LEXICAL TONES ARE TIMED TO ARTICULATORY GESTURES

Francesco Burroni

Institute of Phonetics and Speech Processing, Spoken Language Processing Group, LMU, Munich
francesco.burroni@phonetik.uni-muenchen.de

ABSTRACT

We investigate whether speakers more stably time f0 movements, associated with lexical tones, to articulatory gestures or to the acoustic outputs of articulatory gestures. Acoustic and articulatory data collected from eight Bangkok Thai speakers show that f0 movements are more stably timed to the articulatory onsets of vowels. F-tests show that the lag between f0 movements and vocalic gestures onsets has the lowest variance compared to the lag of f0 movements onsets and any other acoustic/articulatory landmark. Additionally, Gaussian Process Regression models trained with articulatory features outperform models trained with acoustic features in providing a generative model of f0 movements onset timing. Taken together these two lines of evidence suggest that speakers time f0 movements onsets to articulatory gestures rather than their acoustic consequences, like the acoustic onsets of syllables/vowels. Issues regarding surface f0 timing, articulation, and perception are also discussed to shed light on the uncovered patterns.

Keywords: lexical tone, timing, articulation, anchoring, articulatory phonology

1. INTRODUCTION

An open question in the phonetic literature is how speakers control the timing of f0 movements, associated with lexical tones and pitch accents, relative to the movements of supralaryngeal articulatory gestures and their acoustic consequences.

Research on tone languages holds that f0 movements are synchronized with syllables e.g., [1]–[4]. Yet, it is not clear whether "syllable" should be interpreted in terms of articulatory gestures affiliated with an articulatory syllable or their acoustic consequences associated with an acoustic syllable. Research on the timing of f0 peaks and valleys of pitch accents has uncovered regularities in their timing relative to anchoring sites in the acoustic segmental string e.g., [5]–[9]. These findings were interpreted as evidence that speakers time relevant f0 movements to acoustic landmarks, such as consonantal closures' or stressed vowels' acoustic onsets. This hypothesis was termed Segmental Anchoring Hypothesis (SAH). Further work on testing the *SAH*, however, showed that the timing of f0 movements to acoustic landmarks can be influenced by linguistic and paralinguistic factors, such as syllable structure, speech rate, and diatopic variation e.g., [8], [10]–[12]. These findings led several researchers to hypothesize that f0 movements may be anchored to articulatory rather than acoustic events e.g., [13]–[15]. Non-linearities in the articulatory to acoustic mapping could then be responsible for variability in f0 movements timing to acoustic landmarks. Investigations testing whether f0 movements are more stably aligned to acoustic or articulatory events, however, failed to provide conclusive evidence e.g., [13], [16].

Meanwhile, researchers investigating both acoustic and articulatory signals, especially from the perspective of gestural phonology [17], assumed that f0 movements associated with lexical tones and pitch accents can be treated on par with articulatory gestures and that these are timed to the supralaryngeal articulatory gestures associated with the production of consonants and vowels e.g., [18]–[24]. Crucially, the timing of f0 movements to supralaryngeal articulatory gestures was posited without assessing whether acoustic or articulatory landmarks provided more stable "anchoring" sites.

The main research question we investigate is the following: do speakers time f0 movements associated with lexical tones to the articulatory gestures underlying the production of consonants and vowels or to the acoustic consequences of said gestures, such as the acoustic onset of consonants/vowels/syllables? We can entertain two hypotheses. ***H1*****:** we expect speakers to time relevant f0 movements to articulatory gestures. ***H2***: we expect speakers to time relevant f0 movements to the acoustic consequences of articulatory gestures, e.g., the (acoustic) onsets of syllables or vowels. The predictions associated with ***H1*** are that, if speakers time f0 movements to articulatory gestures, the variance in lags between f0 movements onsets, associated with lexical tones, and articulatory gestures onsets, associated with consonants and vowels, should be the lowest; indicating that the two are produced with stable relative initiations. Additionally, knowledge of the time of initiation of the articulatory gestures should help predict the timing of the tonal f0 movements initiations. Conversely, the predictions associated with ***H2*** are that, if speakers time f0 movements to the acoustic consequences of articulatory gestures,

the variance in lags between f0 movements onsets, associated with lexical tones, and acoustic onsets of consonants/syllables and vowels should be the lowest. Additionally, knowledge of acoustic onsets of consonants/syllables and vowels should help predict the timing of the tonal f0 movement initiation. A production study with acoustic and articulatory data collected from eight Bangkok Thai (BKKT) speakers was conducted to test ***H1***/***H2*** and their predictions.

2. METHODOLOGY

Eight L1 speakers of Thai participated in the experiment. They did not disclose any speech or hearing impairment. All speakers were screened for nativeness in BKKT by a native speaker trained in phonetics. Participants produced BKKT Falling and Rising tones (F/R), w2, followed by all five tones of the language (Mid, Low, Falling, High, Rising), w3. The disyllabic targets were embedded in a carrier sentence with a fixed number of words and syllables; all carrier words are Mid-toned Table 1.

w1	w2	w3	w4	w5
dū:	mî:	mā:	bōn	dā:w
	mǐ:	mà:		dīn
		mâ:		lāŋ
		má:		
		mǎ:		

Table 1: Experimental materials

F/R in w2 were chosen because their f0 trajectory resembles an articulatory trajectory and can easily be landmarked. All five BKKT tones were used in w3 to introduce variation due to coarticulatory effects. The transition between [u] and [i] in the carrier was chosen to maximize tongue movement in the longitudinal direction and facilitate locating vowels onsets of the target word [mi:]. Participants were asked to produce the stimuli at one of five rates: very slow, slow, normal, fast, or very fast to introduce variation. W5 was varied at every trial to function as a distractor. In total participants produced 2 (F/R) × 5 (M/L/F/H/R) × 3 (w5) = 30 unique stimuli × 5 (rates) × 3 (repetitions) = 450, distributed in 9 blocks. Not all participants completed the entire experiment due to time constraints. Not all tokens for all participants could be analyzed due to errors, equipment malfunctions in articulatory data tracking, and a coding oversight. After screening all the data, a total of 3157 tokens were retained for analysis.

For the experiment, participants sat in a 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). Articulatory data sensors were adhered midsagittally on the lower and upper lip (LL, UL) vermilion border; on the lower right incisor to capture jaw movement (JAW). Two sensors were placed on the tongue, one posterior to the tongue apex, about 1 cm, to measure tongue tip (TT) movement, and one posterior to the TT sensor of approximately 6–7 cm to measure tongue body (TB) movement. Reference sensors were positioned on the nasion and left and right mastoid processes.

Speaker-specific monophone Hidden Markov Models were trained in Kaldi [25] and used to perform forced alignment by speaker, followed by manual checking and correction. F0 was extracted using the Sum of Residual Harmonic algorithm [26] 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 [100 400] Hz range. The raw f0 trajectories were cleaned of f0 "jumps" greater than 20 Hz, interpolated with spline interpolation, and smoothed using a moving median followed by a moving average filter.

Articulatory trajectories' missing values 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 cutoff. In this paper, we focus on the trajectories involved in the production of [m] and [i:] 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. The formation of [i:] was studied from the horizontal component of the tongue body (TB) movement. We focus on the horizontal component because the carrier sentence has a transition from [u:] to [i:] that is maximally differentiated on the horizontal plane. Tones were landmarked based on the f0 trajectory. We first located an inflection/midpoint, then proceeded to locate velocity extrema that precede and follow the inflection/midpoint. We then located positional extrema that precede and follow the first and second velocity extrema, respectively. Within a region spanning the first positional extremum and the first velocity extremum, we located the onset of the movement, defined as the first time point at which the movement velocity surpasses 20% of the maximum velocity. 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. An example of landmarked articulatory and (acoustic) f0 trajectories is presented in Figure 1.

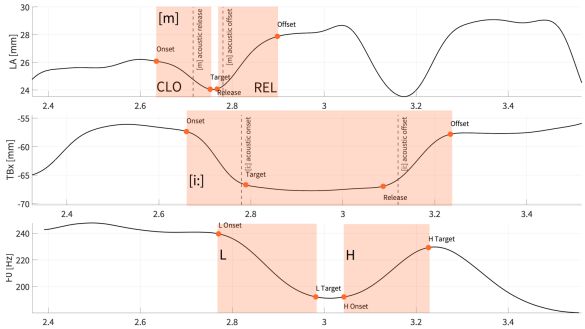

Figure 1: Example of LA, TBx, and F0 landmarking

All acoustic landmarks were obtained from forced alignment boundaries. All landmarking was manually checked and corrected. From the landmarks, second-order landmarks were derived. The C-T center was calculated as the midpoint between the closure (CLO) and tonal (T) onsets. Two versions of the C center (1 and 2) were also calculated. C center 1 is the midpoint of CLO onset and offset, while C center 2 is the midpoint of closure and release. Pairwise lags among the onset of different acoustic and articulatory gestures were calculated as illustrated in Figure 2.

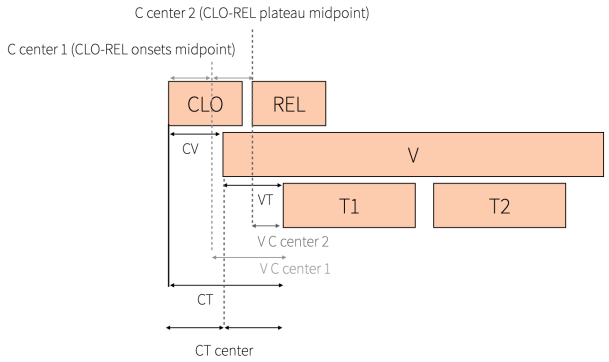

Figure 2: Lags and second-order landmarks

We examined the means and variances of lags between the tonal onset and all other articulatory and acoustic landmarks. F-tests for variance were used to assess differences in variance, indicating different stability of f0 movements onset to the other landmark forming the lag. Gaussian Process Regression (GPR) models were trained to predict the onset of f0 movements based on three sets of features:

1. Control Features: (1-5): (1) Subject, (2) Block, (3) Utterance Duration, (4) Tone, (5) Following Tone
2. Acoustic Features: (6-9) (6) C onset, (7) C Offset, (8) V Onset, (9) V Offset
3. Articulatory Features: (10-14): (10) LA movement onset, (11) LA mov.t target, (12) TB (TB) mov. onset, (13) TB mov. target, (14) C center 1

The control GPR model was used to obtain a root mean squared error (RMSE) baseline based on speakers' nonstationary behavior [27]. Models trained with acoustic and/or articulatory features were used to compare how acoustic and articulatory information could improve the prediction of f0 movements onsets. The accuracy reported in the paper is from tenfold cross-validation.

3. RESULTS

3.1 Variability analysis

The main finding is that the landmark with the lowest variability compared to the f0 movement onset is the articulatory onset of vowels, with a lag standard deviation estimated at 55.8 ms. The second lowest variance is with the acoustic onset of the consonant/syllable, with a standard deviation estimated at 62 ms. An F-test for equal variance ($F = .81$, df_1, $df_2=3156$, $p < .0001$) reveals that the lag between the vowel articulatory onset and the tonal f0 movement has lower variance than the lag with the consonant/syllable acoustic onset.

Variance patterns do not align with the mean patterns. The onsets of tonal f0 movements are more stably timed to the onsets of the vocalic gestures, however, the vocalic gesture initiation *precedes* the f0 movement initiation by 48.4 ms on average. On the other hand, the acoustic onset of the consonant syllable precedes the f0 movement by 2.4 ms on average. To sum up, the onset of f0 movements associated with lexical tones is more stably timed to the articulatory onset of vowels, but it is closest in time to the consonant/syllable acoustic onset. These results are summarized in Figure 3.

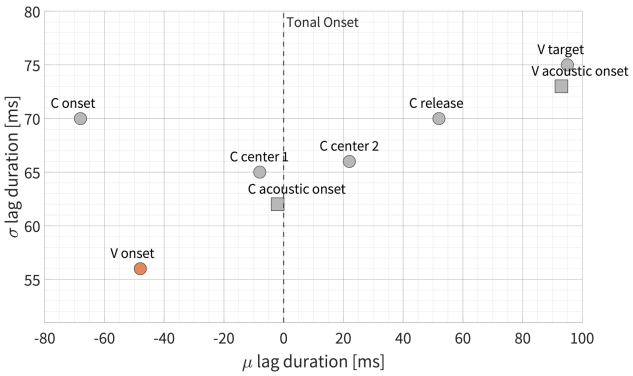

Figure 3: Plot of lag std.dev. (σ) *vs*. lag mean (μ) duration

3.2 Gaussian Process Regression analysis

The main findings are that adding acoustic and/or articulatory information reduces the RMSE in predicting the f0 movement onset by around 5.5

times. Second, if we compare the addition of acoustic *vs*. articulatory information to the control feature set, we observe that articulatory information yields better predictions, in line with the variability analysis. Third, the combination of both acoustic and articulatory information results in a ~1 ms improvement performance of the GPR model. This last finding suggests that both types of information play a role in determining the timing of f0, yet, when articulatory information is available, the addition of acoustic information does not lead to a huge improvement and vice versa. The results of the Gaussian Process are summarized in Figure 4.

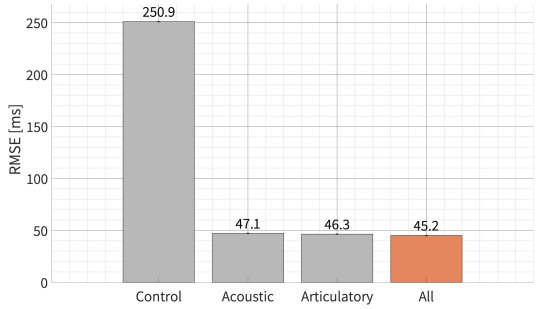

Figure 4: GPRs RMSE for different training features

4. DISCUSSION

Both the variability analysis and the GPR models suggest that f0 movements seem to be timed to articulatory events rather than their acoustic consequences. This is in line with the lower variability observed between f0 movements onsets and the articulatory onset of vocalic gestures rather than any other acoustic/articulatory landmark. The GPR models show that predictions of f0 movements onsets achieve better results based on the articulatory, rather than acoustic, landmarks with a lower RMSE of around ~1 ms. Additionally, the combination of both acoustic and articulatory information only leads to a reduction in RMSE of around ~1 ms, despite a larger number of parameters. We can, thus, conclude that articulatory timing of f0 movements onsets is more in line with the data.

F0 movements are, however, closer in time, with a mean lag of ~2 ms, to the acoustic onset of syllables. The fact that, on average, f0 movements onsets are closest to the acoustic onset of syllables could be the reason why previous work has considered acoustic boundaries as possible anchoring sites for f0 movements. However, the near-synchronization between f0 movements onsets and acoustic landmarks goes hand in hand with a higher variability compared to articulatory landmarks. This is an additional problem for a hypothesized timing to acoustic landmarks since lower means in speech timing are expected to correlate with *lower*, not higher, standard deviations e.g., [28]. We can hypothesize that f0 movements are timed, but not synchronized, to articulatory gestures because f0 movements onsets are acoustic, however, these acoustic changes may be delayed compared to the laryngeal articulatory adjustments necessary to produce them. For instance, consonantal articulatory onsets occur ~68 ms before their acoustic onsets in the data presented. If this reasoning is correct, the articulatory onset of the laryngeal adjustments resulting in f0 movements could be closer in time or synchronized, to the onset of the vocalic gesture.

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 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 [29]. If this is correct, 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 movements onsets is that tones are produced synchronized to acoustic syllables, thus, largely, over perceptually salient rhymes where f0 is unperturbed. Tones could then be another case where "gestural" timing is constrained by perceptual recoverability [30]. Whatever the exact mechanisms, several factors may be responsible for the uncovered pattern whereby f0 movements onsets are stably timed to articulatory gestures, but, at the surface, they look near-synchronous to acoustic syllable boundaries.

A final note of caution is in place. Articulatory gestures and their acoustic consequences have a causal relationship that makes identifying the control structure of speech timing difficult: the standard deviation differences and model accuracy differences we reported are admittedly modest in size. However, they do suggest that speakers time f0 movements to the movements of articulators producing consonants and vowels, not to their acoustic consequences. More work on tonal languages beyond Thai and on pitch accents languages will help assess the generality and merits of the findings presented in this paper, as well as the possible cross-linguistic instantiations of f0 movements timing.

5. REFERENCES

[1] Y. Xu and S. Prom-on, "Degrees of freedom in prosody modeling," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Springer, 2015, pp. 19–34.

[2] Y. Xu, F. Liu, and others, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 125, 2006.

[3] Y. Xu, "Syllable is a synchronization mechanism that makes human speech possible." PsyArXiv, Mar. 2020. doi: 10.31234/osf.io/9v4hr.

[4] Y. Xu and A. Lee, "Tonal Processes Defined as Articulatory-based Contextual Tonal Variation," in *The Cambridge Handbook of Chinese Linguistics*, C.-R. Huang, Y.-H. Lin, I.-H. Chen, and Y.-Y. Hsu, Eds. Cambridge University Press, 2022, pp. 275–290. doi: 10.1017/9781108329019.016.

[5] A. Arvaniti, D. R. Ladd, and I. Mennen, "Stability of tonal alignment: the case of Greek prenuclear accents," *Journal of Phonetics*, vol. 26, no. 1, pp. 3–25, 1998.

[6] D. R. Ladd, D. Faulkner, H. Faulkner, and A. Schepman, "Constant 'segmental anchoring' of F0 movements under changes in speech rate," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1543–1554, 1999.

[7] D. R. Ladd, I. Mennen, and A. Schepman, "Phonological conditioning of peak alignment in rising pitch accents in Dutch," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2685–2696, 2000.

[8] M. Atterer and D. R. Ladd, "On the phonetics and phonology of 'segmental anchoring' of F0: evidence from German," *Journal of Phonetics*, vol. 32, no. 2, pp. 177–197, 2004.

[9] T. Ishihara, "Tonal alignment in Tokyo Japanese," The University of Edinburgh, Edinburgh, 2006.

[10] C. Petrone and R. Ladd, "Sentence-domain effects on tonal alignmenr in Italian," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, 2007, pp. 1253–1256.

[11] P. Welby and H. Løevenbruck, "Anchored down in Anchorage: Syllable structure and segmental anchoring in French," *Italian Journal of Linguistics/Rivista di linguistica*, vol. 18, pp. 74–124, 2006.

[12] P. Prieto and F. Torreira, "The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish," *Journal of Phonetics*, vol. 35, no. 4, pp. 473–500, 2007.

[13] M. D'Imperio, R. Espesser, H. Loevenbruck, C. Menezes, N. Nguyen, and P. Welby, "Are tones aligned to articulatory events? Evidence from Italian and French," *Laboratory Phonology*, vol. 9, Jan. 2007.

[14] D. R. Ladd, "Segmental anchoring of pitch movements: Autosegmental association or gestural coordination?," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 19, 2006.

[15] M. D'Imperio, N. Nguyen, and K. G. Munhall, "An Articulatory Hypothesis for the Alignment of Tonal Targets in Italian," in *Proceedings of the 15th international congress of phonetic sciences*, 2003, pp. 253–256.

[16] D. Mücke, M. Grice, J. Becker, and A. Hermes, "Sources of variation in tonal alignment: Evidence from acoustic and kinematic data," *Journal of Phonetics*, vol. 37, no. 3, pp. 321–338, 2009, doi: https://doi.org/10.1016/j.wocn.2009.03.005.

[17] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, 1992.

[18] M. Gao, "Mandarin tones: An articulatory phonology account," Ph.D. Thesis, Yale University, 2008.

[19] D. Mücke, H. Nam, A. Hermes, and L. Goldstein, "Coupling of tone and constriction gestures in pitch accents," *Consonant clusters and structural complexity*, vol. 26, p. 205, 2012.

[20] D. Mücke, A. Hermes, and S. Tilsen, "Strength and Structure: Coupling Tones with Oral Constriction Gestures.," in *INTERSPEECH*, 2019, pp. 914–918.

[21] R. P. Karlin and S. Tilsen, "The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai," in *Proceedings of Meetings on Acoustics 168ASA*, 2014, vol. 22, no. 1, p. 060006.

[22] H. Yi, "Lexical tone gestures," PhD Thesis, Cornell University, 2017.

[23] H. Niemann, D. Mücke, H. Nam, L. Goldstein, and M. Grice, "Tones as Gestures: The Case of Italian and German.," in *ICPhS*, 2011, pp. 1486–1489.

[24] R. P. Karlin, "Towards an articulatory model of tone: a cross-linguistic investigation," Ph.D. Thesis, Cornell University, 2018.

[25] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011, no. CONF.

[26] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of the Annual Conference of the International Speech Communication Association*, Florence, 2011, pp. 1973–1976.

[27] S. Tilsen, "Structured nonstationarity in articulatory timing.," in *ICPhS*, 2015.

[28] J. A. Shaw, A. I. Gafos, P. Hoole, and C. Zeroual, "Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters," *Phonology*, vol. 28, no. 3, pp. 455–490, 2011.

[29] T. Yamashita, E. A. Nash, Y. Tanaka, and C. L. Ludlow, "Effects of stimulus intensity on laryngeal long latency responses in awake humans," *Otolaryngology—Head and Neck Surgery*, vol. 117, no. 5, pp. 521–529, 1997.

[30] C. P. Browman and L. Goldstein, "Competing constraints on intergestural coordination and self-organization of phonological structures," *Les Cahiers de l'ICP. Bulletin de la communication parlée*, no. 5, pp. 25–34, 2000.