Pitch Target Representation of Thai Tones

Santitham Prom-on [1,2], *Yi Xu* [1]

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

Abstract

This paper presents a quantitative representation of Thai tones in the form of pitch targets. The optimal pitch targets of Thai tones were estimated by using a combination of the quantitative Target Approximation (qTA) model and a stochastic learning algorithm. With the advent of this technique, the representation of communicative functions can be systematically and quantitatively studied. The statistical analysis of estimated parameters shows clear contrastive differences between tone categories, indicating unique representation of Thai tones. The results of the modeling experiment demonstrate the adequacy of unitary (as opposed to compositional) underlying targets for generating continuous tonal contours in Thai.
Index Terms: Thai tone, pitch target, modeling

1. Introduction

Thai has five lexical tones, namely, common or Mid (M, 0), Low (L, 1), Falling (F, 2), High (H, 3) and Rising (R, 4). Thai tones are traditionally classified as either static or dynamic based on their pitch movements [1, 2]. There have been a number of empirical studies of Thai tones to identify their properties and interactions with various factors [1,3-9]. In particular, regarding tone representation, it has been previously proposed to be either a unified contour unit [3,5,10-12] or a sequence of H and L autosegments [13-15]. Different accounts of this argument have been offered, but the evidence is still not conclusive. This is particularly because most of the evidence is derived from interpretation of perceptual tests and there is a lack of tools to learn tone representations directly from data. This study uses a combination of the quantitative Target Approximation (qTA) model [16] and the simulated annealing algorithm [17] to learn the pitch targets underlying the F_0 movements of Thai tones in continuous speech. The study aims to explore the contrastive properties of Thai tone by means of computational modeling.

2. Method

2.1. Corpus

The corpus consists of 2500 four-syllable utterances recorded by five native Standard Thai speakers (3 males and 2 females). All speakers were undergraduate students, aged 20-25, studying at King Mongkut's University of Technology Thonburi. They all grew up in the Greater Bangkok region.

In this paper, Thai lexical tone and vowel length were manipulated in a full factorial design. Each utterance consisted of 4 syllables, with the tones of the two middle syllables varying across all 5 tones and 2 vowel lengths. The first and the last syllables were always M tones to minimize both carryover and anticipatory influences on the two middle syllables. Thus there were 100 tone and vowel length combinations in total. Each utterance was repeated five times by each speaker. Table 1 shows the sentence structure of the corpus.

Table 1. *Sentence structure of the corpus.*

1st	2nd	3rd	4th†
kʰun0 M	ʔa:0/nim0 M	la:0/loŋ0 M	ŋa:n0 or ma:0 M
	no:j1/mam1 L	ʔa:n1/man1 L	
	ma:2/nim2 F	wa:ŋ2/maj2 F	
	na:3/miŋ3 H	ne:n3/lom3 H	
	la:n4/jiŋ4 R	ha:4/loŋ4 R	

† The word of the 4th syllable depends on the preceding vowel (ŋa:n0 if it is preceded by a long vowel or ma:0 if it is preceded by a short vowel).

2.2. PENTATrainer

Based on the Parallel Encoding and Target Approximation (PENTA) framework [18], PENTATrainer (*pen·ta·train·ner*) is a Praat script integrated with a Java program that facilitates investigation of the encoding schemes of communicative functions in any language. It encapsulates the quantitative Target Approximation (qTA) model, which represents the dynamic F_0 control [16], and the simulated annealing optimization, which is a stochastic learning algorithm [17] used to globally optimize the functional parameters. Provided with the sound files and their annotations, the program automatically learns the optimal parameters of all possible functional combinations that the user has annotated. After the optimization, the learned functional parameters can be used to synthesize F_0 contours according to any of the given communicative functions. A brief summary of the modeling technique will be discussed in the following sections. The program can be freely downloaded from (http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer2/).

2.3. Functional annotation

The PENTA framework assumes that communicative functions are encoded in parallel in terms of pitch target parameters, which are the inputs to the target approximation process [18]. PENTATrainer implements this parallel encoding scheme by allowing users to create tiers of factors as illustrated in Figure 1. These factors can represent either communicative functions such as tone, focus, or sentence type, or other factors that may influence the prosody such as vowel length or part of speech. In this study, as shown in Figure 1, the factors under scrutiny are tone and vowel length. The first tier is an optional one on which users can set a constraint on the type of the target (e.g., S for static target and D for dynamic target). In this study, all the target types were set to dynamic so that both slope and height were automatically learned.

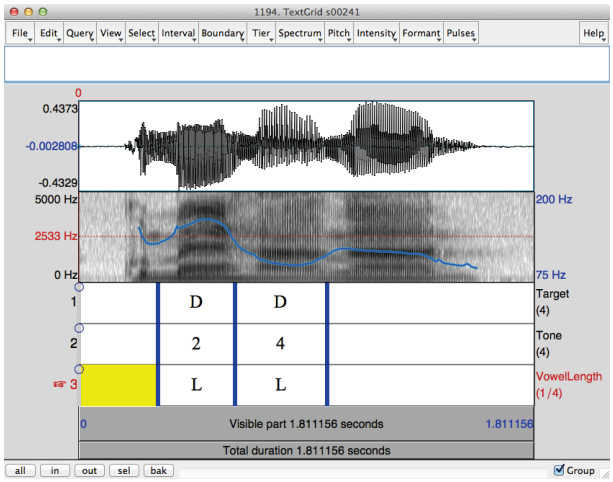

Figure 1: *A snapshot of functional annotation window in PENTATrainer.*

2.4. Modeling

PENTATrainer uses the qTA model to represents F_0 control as a process of target approximation. Figure 2 illustrates the concept of target approximation. F_0 contours (black solid curve) are the responses of the target approximation process. Pitch targets (gray dashed line), which are synchronized to the host interval (demarcated by the boundaries represented by the vertical gray lines), represent the goals of the F_0 control. The F_0 dynamic state at the end of an interval is transferred to the next one. For more details about the formulation of qTA, please refer to [16].

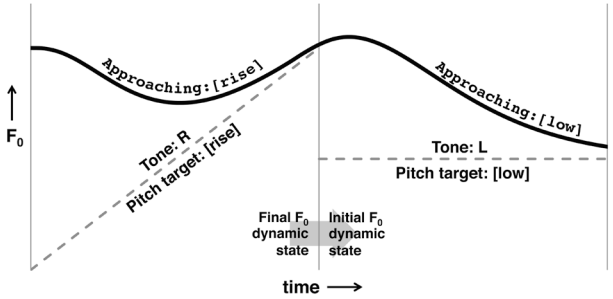

Figure 2: *An illustration of the target approximation process* [20].

In qTA, there are three model parameters controlling the F_0 trajectory of each interval, including target slope (m), target height (b), and the rate or strength of target approximation (λ). m and b specify the form of the pitch target. For example, the Mandarin rising and falling tones are found to have positive and negative m values, respectively [16,19]. The current version of PENTATrainer represents b as relative to the speaker F_0 mean. λ indicates how rapidly a pitch target is approached. The higher the value of λ the faster F_0 approaches the target. For example, λ of the Mandarin neutral tone has been found to be smaller than other tones [19].

Despite its success in modeling F_0 contours of individual utterances, there are still difficulties when it comes to summarizing parameters of all syllables into functional categories. Because of the trade-off between model parameters, their nonlinear interplays in the model and the differences in the optimum conditions of parameter estimation process, a simple averaging procedure could sometimes result in representations that do not reflect globally optimal solutions. This issue was addressed in the latest implementation of PENTATrainer. Instead of modeling F_0 contours of each individual utterances and summarizing afterward, the parameters of all functional categories are optimized simultaneously, using the simulated annealing algorithm [17]. At the initial stage, the algorithm randomly generates the parameters of all functional categories. These parameters were then repeatedly modified and tested to decide whether to accept or reject the proposed changes. The probability of acceptance/rejection depends on the temperature parameter of the algorithm. At the initial round, the temperature is set to a high value. It then gradually reduces as the procedure is repeated. This allows the solution to evolve and converge to the global optimum over the iterations. Since the final parameters may differ slightly due to the randomness of the stochastic method, in this study, the learning process was repeated 5 times for each speaker to obtain more stable solutions. The median values of m, b, and λ were calculated across the repetitions for each functional category of each speaker.

Estimated parameters were analyzed using two-factor analysis of variance (ANOVA) in SPSS. Afterward, post-hoc analysis of each functional category was performed using Scheffé's post-hoc test to determine the nature of the differences between groups. The parameter distributions were further analyzed using Student's t-test.

3. Results

3.1. Synthesis accuracy

As the functional parameters are adjusted during the optimization process, the errors between the original and synthesized F_0 change accordingly. Figure 3 shows what happened to the solution during one run. The overall error generally reduced over iterations, while the local minima were skipped. This clearly shows the robustness of the learning algorithm.

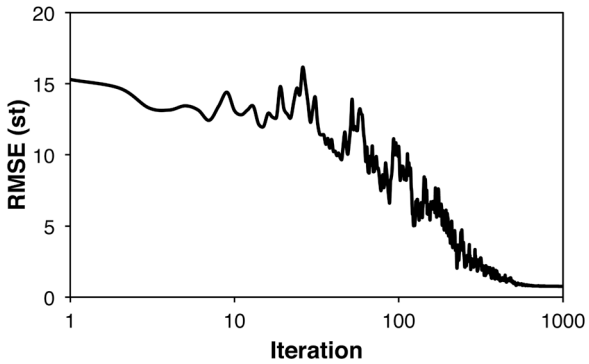

Figure 3: *An example of the per-utterance error reduction in the learning process. The data were obtained from one simulation run.*

Table 2 shows means and standard errors of root mean square error (RMSE) and Pearson's correlation coefficient, comparing between original and synthesized F_0 contours, when parameters were summarized for speaker dependent and independent conditions. The speaker dependent parameters yielded significantly better synthesis accuracies than speaker dependent parameters (RMSE: $t(4)=3.55$, $p=0.024$; Correlation: $t(4)=3.74$, $p=0.020$), while the number of parameter sets of the speaker dependent condition was five times higher than that of the speaker dependent condition. This indicates that certain speaker specific characteristics may have been encoded by the speaker dependent parameters.

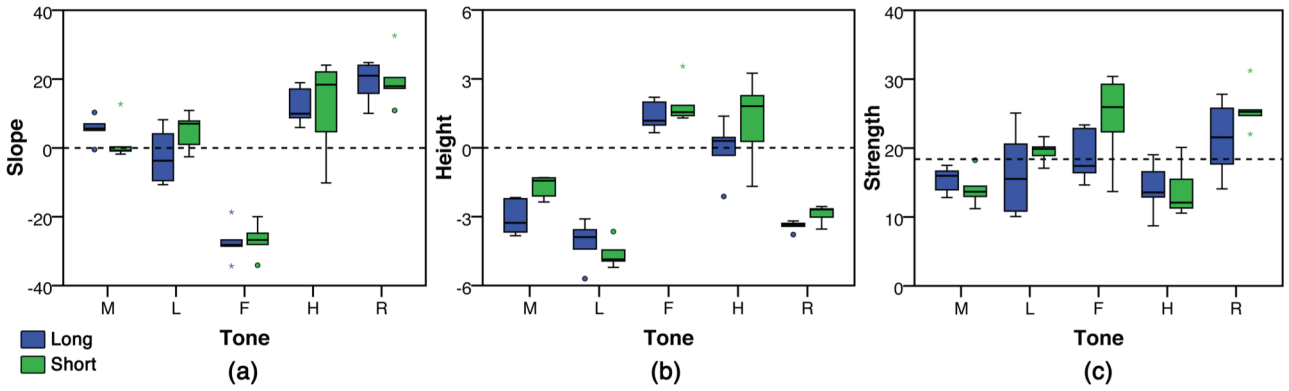

Figure 4: *Parameter distributions of each tone compared to reference values (dashed line) which are 0 in the cases of m and b and the total mean in the case of λ. Blue color indicates the parameters from long vowels while the green color indicates the parameters from short vowels.*

3.2. Parameter analysis

Table 3 shows learned parameters of all Thai tones in different vowel lengths. All parameters significantly differ depending on the tonal categories (*m*: F(4,49)=56.81, $p<0.001$; *b*: F(4,49)=71.07, $p<0.001$; *λ*: F(4,49)=9.23, $p<0.001$). While no difference was found in *m* or *λ*, *b* was significantly different between vowel lengths (F(1,49)=5.37, $p=0.026$),. In particular, M tone was significantly higher in short vowels than in long vowels. This could be a result of undershoot of M tone in short vowels. There were no interactions between tone and vowel length. This indicates the independence of the mechanisms controlling the two factors.

Post-hoc analysis of the *m* revealed interesting patterns in categorical arrangements of tones. Static tones were generally not significant difference from one another, although there was a marginal difference between H and L. (M-L: $p=0.968$, M-H: $p=0.205$, L-H: $p=0.050$). *m* of M and L tones significantly differ from those of dynamic tones (M-F: $p<0.001$; M-R: $p=0.001$; L-F: $p<0.001$; L-R: $p<0.001$). *m* of H tone, however, was not different from that R tone (H-R: $p=0.293$), but significantly different from F tone (H-F: $p<0.001$). This result conforms to the traditional classification of Thai tone into a static-dynamic dichotomy [1].

Comparing the parameter distributions of each tone to the reference values (0 for *m* and *b*, total mean for *λ*) revealed more distinctive properties of each tone as shown in Figure 4. Dynamic tones, including F and R, has their slopes significantly lower and higher than zero respectively, regardless the vowel length. (F-Long: $t(4)=10.85$, $p<0.001$; F-Short: $t(4)=-11.66$, $p<0.001$; R-Long: $t(4)=6.92$, $p=0.002$; R-Short: $t(4)=5.57$, $p=0.005$). This indicates the distinctive property of dynamic tones. On the other hand, *m* of L tone were not significantly different from zero regardless of vowel length (L-Long: $t(4)=0.63$, p=0.565; L-Short: $t(4)=1.98$, $p=0.119$). *m* of M and H tones was significantly higher than zero only in long vowels but not in short vowels (M-Long: $t(4)=3.15$, $p=0.035$; H-Long: t(4)=4.84; $p=0.008$; M-Short: $t(4)=0.70$, $p=0.523$; H-Short: $t(4)=1.83$, $p=0.141$). Further inspection of the means of *m* in Table 3 suggests that H should have a shallow rise target while M should have a static target. For target height, only H was found to be not significantly different from zero regardless of the vowel length (H-long $t(8)=1.72$, $p=0.123$). M, L and R tones have *b* values significantly lower than the total mean (M-Long: $t(4)=8.57$, $p=0.001$; M-Short: $t(4)=7.68$, $p=0.002$; L-Long: $t(4)=9.27$, $p=0.001$; L-Short: $t(4)=17.03$; $p<0.001$; R-Long: $t(4)=33.89$, $p<0.001$; R-Short: $t(4)=16.16$, $p<0.001$), while only F tone has *b* significantly higher than zero (H-Long: $t(4)=4.74$, $p=0.009$; H-Short: $t(4)=4.64$, $p=0.010$). For strength, only M has significantly lower λ compared to the total mean (M-Long: $t(4)=3.47$, $p=0.026$; M-Short: $t(4)=3.67$, $p=0.021$). These contrastive properties in model parameters indicate unique representation of Thai tones. This also indicates the effectiveness of PENTATrainer in learning the abstract representation of communicative functions.

Figure 5 shows comparison of mean time-normalized F_0 contours between originals (solid black line) and synthetics (red dotted line) using the parameters in Table 3. The overall close fit between the two indicates that PENTATrainer can generate most of the contextual tonal variations with the learned categorical targets. What cannot yet be simulated,

Table 2. *Average RMSE in semitone and correlation coefficients comparing between the speaker dependent and speaker independent parameters.*

Conditions	Number of Parameter	RMSE (st)	Correlation
Speaker Dependent	50	0.78 ± 0.05	0.889 ± 0.012
Speaker Independent	10	0.90 ± 0.06	0.871 ± 0.014

Table 3. *Means and standard errors of parameters of the tone function in different vowel lengths.*

Tone	Vowel Length	*m* (st/s)	*b* (st) †	*λ*
M	Long	5.5 ± 1.8	-3.0 ± 0.4	15.4 ± 0.9
	Short	1.9 ± 2.7	-1.7 ± 0.2	14.1 ± 1.2
L	Long	-2.3 ± 3.7	-4.1 ± 0.4	16.4 ± 2.9
	Short	4.8 ± 2.4	-4.6 ± 0.3	19.5 ± 0.8
F	Long	-27.3 ± 2.5	1.4 ± 0.3	18.9 ± 1.8
	Short	-26.7 ± 2.3	1.9 ± 0.4	24.3 ± 3.0
H	Long	12.1 ± 2.5	-0.1 ± 0.6	14.2 ± 1.7
	Short	11.8 ± 6.5	1.2 ± 0.9	13.9 ± 1.8
R	Long	19.1 ± 2.8	-3.4 ± 0.1	21.4 ± 2.5
	Short	19.8 ± 3.6	-2.9 ± 0.2	25.8 ± 1.5

† *b* is relative to the F_0 onset of the utterance.

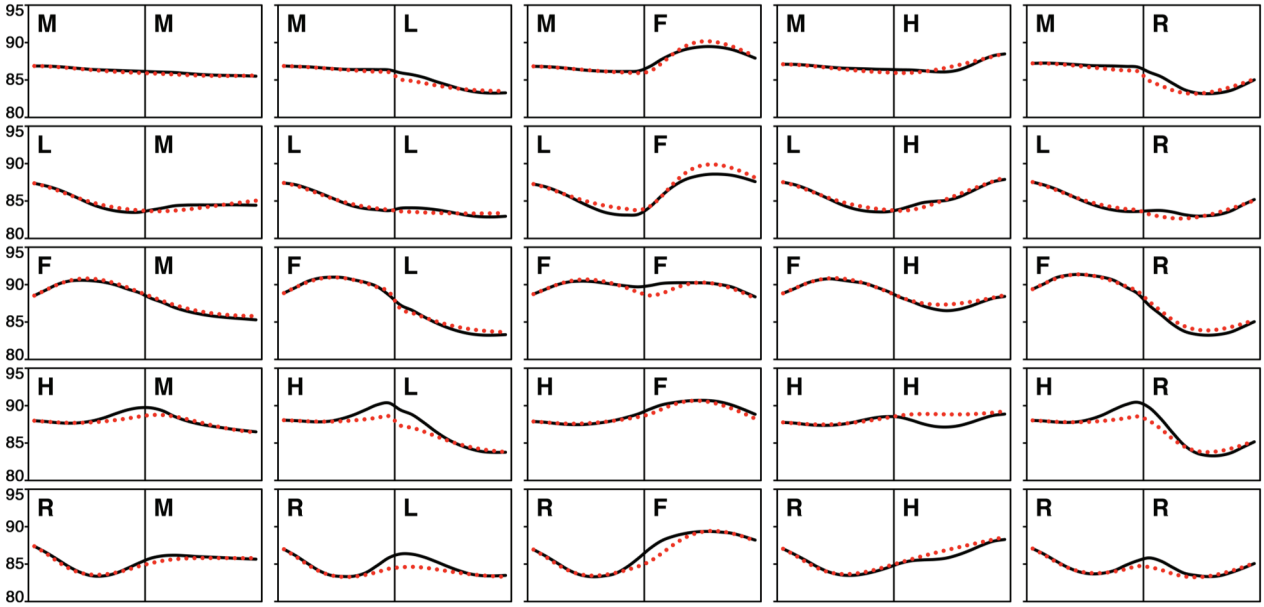

Figure 5: *Mean time-normalized F_0 contours averaged in semitone across 4 vowel length conditions, 5 repetitions, and 5 speakers. Y-axis displays F_0 values in semitone. In each panel, the solid black line is the mean original F_0 contour while the dotted red line is the mean synthesized F_0 contour from PENTATrainer using the parameters in Table 3.*

however, is anticipatory dissimilation [8], which is not currently modeled. For example, when H tone was followed by tones that approach a relatively low F_0 such as M, L or R tones, the original contours of H tones are higher than the synthesized contours.

4. Conclusion

This study has shown Thai tone can be represented by categorical underlying representations in terms of pitch targets. Based on the results from computational modeling and visual inspection, a set of pitch target parameters can be used to simulate Thai tone quite accurately. Distinctive contrasts in model parameters were found for each tone. The results have also demonstrated the effectiveness of new version of PENTATrainer in modeling prosody as communicative functions. Unlike the earlier implementation qTA in [16], in which constraints were imposed on the slope of the tonal targets, here the categorical target parameters were learned without any prior parameter constraints. This is one step forward toward fully automatic extraction of tonal parameters from continuous speech.

5. Acknowledgements

The authors would like to thank the Royal Society and the Royal Academy of Engineering for the financial support through the Newton International Fellowship Scheme.

6. References

[1] Abramson, A. S., "The vowels and tones of Standard Thai: acoustical measurements and experiments", Int. J. Am. Linguist., 28(2):pt.3, 1962.

[2] Luksaneeyanawin, S., "Intonation system in Thai", in D. Hirst and A. Di Cristo [Eds], Internation Systems: A survey of Twenty Languages, 376-395, Cambridge University Press., 1998.

[3] Abramson, A. S., "The tones of Central Thai: some perceptual experiments", in J. G. Harris and J. Chamberlain [Eds], Studies in Tai linguistics, 1-16, Bangkok: Central Institute of English Language, 1975.

[4] Abramson, A. S., "Lexical tone and sentence prosody in Thai", in proceedings of the 9th ICPhS, 380-387, 1979.

[5] Gandour, J., "On the Interaction between tone and vowel length: evidence from Thai dialects", Phonetica, 34(1):54-65, 1977.

[6] Gandour, J., Potisuk, S., Ponglorpisit, S. and Dechongkit, S., "Inter- and intraspeaker variability in fundamental frequency of Thai tones" Speech Commun., 10(4):355-372, 1991.

[7] Gandour, J., "Effects of speaking rate on Thai tones", Phonetica, 56(3-4):123-134, 1999.

[8] Potisuk, S., Gandour, J. and Harper, M. P., "Contextual variations in trisyllabic sequences of Thai tones" Phonetica, 54(1):22-42, 1997.

[9] Thepboriruk, K., "Bangkok Thai tones revisited", J. Southeast Asian Linguist. Soc., 3(1):86-105, 2010.

[10] Abramson, A. S., "Static and dynamic acoustic cues in distinctive tones, Lang. Speech, 21(4):319-325, 1978.

[11] Xu, Y., "Consistency of tone-syllable alignment across different syllable structures and speaking rates", Phonetica, 55(4):179-203, 1998.

[12] Xu, Y., "Understanding tone from the perspective of production and perception" Lang. Linguist., 5:757-797, 2004.

[13] Morén, B. and Zsiga, E. "The lexical and post-lexical phonology of Thai tones", Nat. Lang. Linguist. Th., 24(1):113-178, 2006.

[14] Zsiga, E. and Nitisaroj, R. "Tone features, tone perception, and peak alignment in Thai", Lang. Speech, 50(3):343-383, 2007.

[15] Zsiga, E. "Modeling diachronic change in the Thai tonal space", U. Penn Working Papers in Linguistics, 14(1):article 30, 2008.

[16] Prom-on, S., Xu, Y. and Thipakorn, B., "Modeling tone and intonation in Mandarin and English as a process of target approximation" J. Acoust. Soc. Am., 125(1):405-424, 2009.

[17] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P., "Optimization by simulated annealing", Science, 220(4598):671-680, 1983.

[18] Xu, Y., "Speech melody as articulatory implemented communicative functions", Speech Commun., 46(3-4):220-251, 2005.

[19] Prom-on, S., Liu, F. and Xu, Y., "Functional modeling of tone, focus, and sentence type in Mandarin Chinese", in proceedings of the 17th ICPhS, 1638-1641, 2011.

[20] Xu, Y. and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Commun., 33(4):319-337, 2001