TIME DELAYS IN TONE PRODUCTION: A COMPUTATIONAL STUDY OF THAI TONES

Khantaphon Chaiyo [1], Yi Xu [2] and Santitham Prom-on [1,2]

Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand [1]
Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom [2]
khantaphon.c@mail.kmutt.ac.th, yi.xu@ucl.ac.uk, santitham.pro@mail.kmutt.ac.th

ABSTRACT

Different speech rates produce different effects in syllable shortening or lengthening. These effects are intertwined with the convergence of fundamental frequency (F_0) movements in the target approximation (TA) model. This paper presents a study of the effects of speech rates as time delays in tone production that manifest in the variations of surface F_0 contours. In TA, the time delay can be defined as an event that the actual implementation of the pitch target is delayed in order to accommodate the lengthening of the syllable. Results indicate apparent effects of speech rates that can be explained as time delays in the TA model. They also reveal different time delay patterns comparing between normal and weak tonal elements. This suggests the potential mechanism of timing control of tone production and thus allows us to generalize the knowledge toward a general speech production framework.

Keywords: time delays, target approximation, speech rates, syllable lengthening, Thai tones

1. INTRODUCTION

Variations of fundamental frequency (F_0) in speech encode multiple-level of information, including communicative meanings, expressions and emotions [1-4]. Segmental variations occur from the distinct articulatory control in every pronunciation [5,6]. One of the variabilities that is critical to modeling speech production is the effect of the speaking rate on F_0 movement. A number of studies have reported the effects of this variation in many languages, including English [7-8], Mandarin Chinese [9, 10] and Thai [11]. It is also closely related to syllable lengthening and shortening [12-13].

While speaking rate has been known to affect the F_0 movement in normal and fast speech, the exact nature of the effect is still unclear when the speech rate become slower, causing the lenghening or delay in articulation. Segmental anchoring of F_0 movements offers a plausible explanation of the effect of speech rate on the F_0 dynamics [14-17], but it provides no inner mechanism of the delay. In this hypothesis, the rise and fall patterns of F_0 movement are anchored to specific segmental locations. Thus, as speaking rate increases or decreases, the anchoring points would become closer or farther apart.

While the segmental anchoring hypothesis focuses on the alignment of the specific F_0 pattern to the segmental points, Target Approximation (TA) model offers an explanation of how F_0 movements can be represented as the surface response to successive pitch target implementation [18-20]. In TA, a pitch target represents an underlying goal of the F_0 movement and F_0 is a result of the implementation of the pitch targets through a third-order critically damped linear system [18]. The TA model can represent F_0 quite accurately in speech with various prosodic variations [19-20]. However, the model does not include an explicit control of the F_0 movement when different speech rate is used, especially when it is excessively slowed down.

In the present study, we explores the variability pattern and associations that a time delay action may cause the changes in F_0 signals in relatively long syllables. The long syllables may occur either in a normal lengthened pronunciation or in a deliberately slow speech. This paper aims to study the effects of slowing down speaking rate on F_0 movement and test if there are specific patterns in the F_0 contours that are explainable by a delay action in target approximation. This will allow us to understand more about the manner in which speaking rate affects F_0 movement.

2. METHOD

2.1 Dataset Design

The dataset was designed to elicit variations in tones, vowels and speaking rates at three different paces (normal, slow, and very slow). The following four-syllable sentences were recorded.

*Pud2 wa2 __**X**__ krab3/ka1*
*[Say the word __**X**__]*

where __***X***__ is the targeted monosyllabic word that can be /ma/, /mi/ or /mu/, with one of the five Thai tones. In total, there are 15 lexical combinations created from two conditions

- Vowel: /a/, /i/, or /u/.
- Lexical Tone: Mid (M, Tone 0), Low (L, Tone
- 1), Falling (F, Tone 2), High (H, Tone 3), or Rising (R, Tone 4).

The first and second syllables create a context that will put the emphasis on the third syllable as the keyword. The fourth syllable is the natural sentence ending word. The sentence ending word differ based on speaker's gender.

Each sentence was spoken with three different speaking rates: Normal, Slow A, and Slow B. For the normal speaking rate, speakers were asked to keep speaking rate equivalent to daily conversation speed. For Slow A, speakers were asked to speak slower than normal. And for Slow B, speakers were asked to speak slower than the slowA speed.

2.2 Recording

Eight native Thai speakers, 5 males and 3 females, participated in the study. They have no self-reported speech and hearing disorders. The subjects were King Mongkut's University of Technology Thonburi students. Sound recording was made with 44.1kHz sampling rate and 32-bit floating point resolution. Speakers were asked verbally by experimenter to speak 15 sentences which were arranged in random order. There are 3 utterances for each speech rate and sentence combination. A total of 135 utterances per speaker were obtained, resulting in a corpus size of 1080 utterances.

Visual inspection was made on the extracted F_0 contour so that frequent-occuring features in F_0 contour can be measured.

2.3 Measuring timing features

Syllable boundary marking and time-normalized F_0 contour extraction were made using ProsodyPro [21] that runs with Praat program [22]. F_0 contours of the third syllable in each sentence were analyzed and profiled as shown in Figure 1. The timing of specific F_0 dynamic events were measured and compared between different factors. It should be noted that some measurements were applicable only to specific types of tones. For example, the measurements of F_0 valley and peak locations are valid only for dynamic R and F tones, respectively. Also, the numeric subscriptions (0 and 1) in each measurement indicates whether the time measurement was done with respect to syllable onset or offset. This is to see in detail the delay that may occurs at different speaking rates.

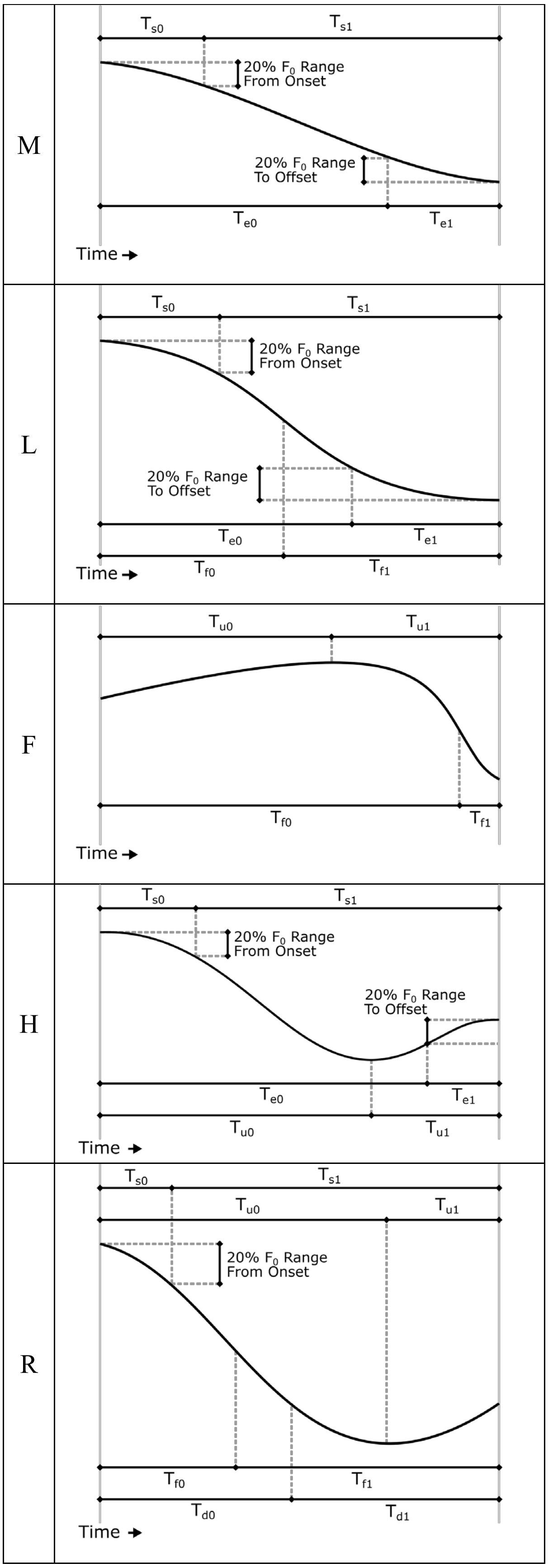

Figure 1: F_0 dynamic measurements on Thai tones.

The first measurement is the movement start time (T_s), the earliest time in syllable when F_0 value differs from syllable onset F_0 level by 20% of the pitch range of that syllable. This indicates the start time of current syllable's target approaching, as F_0 moves from syllable onset's value to approach current syllable's target. The second measurement is the convergence time (T_e). This measurement applies to all static tones (M, L and H). In TA, the target represents the F_0 movement goal. So to measure this, the target level is estimated by fitting the final one third of the syllable with a linear regression function with relative zero position at the end of the syllable. The intercept of the regression is then treated as the target level. After the target of each syllable was obtained, the time in syllable that F_0 value reach target level by a threshold was marked as the target convergence time. The threshold level is 20% of the syllable pitch range. This measurement represents the time that the target is finally approached. The third measurement is the time of fastest F_0 falling (T_f), the time with the maximum speed of F_0 drop. The forth measurement is the time of F_0 turning point (T_u), an F_0 peak where F_0 movement turned from rise to fall, and vice versa. The last measurement is deceleration time (T_d), which is the time when F_0 acceleration is the highest during the falling part of F_0 contour, after which deceleration of the fall starts. R tone is the only tone that this feature was extracted from. Due to the preceding F tone, F tone does not have enough rising momentum for this measurement.

All these timing measurements were made relative to syllable onset and syllable offset. Analysis of Variance (ANOVA) was performed to assess the speaking rate effects on F_0 movement.

3. RESULTS

3.1 Effects of Speaking Rate on Syllable Durations

The distribution of syllable duration across tone and speech rate is shown in Table 1, together with results of repeated measures ANOVA. Slowing speaking rate lengthens syllable durations for all tones. Lexical tone does not affect the syllable duration .

Table 1: Mean of syllable duration (in seconds) as function of speech rate and lexical tone. Values in parentheses are standard deviation.

	Speaking Rate			
Tone	Normal	Slow A	Slow B	ANOVA
M	0.342 (0.076)	0.430 (0.095)	0.538 (0.181)	F=43.11 p<**0.001**
L	0.346 (0.082)	0.430 (0.106)	0.568 (0.190)	F=92.66 p<**0.001**
F	0.342 (0.081)	0.442 (0.096)	0.566 (0.193)	F=143.8 p<**0.001**
H	0.358 (0.082)	0.424 (0.103)	0.535 (0.157)	F=184.4 p<**0.001**
R	0.350 (0.083)	0.446 (0.122)	0.572 (0.207)	F=224.1 p<**0.001**

ANOVA	F=0.4756
	p=0.754
		F=0.2760
		p=0.894
			F=0.5471
			p=0.701	

3.1 M Tone

Features measured on M tone are T_s and T_e. Table 2 shows mean values of these features. For different speaking rate, T_s (onset) is relaltively located at the same location. This indicates that T_s aligns with syllable onset and remains stable across speaking rates. T_e, however, changes with speaking rate. The slower speaking rate delays T_e, but does not align it to syllable offset. This suggests that in slower speaking rate, target approximation of the M tone may start from the later part of the syllable.

Table 2: Mean time position of features found from M tone with different speech rate (in seconds). Value in parentheses are standard deviation.

Feature	Timing Relative to	Speech Rate			ANOVA Result
		Normal	Slow A	Slow B	
T_s	Syllable Onset	0.082 (0.027)	0.077 (0.032)	0.092 (0.052)	F=2.598 p=0.077
	Syllable Offset	-0.260 (0.062)	-0.352 (0.086)	-0.446 (0.177)	F=42.93 **p<0.001**
T_e	Syllable Onset	0.217 (0.055)	0.241 (0.062)	0.282 (0.091)	F=15.18 **p<0.001**
	Syllable Offset	-0.125 (0.048)	-0.189 (0.072)	-0.256 (0.174)	F=24.21 **p<0.001**

3.2 L Tone

Features measured on L tone includes T_s, T_e, and T_f. Table 3 shows mean values of these features across tone and speaking rate, together with results of repeated measures ANOVA. These meausrements indicate that, again, T_s is similar regardless of speaking rate. On the other hand, T_e and T_f are not aligned with either syllable onset or offset when speaking rate changes. Slowing down speaking rate results in the delay of both features toward the end of the syllable.

Table 3: Mean time position of features found from L tone with different speech rate (in seconds). Value in parentheses are standard deviation.

Feature	Timing Relative to	Speech Rate			ANOVA Result
		Normal	Slow A	Slow B	
T_s	Syllable Onset	0.098 (0.034)	0.104 (0.051)	0.133 (0.090)	F=6.278 p=0.022
	Syllable Offset	-0.248 (0.067)	-0.326 (0.095)	-0.435 (0.189)	F=38.21 **p<0.001**
T_f	Syllable Onset	0.142 (0.051)	0.153 (0.076)	0.217 (0.149)	F=11.48 **p<0.001**
	Syllable Offset	-0.204 (0.082)	-0.277 (0.106)	-0.351 (0.194)	F=20.64 **p<0.001**

T_e	Syllable Onset	0.224 (0.060)	0.263 (0.078)	0.344 (0.133)	F=29.42 **p<0.001**
	Syllable Offset	-0.122 (0.042)	-0.167 (0.074)	-0.223 (0.161)	F=16.63 **p<0.001**

3.3 F Tone

Features measured on F tone are T_f and T_u. Turning point in F tone is the peak of the F_0 contour. Table 4 shows mean values of these measurement, together with results of repeated measures ANOVA. T_f aligns consistently with the syllable offset, but T_u aligns to neither syllable onset nor syllable offset. This alignment indicates that the implementation of F tone is delayed until almost the end of the syllable, before the F_0 falling movement starts to take place.

Table 4: Mean time position of features found from F tone with different speech rate (in seconds). Value in parentheses are standard deviation.

Feature	Timing Relative to	Speech Rate			ANOVA Result
		Normal	Slow A	Slow B	
T_u	Syllable Onset	0.137 (0.062)	0.182 (0.072)	0.236 (0.122)	F=21.91 **p<0.001**
	Syllable Offset	-0.205 (0.085)	-0.260 (0.094)	-0.330 (0.196)	F=15.35 **p<0.001**
T_f	Syllable Onset	0.291 (0.088)	0.392 (0.094)	0.513 (0.182)	F=52.76 **p<0.001**
	Syllable Offset	-0.051 (0.055)	-0.050 (0.047)	-0.053 (0.059)	F=0.052 p=0.949

3.4 H Tone

Features extracted from H tone are T_s, T_e, and T_u. Turning point in H tone is the valley of the F_0 contour. Unlike F and R tones, this valley occurs because of the carryover effect of the preceding syllable. Table 5 shows mean values of these features, together with results of repeated measures ANOVA. T_s of various speaking rate were aligned to syllable onset. T_e were aligned to syllable offset, even with different speaking rate. T_u, however, is aligned neither to syllable onset nor offset. Like F tone, this result indicates that the effect of speaking rate persists through the turning point, but the H tone target is implemented roughly by the end of the syllable.

Table 5: Mean time position of features found from H tone with different speech rate (in seconds). Value in parentheses are standard deviation.

Feature	Timing Relative to	Speech Rate			ANOVA Result
		Normal	Slow A	Slow B	
T_s	Syllable Onset	0.059 (0.023)	0.067 (0.054)	0.072 (0.054)	F=1.399 p=0.249
	Syllable Offset	-0.298 (0.072)	-0.358 (0.101)	-0.462 (0.178)	F=39.50 **p<0.001**
T_u	Syllable Onset	0.241 (0.049)	0.265 (0.062)	0.282 (0.097)	F=5.663 **p=0.004**
	Syllable Offset	-0.117 (0.047)	-0.159 (0.082)	-0.253 (0.164)	F=28.88 **p<0.001**
T_e	Syllable Onset	0.231 (0.114)	0.316 (0.137)	0.399 (0.162)	F=25.94 **p<0.001**
	Syllable Offset	-0.127 (0.080)	-0.108 (0.074)	-0.136 (0.109)	F=1.811 p=0.166

3.5 R Tone

Feature extracted from R tone are T_s, T_f, T_u, and T_d. Turning point in R tone is the valley of the F_0 contour. Table 6 shows mean values of these features. T_f align with the syllable onset. Timing of other features are not aligned to either syllable onset or offset. This suggests that the delay starts after the first falling part in the R tones which is before the turning point.

Table 6: Mean time position of features found from R tone with different speech rate (in seconds). Value in parentheses are standard deviation.

Feature	Timing Relative to	Speech Rate			ANOVA Result
		Normal	Slow A	Slow B	
T_s	Syllable Onset	0.076 (0.025)	0.083 (0.050)	0.103 (0.089)	F=3.772 **p=0.025**
	Syllable Offset	-0.274 (0.072)	-0.363 (0.112)	-0.469 (0.197)	F=36.06 **p<0.001**
T_f	Syllable Onset	0.109 (0.031)	0.116 (0.063)	0.133 (0.099)	F=2.332 p=0.100
	Syllable Offset	-0.241 (0.073)	-0.330 (0.108)	-0.438 (0.199)	F=36.83 **p<0.001**
T_u	Syllable Onset	0.270 (0.060)	0.319 (0.080)	0.361 (0.091)	F=23.86 **p<0.001**
	Syllable Offset	-0.079 (0.047)	-0.127 (0.088)	-0.211 (0.181)	F=22.09 **p<0.001**
T_d	Syllable Onset	0.165 (0.044)	0.193 (0.089)	0.219 (0.134)	F=6.050 **p=0.003**
	Syllable Offset	-0.184 (0.082)	-0.253 (0.118)	-0.353 (0.207)	F=24.04 **p<0.001**

4. DISCUSSIONS AND CONCLUSIONS

This paper shows the computational measurements of the effects of speaking rate on the F_0 dynamics of Thai tones. The results in Table 1-6 clearly indicates that speaking rate significantly affects the F_0 contours and the localizations of pitch events in many ways. The results in this paper have shown that the effects vary according to the nature of the tones, but nevertheless, a consistent time delay pattern of F_0 dynamic events can be found throughout all tones. This persistent pattern suggests that there is a time delay after the onset of the target approximation of the Thai tones as speaking rate slows down.

Further study on the modeling of the time delay mechanism is needed. The results show that the delay often extends to the later part of the syllable, which can be directly incorporated into the TA model as an additional mechanism of the model. Simply speaking, for a slow speaking rate, target approximation can be divided into a waiting phase and an implementation phase.

5. REFERENCES

[1] Xu., Y. 1997. Contextual tonal variations in Mandarin. *J. Phon.* 25, 61-83.

[2] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions, *Speech Commun.*. 46(3-4), 220-251.

[3] Bänziger, T., Patel, S., Scherer, K. R. 2014. The role of perceived voice and speech characteristics in vocal emotion communication, *J. Nonverbal Behav.*. 38(1), 31-52.

[4] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. 2008. Encoding emotions in speech with the size code -- A perceptual investigation. *Phonetica* (65), 210-230.

[5] De Jong, K. 2004. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *J. Phon.* (32), 493-516.

[6] Mermelstein, P. 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am.

[7] Kessinger, R.H., Blumstein, S.E. Effects of speaking rate on voice-onset time in Thai, French, and English. *J. Phon.* 24(2), 143-168.

[8] Moon, S. J., Lindblom, B. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.* 96, 40.

[9] Xu, Y. 2001. Fundamental frequency peak delay in Mandarin. Phonetica 58, 26-52.

[10] Sereno, J. A., Lee, H., Jongman, A. 2015. Effects of speaking rate and context on the production of Mandarin tone. *Proc. 18th ICPhS Glasgow*.

[11] Gandour, J., Tumtavitikul, A., Satthamnuwong, N. 1999. Effects of Speaking Rate on Thai Tones. *Phonetica* 56, 123-134.

[12] Edwards, J., Beckman, M. E., 1988. Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica* 45, 156-174.

[13] Wang, C., Zhang, J., Xu, Y. 2018. Compressibility of segment duration in English and Chinese. *Proc. 9th Speech Prosody, Poznań*, 651-655.

[14] Atterer, M., Ladd, D. R. 2004. On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. *J. Phon.* 32, 177-197.

[15] Mücke, D., Grice, M., Becker, J., Hermes, A. 2009. Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *J. Phon.* 37, 321-338.

[16] Ladd, D. R., Faulkner, D., Faulkner, H., Schepman, A. 1999. Constant "segmental anchoring" of F_0 movements under changes in speech rate. *J. Acoust. Soc. Am.* 106, 1543.

[17] Ladd, D. R., 2004. Segmental anchoring of pitch movements: autosegmental phonology or speech production? *LOT Occasional Series* 2, 123-131.

[18] Prom-on, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am.* 125, 405-424.

[19] Prom-on, S., Liu, F., Xu, Y. 2012. Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *J. Acoust. Soc. Am.* 132, 421-432.

[20] Xu, Y., Prom-on, S. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Commun.* 57, 181-208.

[21] Xu, Y. 2013. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. Proc. *TRASP 2013, Aix-en-Provence*, 7-10.

[22] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341-345.