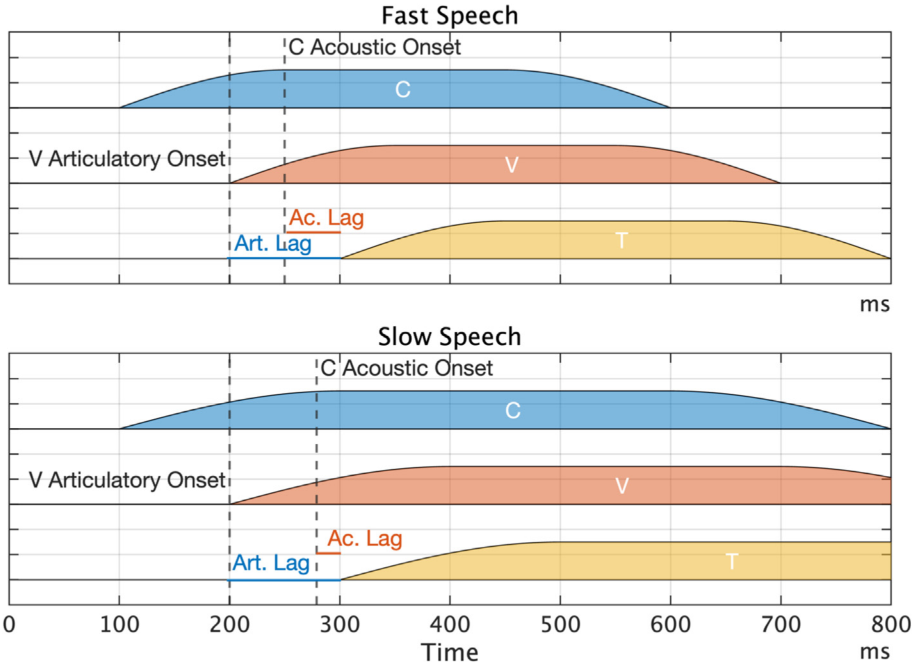

Fig. 25. Schematic illustration of changes in utterance duration and its effects on articulatory and acoustic lags. When gestural duration increases (bottom), as the articulatory vocalic gesture-tone onset lag remains constant across rates (Art. Lag), the acoustic consonant-tone onset lag, on the other hand, shortens (Ac. Lag), this is because the time between the vocalic onset gesture and the acoustic output of a consonant, the distance between dashed lines, increases as well.

previous articulatory work (Burroni, 2023a; Gafos et al., 2014; Shaw et al., 2009, 2011; Shaw & Gafos, 2015). Our interest is in modelling rate effects, specifically, showing that the negative correlation between utterance duration and the lag between the syllable acoustic onset and the tonal onset described above emerges from articulatory timing patterns, without dedicated acoustic specifications. Given that our interest lies in modelling speech rate effects, rather than the effects of different levels of variability (Gafos et al., 2014; Shaw et al., 2009, 2011; Shaw & Gafos, 2015), we opted for introducing stochasticity in the model via bootstrapping rather than in the form of Gaussian error terms. This is a slight departure from previous work, but it is a necessary one. If we had use gaussian error terms we would have needed to estimate noise levels separately for different durations/rate regimes, potentially, confounding the effects of rate and variability. Thus, we have chosen to only model rate explicitly and let the bootstrapping procedure introduce a stochastic variability component in the model. Notwithstanding this subtle conceptual difference, our model follows in all respects standard simulations of articulatory timing (Burroni, 2023a; Gafos et al., 2014; Shaw et al., 2009, 2011; Shaw & Gafos, 2015).

There are three gestures involved in the timing pattern of interest: a consonantal gesture (C), a vocalic gesture (V) and a tone gesture (T). The last one is estimated from f0 contours, but it is also potentially observable articulatorily from laryngeal adjustments. Moreover, an acoustic landmark is also involved: the syllable consonantal acoustic onset ($C_{Ac\ Ons}$). We model the temporal dynamics between these events using the following set of relationships. The consonantal gesture onset (C_{Ons}) can be considered a starting point for local articulatory timing relationships, we generate it from a random normal distribution with a mean of 0 and a standard deviation of 25. The value is chosen to give meaningful local relationships and follows previous work in this respect. The vocalic gesture onset (V_{Ons}) can then be generated from the C_{Ons} plus a $C_{Ons} – V_{Ons}$ lag that is stochastically estimated from data using bootstrapping of lag values via sampling with replacement from the empirically observed datapoints. Following the proposal put forth in this paper, the tonal onset (T_{Ons}) timing can then be stochastically generated from V_{Ons} and a $V_{Ons} – T_{Ons}$ lag also estimated using bootstrapping as described above. Finally, the consonantal acoustic onset ($C_{Ac.Ons}$) is estimated to be (nearly) time locked to the consonantal target (C_{Targ}) attainment, the moment where the lips have approached closure. $C_{Ac.Ons}$ is, thus, generated from C_{Ons} and closure formation duration, with the duration also estimated again using bootstrapping.

The articulatory relationship assumed in our models are summarized below, $*$ indicates bootstrapped values:

$$C_{Ons} = N(0, \sigma = 25)$$

$$V_{Ons} = C_{Ons} + C - V\,Lag*$$

$$T_{Ons} = V_{Ons} + T - V\,Lag*$$

$$C_{Ac.Ons} = C_{Ons} + C_{Ons} - C_{Targ}\,Lag*$$

Crucially, the simulations of articulatory timing are run twice with one hundred sets of lags generated each time. The first time, the model is run using relatively fast speech, by subsetting the data based on an utterance duration smaller than -1 z-score, and then a second time using relatively normal/slow speech, by subsetting the data based on an utterance duration greater than 0 z-score. We chose our thresholds for speech rate so that we have enough datapoints to conduct meaningful simulations of relatively fast and relatively normal or slow speech. Using different thresholds does not, however, qualitatively affect the model predictions.

Once we simulate timing patterns from the model described above, the model predicts exactly the observed negative correlation between utterance duration and the lag between the syllable acoustic onset and the tonal onset, Fig. 26. Note, specifically, how the models replicate information present in the data, such as a more synchronous C–V lag, 3.1.1, at slower rate, and, on the basis of purely articulatory relationships and their mapping to acoustics, the model also predicts the otherwise puzzling acoustic pattern observed in the data: a negative lag between the tone onset and the consonant/syllable acoustic onset is observed in slower speech, Fig. 26.

The upshot of the model presented above is that the not entirely obvious patterns of *shorter* and *more negative* acoustic lags between tones and consonantal acoustic onset at *longer* utterance durations is not only interpretable, but fully predicted under an articulatory model of tonal timing based on onset-to-onset timing of the tone to the vocalic gesture. This pattern follows from the tone being stably timed to a vowel, which slides earlier in time together with the tone in slower speech, and an acoustic syllable onset that is delayed, as the closure formation takes longer in slower speech.

Thus, several lines of evidence of lower variability, probed in different ways; greater stability under rate and tonal context manipulations; and higher informativity, probed via MI comparisons, all point to a more stable timing of tonal onset to the vocalic gesture onsets. Model simulation based on articulation also predicts some of the acoustic patterns present in the data, based on articulation. Thus, a broader implication of our findings is that, in the production of Thai lexical tones and their unfolding in real time, speakers seem to coordinate onsets of quasi-simultaneous articulatory events rather than their acoustic output. Such an interpretation of the evidence is compatible again with models that operate on coordination of the onsets of articulatory movements, like the coupled oscillator model of articulatory phonology (Burroni, 2022; Nam, 2007a, 2007c; Nam et al., 2009; Nam & Saltzman, 2003; Saltzman & Byrd, 2000; Tilsen, 2017, 2018, 2022) or recent versions of the PENTA model (Xu, 2020; Xu et al., 2022), over models which hypothesize that speakers' coordination is based on acoustic events, and, more specifically, targets of acoustic events such as the SAH (Arvaniti et al., 1998; Flemming & Cho, 2017; Ladd et al., 1999) and models of production and lexical access based on acoustic features (Stevens, 2002; Turk & Shattuck-Hufnagel, 2020a, 2020b).

Additionally, our findings also suggest that Thai lexical tones, as actions conducted in the production of speech, are fully integrated into articulatory plans that include oral gestures. In other words, Thai speakers produce lexical tones with regular timing regimes to surrounding articulatory gesture and the quasi-systematic relationship to acoustic landmarks is in part epiphenomenon to the former. This is showcased not only by lower variability, greater stability, and higher mutual information, but also by the fact that an articulatory model of tonal timing predicts otherwise puzzling facts in acoustic timing, particularly the larger negative lags between acoustic syllable onsets and tonal onsets at slower rates.

Findings like the one we have presented challenge, in our opinion, the notion that speech production can be understood purely in terms of an articulatory system yoked to produce certain acoustic outputs. There can be little doubt that the speech production system for the production of consonants, vowels, and tones is tuned throughout development using acoustic feedback to learn motoric routines. Yet, our data suggest that being a fluent speaker of a tonal language like Thai entails learning motor routines that coordinate laryngeal and supralaryngeal events to produce speech. Crucially, these timing patterns appear to be more precisely controlled than purely acoustic timing patterns.

This is an important consideration that should be kept in mind when equating lexical tones to f0 contours synchronized to acoustic syllables. That common approach is of course

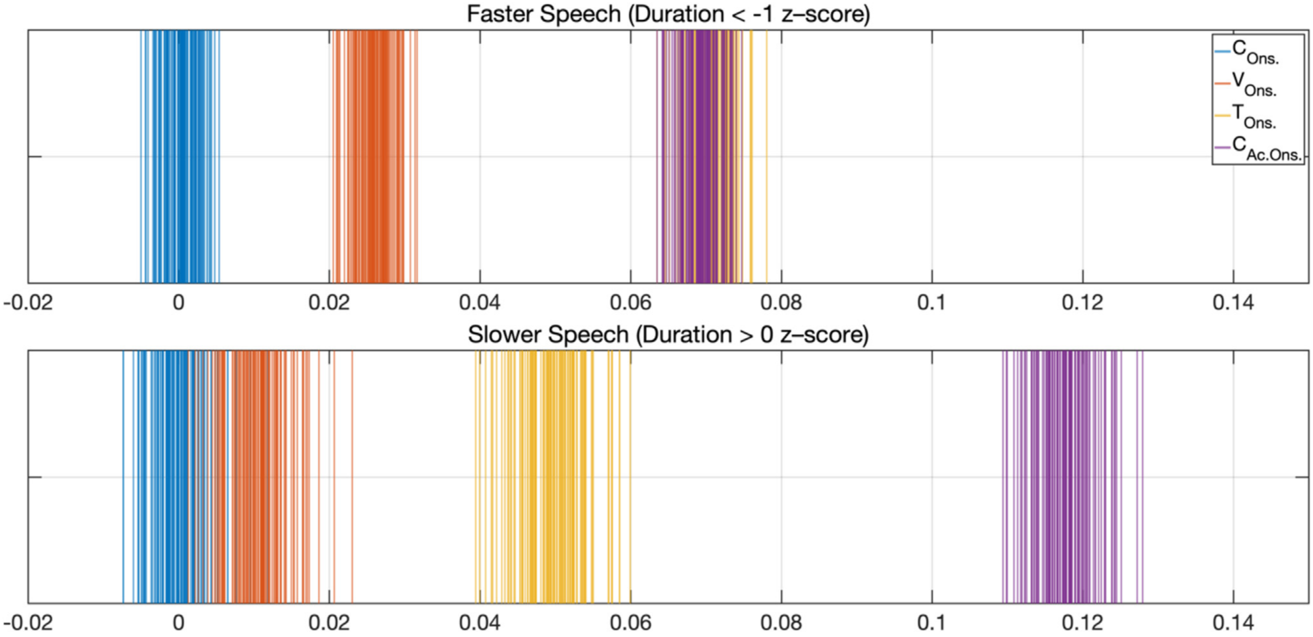

Fig. 26. Stochastic time model simulation of articulatory and acoustic timing relationships between consonant articulatory onset (C_{Ons}), vowel articulatory onset (V_{Ons}), tone onset (T_{Ons}) and consonantal acoustic onset (C_{Ons}) in faster (top) vs. normal/slower speech (bottom). Each lines represents one of a hundred datapoints for each lag.

practical in terms of phonetic measurements, but it is only indirect route to investigating the processes underlying tone production. We should thus be careful in attributing an ontological status to the synchronization of tones to syllables.

As a matter of fact, the relative timing patterns we have discovered in articulation does not entail temporal coincidence at all, contra what is assumed in virtually all models of tonal production. We, thus, turn to this final aspect before concluding.

4.3. A final wrinkle: Temporal coincidence vs. temporal locking of tones to supralaryngeal articulatory?

One interesting aspect of our results is that, even though the tone onset to vocalic gesture onset lag is the least variable, it is not an example of temporal coincidence of events. This finding may be unexpected because lags, at least relatively long ones, may be expected to display lower variability than lags with longer mean values (Schöner, 2002; Shaw et al., 2011; Wing & Kristofferson, 1973). However, this is not the case here, as the articulatory lag of interest has a lower variability than the usually assumed lag with the acoustic syllable onset; even though the latter is on average closer to temporal coincidence. We find it likely that the tonal coordination we observe is a stable coordination regime obeying a positive-lag timekeeping mechanism. That is to say, the initiation of tonal f0 movement is stably locked to sometime after the articulatory vowel onset. On the other hand, a lower central tendency value for the acoustic lag is perhaps not so meaningful, given the large fluctuations that encompass both highly negative and positive lags and lead to a high variability.

This uncovered positive lag of around ~50 ms that has been observed for the tonal onset compared to the vowel articulatory onset requires some further consideration. Why do speakers coordinate tonal onsets to vowel articulatory onsets but not in a temporally coincident manner?

In some models, like the coupled oscillator model of Articulatory Phonology, the fact that tones are initiated later than vocalic gestures is a consequence of the couplings existing not only between tones and vowels but also between tones and consonants. Tones are hypothesized to be coupled in-phase with a vocalic gesture and anti-phase with consonantal gestures in an onset-to-onset fashion. This coupling graph leads to a c-center coordination regime, where, roughly speaking, the vowel onset is initiated after the consonantal onset, produced, first, and before the tonal onset, produced last. This timing regime, where the vowel starts before the tone, is observed in Thai as well, as the positive V–T lag we have described illustrates. The Thai c-center is, however, asymmetric: the consonantal and the vocalic gesture onsets are much closer to each other than the vowel is to the tone onset, as the C–V lag reported in Section 3.1.1 shows and as we also illustrated in Fig. 26, where not much of a c-center is observed in slower speech.

The Thai facts are, thus, amenable to different interpretations that do not entail a c-center pattern. First of all, the c-center pattern for tone was posited on the basis of a long lag (~50 ms) between consonantal and vocalic gesture onsets in Mandarin, which was compared to that of English clusters, leading to the idea that tones act like a second onset consonant. Yet, the lag between consonantal and vocalic gestures in Thai is much smaller (~20 ms, Section 3.1.1). Such a short lag is reminiscent of the short lag observed for simplex CV syllables in English for which a split-gesture model, where the closure and release phase of a single consonant have been proposed to trigger c-center (Nam, 2007a, 2007b, 2007c; Tilsen, 2017). If the consonant–vowel lag is due to a single consonant, then no evidence for coupling of tones to consonants exists, as we have noted in our previous work (Burroni, 2023a). Additionally, recent work found comparable

C–V lags for Mandarin and English (Kramer et al., 2023). Moreover, speakers of diaspora Tibetan display a C–V lag, no matter whether they produce tone or not (Geissler, 2021). Similarly, in work on pitch accents in Swedish, the authors have also suggested that the differences in C–V lags may not reflect the distinction of tonal and non-tonal languages, but rather individual preferences in articulation (Svensson Lundmark et al., 2021). This has in fact been explicitly modelled for Thai using coupling strength differences between articulatory gesture to capture individual differences (Burroni, 2023a). In sum, all these facts taken together suggest that C-V lags may not be due to a competitive coupling of tone because sometimes these are absent.

If no C–T couplings are present, they cannot be the rationale for the long V–T lag observed in our data. Thus, what is the origin of this timing pattern?

One possibility is that lexical tones f0 movements are indeed near-synchronized by speakers to articulatory gestures, but this timing regime is not observed because f0 movement onsets are estimated from acoustic events. These acoustic changes could be delayed compared to the laryngeal articulatory adjustments necessary to produce them. For instance, consonantal articulatory onsets occur $\sim$ 68 ms before their acoustic onsets in the data presented in this work. If this reasoning is correct, the articulatory onset of the laryngeal adjustments resulting in f0 changes could be closer in time or even be synchronized to the onset of the vocalic gesture. If we take EMG estimates of a lag of around 60–100 ms between laryngeal muscle activity for f0 raising and lowering (Erickson, 2011), respectively, we would have, after all, a synchronization between vocalic gesture onset and the beginning of laryngeal muscle activity. Accordingly, we think that establishing whether tones exhibit synchronization or lack thereof is a question that will need to be solved with physiological studies that combine the tracking of supralaryngeal articulation with electromyographic work; or, possibly, via imaging of the entire vocal tract using real time MRI.

In the absence of direct evidence, we could consider additional methods, beyond landmarking, to probe tonal onsets. Several lines of work have suggested looking at divergences in articulatory time series to identify oral gestural initiation using minimal pairs (e.g., Boyce, 1990; Liu et al., 2022; Tilsen, 2020). This could be extended to tone as we quickly illustrate below. Recall that the tonal timing analyzed in this paper is from a minimal pair [mîː]/[mǐː]. Accordingly, we can attempt to determine when in time the tones produce a difference in tongue body and jaw vertical position (as it has been established for Mandarin, cf. Erickson et al., 2004; Hoole & Hu, 2004; Shaw et al., 2016). If we inspect (time-warped) 95% confidence intervals of these two articulatory dimensions over the acoustic boundaries of the word as a function of tone, we observe that they diverge over time. More specifically, the tongue and body vertical position has lower values for the R tone, quite early, around 20% for the tongue body movement, and from the very beginning for the jaw movement, Fig. 27.

The lower values for the tongue body and jaw vertical movement for the R tone, compared to the F tone, may be due to laryngeal adjustments conducive to f0 lowering. In particular, low f0 is due to vocal folds shortening and relaxation, which is accomplished via lowering of the hyoid-larynx complex. In turn, lowering of the hyoid complex leads to lowering of both the jaw and the tongue surface (Honda et al., 1999). If we take the lowering of tongue body and jaw to be a manifestation of tonal production, then, the production of tone and vowel may indeed be quasi-synchronous, especially in virtue of the effects on the jaw. Future work could be dedicated to test with dedicated experiments and more sophisticated modelling whether "coarticulatory effects" of laryngeal and supralaryngeal muscles may shed further light on the timing of oral vocalic and laryngeal tonal adjustments.

Maintaining for the moment the asynchrony of tonal and oral gestures, there are other hypotheses that we could consider in order to explain this pattern. A second more speculative hypothesis worth considering is that the muscles controlling laryngeal adjustments may have higher latencies compared to the muscles controlling oral movements of e.g., jaw, lips, and tongue. The physiological underpinning of these longer latencies may be the greater length of the recurrent laryngeal nerve, controlling most laryngeal muscles, compared to the hypoglossal and trigeminal nerves, controlling the jaw, tongue, and lips. For instance, bilateral adduction of vocal folds, time-locked to a longer-latency response (R2) of recurrent laryngeal nerve stimulation, occurs around 60 ms after external stimulation via electromyography (Yamashita et al., 1997). Of course, such longer latencies do not hold for the cricothyroid muscle which is innervated by the external superior laryngeal nerve and, thus, this muscle can respond faster but still with some latency (Lee et al., 2021; Ludlow et al., 1992). If laryngeal innervation is responsible for the observed lag, future work may try to probe whether tones with high onset values and low onset values display different lags, with the low tones have even longer lags than high tone due to their production being primarily implemented by muscles innervated by the recurrent laryngeal nerve. At any rate, the long latency in innervation could be another way to reconcile the notion that speakers may co-plan movements underlying f0 changes and oral articulatory gestures, but these are not synchronous due to intrinsic differences in their innervation response latencies.

A third hypothesis worth considering is tied to perceptual effects. A serendipitous consequence of the latency between vocalic gestures' articulatory onsets and acoustic f0 changes is that tones end up being produced during periods where the glottal specification for vowels allows for voicing. Thus, they are largely produced over perceptually salient rhymes where f0 is unperturbed. Tones could then be another case where "gestural" timing is constrained by perceptual recoverability (Browman & Goldstein, 2000). Whatever the exact mechanisms, several factors may conspire to produce the uncovered pattern whereby f0 movements onsets are stably timed to the vocalic gesture onset, but they look at as if they are quasi-synchronous to acoustic syllable boundaries.

4.4. Remaining issues, limitations, and future work

By exploring a variety of possible landmarks for tonal timing, as well as the most stable landmarks within each modality against each other, we have provided support for the notion that Thai speakers time their production of tones to that of an articulatory gesture, the vocalic gesture onset. We have also proposed an explicit model that can derive a puzzling aspect

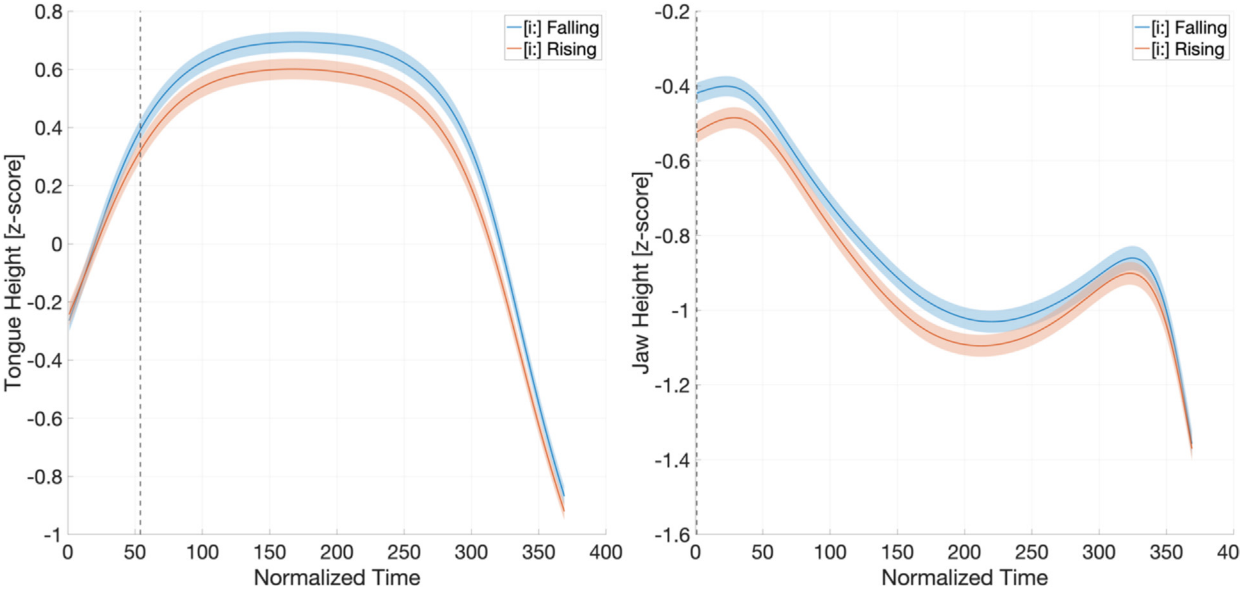

Fig. 27. Left: tongue body vertical movement over the vocalic gesture solid line represent mean values and shaded areas represent 2 standard errors. Right: jaw vertical movement over the vocalic gesture solid line represent mean values and shaded areas represent 2 standard errors.

of the acoustic results from purely articulatory control and have spelled out the implications of our findings for models of phonology and speech production. However, some important limitations of the present work should also be acknowledged.

First of all, it must be stressed that the differences in variability or informativity between acoustic and articulatory timing are admittedly modest in size. This is because articulatory gestures and their acoustic consequences have a causal relationship, which is captured in our model. This fact always makes identifying the control structure of speech timing difficult. This is also the reason why, in this paper, we took a multi-faceted approach to characterizing lexical tone timing by looking at variability combined with stability measures under manipulations in rate and tonal contexts, as well as informativity, estimated from mutual information analyses. We also supplemented these analyses with explicit modelling. Hopefully, a combined approach of this type will be useful for other areas of speech timing where identifying control structures has proven difficult.

A second area where further work is needed concerns the tonal onset and its lag with the vocalic onset. In this paper, the tonal onset was estimated by landmarking the f0 trajectory and deriving tonal onset measure from it. If the tonal onset is measured in this way, the picture that emerges is that of a positive-lag timekeeping mechanism between the vocalic gesture onset and the tonal gesture onset. At the same time, this question could be probed in more detail by trying to isolate tonal onset on the basis of its “coarticulatory effects” on relevant articulatory dimensions such as the vertical position of tongue body and jaw. Once these dimensions are inspected, however, we observe that the effect of tones on articulators starts very early, around 20% of the vowel, and synchronously with the vowel for the jaw. Accordingly, the tone and the vowel seem to be initiated quasi-synchronously in the articulatory signals, but the tonal onset is delayed when estimation is based on the acoustic f0 contours. Future work should try further probing the question of how acoustically estimated onset may be delayed compared to the tones as articulatory gestures. This means either obtaining direct data on laryngeal adjustments via EMG, real-time MRI (for vertical movement), or laryngoscopy (for changes in vocal fold postures). Alternatively, tonal onsets can be probed indirectly via their effect on other structure such as the jaw and the tongue; yet, in this respect, more groundwork needs to be done to estimate how consistent these effects are across different tones, vowels, individual speakers, and languages, as differences based on vowel quality associated with tones have been reported and the physiological bases of the effects questioned (Shaw et al., 2016).

An additional limitation involves defining tonal onsets in the context of contour tones. Ideally, when studying contour tones, both movement components of the contour should be examined. However, in practice, this was not feasible because the tonal context following the contour affects the estimates of velocity peaks. Consequently, this impacts the determination of tonal onsets and targets of the second component, which are based on thresholds of velocity peaks, as noted in Section 2.2. Despite these challenges, it is reasonable to expect that empirical measurements of the second component's onset would exhibit low variance in their timing lag relative to the target of the first component. This is because the two components cannot be randomly timed, given the necessity of maintaining consistent tonal shapes. A different experimental paradigm might provide a solution. For instance, contour tones produced in contexts where coarticulation is minimized—such as phrase-final positions—or in fixed tonal contexts, could allow for more precise estimates of velocity peaks, onsets, and targets. Such a setup could help investigate whether additional timing mechanisms, beyond the antiphase timing assumed between the two components, are necessary to explain the observed empirical patterns.

Third, we found that Thai speakers time the production of lexical tones to the onset of the vocalic gesture. This does not mean that the same will be observed in all tone languages. It is important to note that the timing regime we described could be a Thai–specific property, as the language has been argued to be sensitive to moraic structure for the purposes of tonal distributions, if not timing and alignment (Morén & Zsiga, 2006; Pittayaporn, 2018; Zsiga & Nitisaroj, 2007). More languages need to be examined to assess the generality of our findings. We must also add that some of the finer details of Thai tonal timing remain beyond the scope of the present work: our results were obtained based on a limited number of speaker and two tones, the F and the R tone; accordingly, future work should investigate more speaker and the remaining three tones of the language. We expect that the remaining tones are likely to conform to the timing patterns presented here. However, they may be less amenable to the landmarking procedure used in this paper, as described in Section 1.6, since they do not resemble articulatory trajectories. We also are not sure that the M and H tone onset can actually be landmarked using the velocity or acceleration signal, as these are mostly flat, with small fluctuations. Thus, landmarking these tones may require different surrounding tonal contexts and/or new methodologies, e.g., divergence across minimal pairs. In turn, these changes may introduce new confounds that will need to be accounted for, e.g., the necessity of landmarking tones and gestures differently or using a completely different strategy to get at tonal onset, for instance, divergence in minimal pair trajectories using GAMMs (Liu et al., 2022) or neural networks (Tilsen, 2020).

Fourth, it must also be noted that Thai tones are a good initial case study because we can assume that tonal articulation is representatively translated into f0. However, there are many cases where tones are also cued using voice quality. For instance, in Northern Vietnamese, Shanghai Chinese, or Mandarin Chinese (e.g., Brunelle et al., 2010; Chen & Gussenhoven, 2015; Kuang, 2017) among many other languages. In these cases, tonal implementation will need to be tracked more directly with imaging, e.g., via laryngoscopy to inspect vocal fold constriction, or by measuring simultaneously f0 and voice quality as acoustic trajectories that both need to be landmarked. The potential issue that certain voice quality cues may not be synchronized to f0 can of course arise, adding complexity to ascertaining tonal timing regimes. This will make for a more difficult but also a necessary investigation to increase our understanding of tonal timing within the wider spectrum of tonal languages that are currently spoken.

Fifth, in this paper, syllable structure was kept constant in the form of a singleton onset followed by a long high front vowel [i:], richer syllable structures with onset clusters, and the addition of coda consonants may be helpful to probe whether tonal timing is affected by consonants as well, thus, ultimately enabling us to assess whether tonal gesture are coupled not only to vocalic gestures but also to consonantal gestures as well.

Notwithstanding these considerations, the data we have presented offers suggestive evidence for the hypothesis that tones are tightly integrated with articulatory gestures and timed to them in an onset-to-onset fashion, specifically to the vocalic gesture. Thus, future work probing the details the articulatory timing of different tones, in different languages, speakers, and in a wider variety of phonological environments will represent a fruitful and fully justified enterprise that will help us go beyond a widely-held – yet in some ways problematic – assumption that tones are simply f0 contours synchronized to the acoustic boundaries of a syllable or a word.

5. Conclusion

In this paper, we have used acoustic and articulographic data to address the question of how speakers of a tonal language, Thai, time the production of tones. Specifically, we focused on two questions: what sorts of landmarks are most relevant for controlling tonal timing, and are those landmarks best represented in the articulatory or acoustic domain. These are, in our opinion, basic questions that are relevant for phonological theories and for models of speech production. Unfortunately, they have often been left unaddressed in previous work. To answer these questions a wide set of analyses was applied and developed, together with model simulations.

The data presented in this paper suggest that, generally, onset-to-onset timing is more stable in tonal production than target-to-target timing, both in articulation and acoustics. In reaching this conclusion, we also acknowledge that determining tonal targets is a difficult task and that further research in this area is required.

Additionally, by comparing the lags that are most stable in acoustics and articulation against each other, we observed that the lag with the lowest variability is the lag between tonal onset and an articulatory event, the vocalic gesture onset, rather than the consonant/syllable acoustic onset. Additionally, our data showed that the articulatory lags are stable under rate and most tonal context manipulations, while the acoustic lags are not. We also demonstrated that an aspect of acoustic timing – a puzzling shorter and more negative lag between tonal onset and acoustic syllable onset in slower speech where durations are longer – is actually predicted by a stochastic time model of tonal timing that only relies on articulatory temporal control. Finally, the strong dependency between tonal onset timing and vowel articulatory onset is also showcased by a higher mutual information observed between the two. This new analysis, borrowed from the coarticulation literature, was extended to the timing in this paper to try further gauging dependencies in the structure of speech timing across modalities.

Our findings on tonal timing are an important contribution to the more general debate on onset-to-onset vs. target-to-target timing, and they also suggest that tones and articulatory gestures are tightly integrated in speech production. This last finding calls into question the common view in which tones are f0 contours "perching" over segmental material and are synchronized to syllable acoustic boundaries.

Finally, we have discussed some unexpected issues that have emerged from our study. The onset-to-onset articulatory timing model of tone we have proposed seems to imply positive time-locking rather than temporal coincidence. However, this finding may, perhaps, be due to the necessity of estimating tonal onset from acoustic signals. When the effect of tone on articulatory gestures is examined, a pattern closer to true synchrony with the vocalic gesture emerges, especially on the basis of jaw position. An even clearer pattern may emerge from work imaging the larynx directly together with the vocal tract or from pairing articulatory imaging and EMG data.

To conclude, this work investigated a complex issue that is at the center of speech timing in many languages: the timing of lexical tone production. Based on a variety of findings, the conclusion we arrived at is that onset-to-onset gestural models of tones provide the picture that is most compatible with the evidence currently available. Nonetheless, many aspects of these models require further empirical investigation into more languages and phonological environments. Computational modelling of speech timing involving tones also needs to be further developed. It is our hope that the variety of methods applied in this paper and the evidence uncovered will encourage further physiological and motor control work on the issue of tonal production and timing, an area of speech production where our current state of knowledge remains, in many ways, still limited.

CRediT authorship contribution statement

Francesco Burroni: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sam Tilsen:** Writing – review & editing, Supervision, Software, Resources, Project administration, 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.

Acknowledgements

We are grateful to the members of the Institute for Phonetics and Speech Processing at the LMU Munich, Northwestern University, and the University of Potsdam, as well as the audiences and reviewers at ICPhS 2023 for feedback on previous versions of this work. We are also grateful to the Editor-in-Chief , Taehong Cho, and three anonymous reviewers whose comments and suggestions greatly improved the quality of this work. Finally, we are grateful to Bruce McKee for technical support and equipment-related assistance.

Appendices

Appendix A: **Number of tokens**

Table 9

Number of Falling tone tokens by subject and Tonal condition

	_M	_L	_F	_H	_R	Total
SP01	27	54	63	45	36	225
SP02	54	63	45	45	54	261
SP03	26	45	41	24	34	170
SP04	54	36	54	45	34	223
SP05	18	27	27	45	54	171
SP06	63	45	18	45	72	243
SP07	54	9	18	63	72	216
SP08	45	36	36	18	54	189
Total	341	315	302	330	410	1698

Table 10

Number of Rising tone tokens by subject and Tonal condition

	_M	_L	_F	_H	_R	Total
SP01	18	99	27	27	54	225
SP02	45	36	36	27	45	189
SP03	43	54	34	43	44	218
SP04	53	27	63	63	18	224
SP05	81	45	45	27	81	279
SP06	27	36	45	63	36	207
SP07	45	54	45	36	54	234
SP08	27	54	45	63	72	261
Total	339	405	340	349	404	1837

Appendix B:. Bootstrapped Brown-Forsythe test

The bootstrapped test is accomplished as follows in 8 steps (Lim & Loh, 1996):

1. We first calculate the test Brown-Forsythe test statistics, T, based on the entire dataset.
2. We initialize the number of iterations, R, where the bootstrapped test statistic, T^*, exceeds the pooled test statistic, T, as 0.
3. For each group, $i = 1 \cdots I$ for all observations $j = 1...J$, we compute the residuals $e_{ij} = x_{ij} - \widehat{\mu}_i$, where $\widehat{\mu}_i$ is the 20% trimmed mean of group i.
4. We then draw N bootstrap samples $e^*{}_{ij}$ with replacement from the set of randomized pooled residuals $S = \{e_{ij} : i = 1 \cdots I, j = 1...J\}$
5. We then set $x^*{}_{ij} = e^*{}_{ij}$. If, however, for some group, $i = 1 \cdots I$, the number of samples is smaller than 10, $n_i < 10$, we set $x^*{}_{ij} = (12/13)^{1/2}\left(e^*_{ij} + vU\right)$, where $v^2 = N^{-1}\sum\sum\left(x_{ij} - \bar{x}\right)$ and U is a uniform random variable with supports $\left(-\frac{1}{2}, \frac{1}{2}\right)$
6. We then calculate a Brown-Forsythe bootstrapped test statistics, T^*, based on the bootstrapped sample. If the bootstrapped statistics is greater than the pooled one, $T^* > T$, we increment R by 1.
7. We repeat steps 4-6 A times, where A is the number of iterations.
8. The bootstrapped p-value is given by $p^* = R/A$. The null hypothesis is rejected if the p-value is smaller than a significance threshold, α, $p^* < \alpha$.

The number of bootstrapped samples (n = 20) and bootstrapping iterations (iter = 1000) follows Lim and Loh (1996). The test statistic for the Brown-Forsythe tests is defined below:

$$T = \frac{(N - I)\sum_{i=1}^{I} N_i\left(\bar{Z}_{i\cdot} - \bar{Z}_{\cdot\cdot}\right)^2}{(k - 1)\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Z_{ij} - \bar{Z}_{i\cdot}\right)^2} \tag{3}$$

Where $\bar{Z}_{i\cdot}$ is the median of group, i, and $\bar{Z}_{\cdot\cdot}$ is the grand median.

Appendix C:. Final statistical models by variables

Falling Art. Lag	artLag∼T2+(1|SP)+(-1+CarrDur|SP)+(-1+T2|SP)
Falling Ac. Lag	acLag∼CarrDur+T2+(1|SP)+(-1+CarrDur|SP)+(-1+T2|SP)
Rising Art. Lag	artLag∼1+(1|SP)+(-1+CarrDur|SP)+(-1+T2|SP)
Rising Ac. Lag	acLag∼CarrDur+(1|SP)+(-1+CarrDur|SP)+(-1+T2|SP)

Appendix D. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wocn.2024.101389.

References

Abramson, A. S. (1962). *The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments*. Univ. https://books.google.co.th/books?id=QQMOAAAAYAAJ.

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics, 26*(1), 3–25.

Atterer, M., & Ladd, D. R. (2004). On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. *Journal of Phonetics, 32*(2), 177–197.

Bennett, R., Henderson, R., & Harvey, M. (2023). Vowel deletion as grammatically controlled gestural overlap in Uspanteko. *Language, 99*(3), 399–456.

Bombien, L., Mooshammer, C., & Hoole, P. (2013). Articulatory coordination in word-initial clusters of German. *Journal of Phonetics, 41*(6), 546–561.

Boos, D. D., & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics, 31*(1), 69–82.

Boyce, S. E. (1990). Coarticulatory organization for lip rounding in Turkish and English. *The Journal of the Acoustical Society of America, 88*(6), 2584–2595.

Browman, C. P. (1994). Lip aperture and consonant releases. *Phonological Structure and Phonetic Form. Papers in Laboratory Phonology, III*, 331–353.

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

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*(2), 201–251.

Browman, C. P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 341–376).

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*(3–4), 155–180.

Browman, C. P., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Les Cahiers de l'ICP. Bulletin de La Communication Parlée, 5*, 25–34.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology, 3*, 219–252.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*(346), 364–367.

Brunelle, M., Nguyên, D. D., & Nguyên, K. H. (2010). A laryngographic and laryngoscopic study of Northern Vietnamese tones. *Phonetica, 67*(3), 147–169.

Burroni, F. (2022). A split-gesture, competitive, coupled oscillator model of syllable structure predicts the emergence of edge gemination and degemination. *Proceedings of the Society for Computation in Linguistics, 5*(1), 11–22.

Burroni, F. (2023a). *Dynamics of F0 Planning and Production: Contextual and Rate Effects on Thai Tone Gestures* [Ph.D. Thesis]. Cornell University.

Burroni, F. (2023b). Lexical Tones are timed to Articulatory Gestures. *Proceedings of the 20th International Congress of Phonetic Sciences. 20th International Congress of Phonetic Sciences*.

Burroni, F., & Kirby, J. (2023). Speakers adaptively plan and execute f0 trajectories under rate changes: Evidence from Thai contour tones. In *Proc. The Second International Conference on Tone and Intonation* (pp. 49–53).

Burroni, F., & Kirby, J. (under review). *Effects of speaking rate on f0 trajectories: Evidence from Thai lexical tones*.

Burroni, F., Maspong, S., Hoole, P., & Kirby, J. (2024). *Jaw-dropping (and jaw-raising) tones: Tonal effects on vowel articulation in Thai and their implications for sound change*.

Burroni, F., & Tilsen, S. (2022). The online effect of clash is durational lengthening, not prominence shift: Evidence from Italian. *Journal of Phonetics, 91* 101124.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149–180. https://doi.org/10.1016/S0095-4470(02)00085-2.

Chen, W.-R., Chang, Y., & Iskarous, K. (2015). Vowel coarticulation: Landmark statistics measure vowel aggression. *The Journal of the Acoustical Society of America, 138* (2), 1221–1232.

Chen, Y., & Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association, 45*(3), 321–337.

Cho, H. (2010). *A weighted-constraint model of F0 movements* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/62312.

DiCanio, C. T., Amith, J., & García, R. (2014). In *May). The phonetics of moraic alignment in Yoloxóchitl Mixtec*, 4th Symposium on Tonal Aspects of Language (TAL). https://doi.org/10.13140/2.1.1254.0807.

D'Imperio, M., Espesser, R., Loevenbruck, H., Menezes, C., Nguyen, N., & Welby, P. (2007). Are tones aligned with articulatory events? Evidence from Italian and French. In J. Cole (Ed.), *Papers in Laboratory Phonology 9* (pp. 577–608). Mouton de Gruyter.

Drugman, T., & Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. *Proceedings of the Annual Conference of the International Speech Communication Association*, 1973–1976.

Elie, B., Lee, D. N., & Turk, A. (2023). Modeling trajectories of human speech articulators using general Tau theory. *Speech Communication, 151*, 24–38. https://doi.org/10.1016/j.specom.2023.04.004.

Elie, B., & Turk, A. (2023). Estimating virtual targets for lingual stop consonants using general Tau theory. *Interspeech, 2023*, 3083–3087 10.21437/Interspeech.2023-953.

Erickson, D. (1976). *A physiological analysis of the tones of Thai*. University of Connecticut.

Erickson, D. (2011). Thai Tones Revisited. *Journal of the Phonetic Society of Japan, 15* (2), 74–82.

Erickson, D. (2013). F0, EMG and Tonogensis in Thai. *名古屋学院大学論集 言語・文化篇, 24*(2), 1–13.

Erickson, D., Iwata, R., Endo, M., & Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. Walter de Gruyter.

Flemming, E., & Cho, H. (2017). The phonetic specification of contour tones: Evidence from the Mandarin rising tone*. *Phonology, 34*(1), 1–40. https://doi.org/10.1017/S0952675717000021.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8*(1), 113–133.

Fujisaki, H., Ohno, S., & Gu, W. (2004). Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their F0 contours. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, 269–337.

Gafos, A. I., Charlow, S., Shaw, J. A., & Hoole, P. (2014). Stochastic time analysis of syllable-referential intervals and simplex onsets. *Journal of Phonetics, 44*, 152–166. https://doi.org/10.1016/j.wocn.2013.11.007.

Gandour, J. (1974). Consonant types and tone in Siamese. *Journal of Phonetics, 2*(4), 337–350.

Gandour, J. (1976). Aspects of Thai tone. *University of California Working Papers in Phonetics, 33*, 20–22.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics, 22*(4), 477–492.

Gao, M. (2008). *Mandarin tones: An articulatory phonology account [Ph.D. Thesis]*. Yale University.

Gao, M. (2010). Articulatory Phonology (AP) and tonal alignment: Further testing of a proposed AP model of tone-to-segment alignment in Mandarin Chinese. The Fourth European Conference on Tone and Intonation (TIE4), Stockholm, Sweden.

Geissler, C. A. (2021). *Temporal Articulatory Stability, Phonological Variation, and Lexical Contrast Preservation in Diaspora Tibetan*. Yale University [PhD Thesis].

Goldsmith, J. A. (1976). *Autosegmental phonology [PhD Thesis]*. MIT Press London.

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 159–207.

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. *Frontiers in Phonetics and Speech Science, 239*, 249.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*(5), 408–422.

Hirose, H. (2010). Investigating the physiology of laryngeal structures. *The Handbook of Phonetic Sciences, 2*, 130–152.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: From global to local prosody. *Speech Prosody 2004, International Conference*.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). *Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control*. https://doi.org/10.1177/00238309990420040301.

Honorof, D. N., & Browman, C. P. (1995). The center or edge: How are consonant clusters organized with respect to the vowel. *Proceedings of the XIIIth International Congress of Phonetic Sciences, 3*, 552–555.

Hoole, P. (2006). *Experimental studies of laryngeal articulation*. Habilitation: University of Munich.

Hoole, P., Gobl, C., & Chasaide, A. (1999). Techniques for investigating laryngeal articulation. Section A: Investigation of the devoicing gesture. *Coarticulation: Theory, Data and Techniques*, 294–300.

Hoole, P., & Honda, K. (2011). Automaticity vs. Feature-enhancement in the control of segmental F0. *Where Do Phonological Features Come From*, 131–171.

Hoole, P., & Hu, F. (2004). Tone-vowel interaction in standard Chinese. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Hoole, P., Mooshammer, C., & Tillmann, H. G. (1994). Kinematic analysis of vowel production in German. *ICSLP*, 53–56.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*, 82.

Hu, F. (2016). Tones are not abstract autosegmentals. *Speech Prosody, 2016*, 302–306.

Ishihara, T. (2003). A phonological effect on tonal alignment in Tokyo Japanese. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 3–9).

Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America, 134*(2), 1271–1282.

Jones, J. A., & Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. *Journal of Phonetics, 30*(3), 303–320.

Karlin, R. P. (2014). *The articulatory TBU: Gestural coordination of lexical tone in Thai*. Cornell Working Papers in Phonetics and Phonology.

Karlin, R. P. (2018). *Towards an articulatory model of tone: A cross-linguistic investigation* [Ph.D. Thesis]. Cornell University.

Karlin, R. P. (2022). Expanding the gestural model of lexical tone: Evidence from two dialects of Serbian. *Laboratory Phonology, 13*(1).

Karlin, R. P., & Tilsen, S. (2014). The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai. *Proceedings of Meetings on Acoustics 168ASA*, *22*(1), 060006.

Kramer, B. M., Stern, M. C., Wang, Y., Liu, Y., & Shaw, J. A. (2023). Synchrony and Stability of Articulatory Landmarks in English and Mandarin CV Sequences. *Proceedings of the 20th International Congress of Phonetic Sciences*.

Krause, P. A., & Kawamoto, A. H. (2020). On the timing and coordination of articulatory movements: Historical perspectives and current theoretical challenges. *Language and Linguistics Compass, 14*(6) e12373.

Kroos, C. (1996). *Eingipflige und zweigipflige Vokale des Deutschen? Kinematische Analyse der Gespanntheitsopposition im Standarddeutschen*. Munich University: Institut Für Phonetik. Master's Thesis.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America, 142*(3), 1693–1706.

Kuberski, S. R., & Gafos, A. I. (2023). How thresholding in segmentation affects the regression performance of the linear model. *JASA Express Letters, 3*(9).

Ladd, D. R. (2004). Segmental anchoring of pitch movements: Autosegmental phonology or speech production? *LOT Occasional Series, 2*, 123–131.

Ladd, D. R. (2006). Segmental anchoring of pitch movements: Autosegmental association or gestural coordination? *Italian Journal of Linguistics, 18*(1), 19.

Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *The Journal of the Acoustical Society of America, 106*(3), 1543–1554.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America, 107*(5), 2685–2696.

Lee, A., & Mok, P. (2021). Lexical Tone. In *The Cambridge Handbook of Phonetics* (pp. 185–208). Cambridge University Press.

Lee, A., Prom-on, S., & Xu, Y. (2021). Pre-low raising in Cantonese and Thai: Effects of speech rate and vowel quantity. *The Journal of the Acoustical Society of America, 149*(1), 179–190. https://doi.org/10.1121/10.0002976.

Lee, D. N. (2011). Guiding movement by coupling taus. *Ecological Psychology.* https://www.tandfonline.com/doi/abs/10.1080/10407413.1998.9652683.

Lim, T.-S., & Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis, 22*(3), 287–301.

Liu, Z., Xu, Y., & Hsieh, F. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics, 90*. https://doi.org/10.1016/j.wocn.2021.101116 101116.

Lorenc, A., Żygis, M., Mik, Ł., Pape, D., & Sóskuthy, M. (2023). Articulatory and acoustic variation in Polish palatalised retroflexes compared with plain ones. *Journal of Phonetics, 96* 101181.

Lu, J., Li, Y., Zhao, Z., Liu, Y., Zhu, Y., Mao, Y., Wu, J., & Chang, E. F. (2023). Neural control of lexical tone production in human laryngeal motor cortex. *Nature Communications, 14*(1), 6917.

Ludlow, C. L., Van Pelt, F., & Koda, J. (1992). Characteristics of late responses to superior laryngeal nerve stimulation in humans. *Annals of Otology, Rhinology & Laryngology, 101*(2), 127–134.

Luksaneeyanawin, S. (1983). *Intonation in Thai*. University of Edinburgh [PhD Thesis].

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control, 14*(3), 380–407.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)*. John Wiley & Sons.

McGowan, R. S., & Saltzman, E. L. (1995). Incorporating aerodynamic and laryngealcomponents into task dynamics. *Journal of Phonetics, 23*(1–2), 255–269.

Moisik, S. R., & Gick, B. (2017). The quantal larynx: The stable regions of laryngeal biomechanics and implications for speech production. *Journal of Speech, Language, and Hearing Research, 60*(3), 540–560.

Morén, B., & Zsiga, E. (2006). The Lexical and Post-Lexical Phonology of Thai Tones*. *Natural Language & Linguistic Theory, 24*(1), 113–178. https://doi.org/10.1007/s11049-004-5454-y.

Mücke, D., Grice, M., Becker, J., & Hermes, A. (2009). Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *Journal of Phonetics, 37*(3), 321–338. https://doi.org/10.1016/j.wocn.2009.03.005.

Mücke, D., Hermes, A., & Tilsen, S. (2019). Strength and Structure: Coupling Tones with Oral Constriction Gestures. *INTERSPEECH*, 914–918.

Mücke, D., Hermes, A., & Tilsen, S. (2020). Incongruencies between phonological theory and phonetic measurement. *Phonology, 37*(1), 133–170.

Mücke, D., Nam, H., Hermes, A., & Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. In *Consonant clusters and structural complexity* (pp. 205–230). De Gruyter Mouton.

Nam, H. (2007a). *A gestural coupling model of syllable structure. Yale.*.

Nam, H. (2007b). Articulatory modeling of consonant release gesture. *International Congress on Phonetic Sciences XVI*, 625–628.

Nam, H. (2007c). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. *Laboratory Phonology, 9*, 483–506.

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. *Approaches to Phonological Complexity, 16*, 299–328.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. *Proceedings of the 15th International Congress of Phonetic Sciences*.

Niemann, H. (2016). *The Coordination of Pitch Accents with Articulatory Gestures: A Dynamical Approach*. University of Cologne.

Niemann, H., Mücke, D., Nam, H., Goldstein, L., & Grice, M. (2011). Tones as Gestures: The Case of Italian and German. *ICPhS*, 1486–1489.

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLoS Computational Biology, 15*(9) e1007321.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics, 25*(5), 382–407.

Petrone, C., & Ladd, R. (2007). Sentence-domain effects on tonal alignmenr in Italian. *Proceedings of the XVIth International Congress of Phonetic Sciences*, 1253–1256.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation [PhD Thesis]*. Massachusetts Institute of Technology.

Pittayaporn, P. (2018). Quantitative and qualitative restrictions on the distribution of lexical tones in Thai. *Topics in Theoretical Asian Linguistics: Studies in Honor of John B. Whitman, 250*, 371.

Potisuk, S., Gandour, J., & Harper, M. P. (1997). Contextual variations in trisyllabic sequences of Thai tones. *Phonetica, 54*(1), 22–42.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., & others. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, *CONF*.

Prieto, P., Mücke, D., Becker, J., & Grice, M. (2007). Coordination Patterns Between Pitch Movements And Oral Gestures In Catalan. *Proceedings of the 16th International Congress of Phonetic Sciences*. ICPhS 16th, Saarbrücken, Germany.

Prieto, P., & Torreira, F. (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics, 35*(4), 473–500.

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology, 12* (1).

Rose, P. (2014). Mr. White goes to Market-Running Speech and Citation Tones in a Southern Thai Bidialectal. *Proceedings of the 15th Australasian International Conference on Speech Science and Technology.* 15th Australasian International Conference on Speech Science and Technology, Christchurch.

Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science, 19*(4), 499–526.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333–382.

Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research, 126*, 289–306.

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition, 48*(1), 31–51.

Shaw, J., & Chen, W. (2019). Spatially conditioned speech timing: Evidence and implications. *Frontiers in Psychology, 10*, 2726.

Shaw, J., Chen, W., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research, 59*(6), S1566–S1574.

Shaw, J., & Gafos, A. (2015). Stochastic time models of syllable structure. *PloS One, 10* (5) e0124714.

Shaw, J., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: Evidence from patterns of temporal stability in articulation. *Phonology, 26*(1), 187–215. https://doi.org/10.1017/S0952675709001754.

Shaw, J., Gafos, A. I., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: A case study of Moroccan Arabic consonant clusters. *Phonology, 28*(3), 455–490.

Smith, C., Erickson, D., & Savariaux, C. (2019). Articulatory and acoustic correlates of prominence in French: Comparing L1 and L2 speakers. *Journal of Phonetics, 77* 100938.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*(4), 1872–1891. https://doi.org/10.1121/1.1458026.

Story, B. H. (2015). Mechanisms of Voice Production. In *The Handbook of Speech Production* (pp. 34–58). John Wiley & Sons Ltd..

Svensson Lundmark, M., Frid, J., Ambrazaitis, G., & Schötz, S. (2021). Word-initial consonant–vowel coordination in a lexical pitch-accent language. *Phonetica, 78*(5–6), 515–569. https://doi.org/10.1515/phon-2021-2014.

Tanaka, K., Kitajima, K., & Kataoka, H. (1997). Effects of transglottal pressure change on fundamental frequency of phonation: Preliminary evaluation of the effect of intraoral pressure change. *Folia Phoniatrica et Logopaedica, 49*(6), 300–307.

Tang, D. (2024). Using altered auditory feedback to study pitch compensation and adaptation in tonal language speakers. *Frontiers in Human Neuroscience, 18* 1364803.

Teeranon, P., & Rungrojsuwan, R. (2009). Change in the Standard Thai high tone: An acoustic study. *MANUSYA. Journal of Humanities, 12*(3), 34–44.

Thepboriruk, K. (2009). Bangkok Thai tones revisited. *Working Papers in Linguistics, 40* (5), 1–15.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics, 55*, 53–77.

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics, 64*, 34–50.

Tilsen, S. (2018). *Three mechanisms for modeling articulation: Selection, coordination, and intention*. Ithaca, NY: Cornell University.

Tilsen, S. (2020). Detecting anticipatory information in speech with signal chopping. *Journal of Phonetics, 82* 100996.

Tilsen, S. (2022). An informal logic of feedback-based temporal control. *Frontiers in Human Neuroscience, 16*.

Tilsen, S., & Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication, 150*, 73–97. https://doi.org/10.1016/j.specom.2023.05.006.

Tilsen, S., Zec, D., Bjorndahl, C., Butler, B., L'Esperance, M.-J., Fisher, A., Heimisdottir, L., Renwick, M., & Sanker, C. (2012). A cross-linguistic investigation of articulatory coordination in word-initial consonant clusters. *Cornell Working Papers in Phonetics and Phonology, 2012*, 51–81.

Titze, I. R. (2000). *Principles of Voice Production. National Center for Voice and Speech*.

Torres, C., & Fletcher, J. (2020). The alignment of F0 tonal targets under changes in speech rate in Drehu. *The Journal of the Acoustical Society of America, 147*(4), 2947–2958. https://doi.org/10.1121/10.0001006.

Turk, A., & Shattuck-Hufnagel, S. (2020a). *Speech timing: Implications for theories of phonology, speech production, and speech motor control* (Vol. 5) USA: Oxford University Press.

Turk, A., & Shattuck-Hufnagel, S. (2020b). Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production. *Frontiers in Psychology, 10*, 2952.

Wang, Y., Rodríguez de Gil, P., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Pham, T., Nguyen, D., & Romano, J. L. (2017). Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement, 77*(2), 305–329.

Weerathunge, H. R., Alzamendi, G. A., Cler, G. J., Guenther, F. H., Stepp, C. E., & Zañartu, M. (2022). LaDIVA: A neurocomputational model providing laryngeal motor control for speech acquisition and production. *PLoS Computational Biology, 18*(6) e1010159.

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics, 14*(1), 5–12.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 55*(4), 179–203.

Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica, 58*(1–2), 26–52.

Xu, Y. (2004). The PENTA model of speech melody: Transmitting multiple communicative functions in parallel. *Proceedings of from Sound to Sense, 50*, 91–96.

Xu, Y. (2009). Timing and coordination in tone and intonation—An articulatory-functional perspective. *Lingua, 119*(6), 906–927.

Xu, Y. (2020, March). *Syllable is a synchronization mechanism that makes human speech possible*. PsyArXiv. https://doi.org/10.31234/osf.io/9v4hr.

Xu, Y., & Lee, A. (2022). Tonal Processes Defined as Articulatory-based Contextual Tonal Variation. In C.-R. Huang, Y.-H. Lin, I.-H. Chen, & Y.-Y. Hsu (Eds.), *The Cambridge Handbook of Chinese Linguistics* (pp. 275–290). Cambridge University Press. https://doi.org/10.1017/9781108329019.016.

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics, 18*(1), 125.

Xu, Y., Prom-on, S., & Liu, F. (2022). *The PENTA model: Concepts, use and implications*. Search In: Prosodic Theory and Practice. The MIT Press.

Yamashita, T., Nash, E. A., Tanaka, Y., & Ludlow, C. L. (1997). Effects of stimulus intensity on laryngeal long latency responses in awake humans. *Otolaryngology—Head and Neck Surgery, 117*(5), 521–529.

Yi, H. (2014). A gestural account of Mandarin tone sandhi. *The Journal of the Acoustical Society of America, 136*(4), 2144.

Yi, H. (2017). *Lexical tone gestures* [PhD Thesis]. Cornell University.

Yi, H., & Tilsen, S. (2014). Gestural timing in Mandarin tone sandhi. *Proceedings of Meetings on Acoustics 168ASA, 22*(1), 060003.

Yip, M. (2002). *Tone*. Cambridge University Press.

Zhang, M., Geissler, C., & Shaw, J. (2019). Gestural representations of tone in Mandarin: Evidence from timing alternations. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 1803–1807.

Zsiga, E. (2012). Contrastive tone and its implementation. In *The Oxford Handbook of Laboratory Phonology* (pp. 196–207). Oxford University Press.

Zsiga, E., & Nitisaroj, R. (2007). Tone Features, Tone Perception, and Peak Alignment in Thai. *Language and Speech, 50*(3), 343–383.