DISCOVERING UNDERLYING TONAL REPRESENTATIONS BY COMPUTATIONAL MODELING: A CASE STUDY OF THAI

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

[1] *Department of Speech, Hearing and Phonetic Sciences, Division of Psychology and Language Sciences, University College London, UK*

[2] *Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand*

ABSTRACT

In the present study we test a computational method for investigating underlying tonal representations. The representation explored is in the form of simple linear functions as ideal pitch targets, with which close-to-natural F_0 contours can be computationally generated. The estimation of the pitch targets is done with PENTAtrainer2, a hypothesis-driven prosody-modeling tool that combines functional annotation, quantitative Target Approximation and global stochastic optimization. In this study we applied PENTAtrainer2 in an investigation of Thai tones. We applied PENTAtrainer2 on a functionally annotated multi-speaker Thai corpus. The pitch targets learned from the corpus showed clear separation between tonal categories, and the F_0 contours synthesized with these targets showed close resemblance to those of natural speech of different speakers, whether or not a particular speaker's data were used in the training. The results demonstrate that it is possible to establish highly economical tonal representations (three parameters per target per tone) that are both fully contrastive and capable of capturing fine phonetic details of Thai tones. Also demonstrated by the study are the effectiveness of PENTAtrainer2 as a prosody research tool, and the potential of computational modeling in general as a new means of basic research in linguistic science.

SUBJECT KEYWORDS

Thai Tones, Representation, Pitch Target, Modeling, Target Approximation

1. INTRODUCTION

The notion of underlying representation is highly appealing in both phonology and phonetics (Chomski and Halle, 1968; Keating, 1988; Pierrehumbert, 1990; Selkirk, 1984; Kuhl, 1989). This is partly because, despite the large amount of phonetic variability in natural speech, the transmission of distinct linguistic categories from the speaker to the listener is so effective that there must be some kind of underlying forms that are shared between the speaker and the listener. But the notion has also turned out to be an extremely difficult one, as so far no clear consensus has been reached in either phonology or phonetics as to what exactly an underlying representation should be like. Nevertheless, based on the general understanding (Chomski and Halle, 1968; Keating, 1988; Pierrehumbert, 1990; Selkirk, 1984; Kuhl, 1989), certain properties should be essential to an ideal underlying representation. That is, it needs to be a) categorically distinctive, b) descriptively economical, c) capable of linking to phonetic variants and d) generalizable across languages. In the present paper we explore one possible form of underlying tonal representation that may exhibit all these properties, using articulatorily based computational modeling as a testing tool.

Early proposals and debates on tonal representations have mostly been concerned with the development of tone features (Duanmu, 1994; Fromkin, 1972; Gandour and Fromkin, 1979; Wang, 1967; Yip, 1980, 2001), i.e., a unified binary set of speech properties that can distinguish not only all the tones in a language, but also all the tones in all tonal languages. Interactions between tones and other factors were represented as transformative rules of tone features. The development of autosegmental phonology and its application to tones (Goldsmith, 1979) lead to the combinatorial representation of tone as a sequence of high and low tones aligned with autosegments. Autosegmental representations of tones had at that time simplified the linguistic analysis process of tones, rendering tone features unnecessary (Clements et al., 2011). Nevertheless, such a simplification treats the contour tone as a sequence of two level tones instead of a unified unit, and reignited an debate that is still ongoing on whether tone, being either level or contour tones, should have a unified unit or a sequence of autosegmental units (Abramson, 1975; Duanmu, 1990, 1994; Gandour, 1977; Morén and Zsiga, 2006; Thepboriruk, 2010; Xu, 1998, 2004; Zsiga and Nitisaroj, 2007; Zsiga, 2008). In the present study we use Thai as a testing case to examine the issue of tonal representation from a modeling perspective.

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 (Abramson, 1962; Luksaneeyanawin, 1998). There have been a number of empirical studies of Thai tones to identify their properties and interactions with various factors (Abramson, 1962, 1975, 1978, 1979; Gandour, 1977, 1999; Gandour et al., 1991; Potisuk et al., 1997; Thepboriruk, 2010). In particular, it has been proposed that each tone in Thai should be represented as either a unified contour unit (Abramson, 1975; Gandour, 1977; Thepboriruk, 2010) or a sequence of H and L autosegments (Morén and Zsiga, 2006; Zsiga and Nitisaroj, 2007; Zsiga, 2008). These proposals offer two very different accounts, but evidence for neither is conclusive, being mostly based on interpretations of perceptual experiments and symbolic analyses. No attempts have been made to predict surface F_0 contours from the proposed representations. This could be partly due to a lack of tools to facilitate such prediction, but could also be partly because detailed predictions have never generally thought to be necessary for establishing the underlying representation of any language. In the present study, we will use the case of Thai tones to explore the idea that a plausible tonal representation should be not only distinctive, economical, and capable of linking to phonetic variants, as is usually assumed, but also capable of *predicting fine phonetic details* and directly derivable from raw *natural speech data*. We will try to achieve this by adopting a computational modeling approach that goes beyond symbolic operations and acoustical analysis. We will also use the modeling results to explore the issue of generality in tonal representation.

The model we will use is the quantitative Target Approximation (qTA) Model, which is based on the theoretical target approximation model (Xu and Wang, 2001). Figure 1 is an illustration of the basic concept of target approximation. Surface F_0 contours (solid curve in Figure 1) are the responses of the target approximation process to the driving force of pitch targets (dashed lines). These targets represent the goals of the F_0 control and are synchronized to the host syllable. Pitch targets are sequentially implemented syllable by syllable, starting from the beginning of an utterance. At the boundary of two syllables, the F_0 dynamic state at the end of the preceding syllable is transferred to the next syllable.

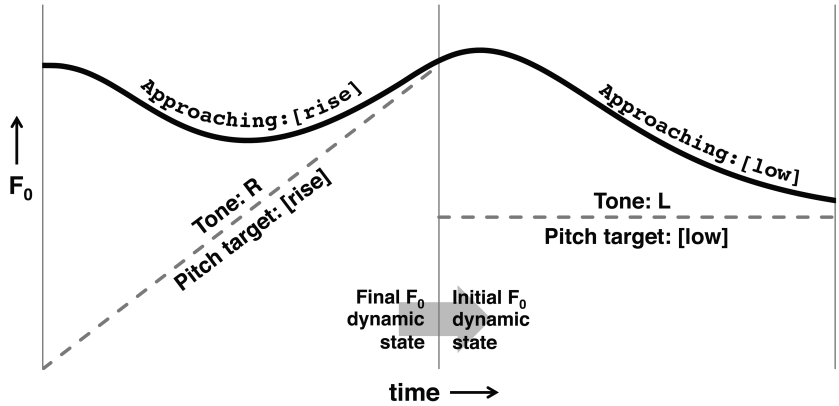

Figure 1. An illustration of the target approximation process (Xu and Wang, 2001). Black solid curve is F_0 contour, gray dashed line indicates the pitch target, and vertical gray line demarcates the syllable boundaries.

In qTA, a pitch target is defined as a forcing function that drives the F_0 movement. It is mathematically represented by a simple linear equation,

$$x(t) = mt + b \tag{1}$$

where m and b denote the slope and height of the pitch target, respectively. t is a relative time from the syllable onset.

The F_0 control is implemented by a third-order critically damped linear system, in which the total response is

$$f_0(t) = x(t) + \left(c_1 + c_2 t + c_3 t^2\right) e^{-\lambda t} \tag{2}$$

where the first term $x(t)$ is the forced response of the system which is the pitch target and the second term is the natural response of the system. The transient coefficients c_1, c_2 and c_3 are calculated based on the initial F_0 dynamic state and pitch target of the specified segment. The parameter λ represents the strength of the target approximation movement. The initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f_0'(0)$, and acceleration, $f_0''(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0. The three transient coefficients are computed with the following formulae.

$$c_1 = f(0) - b \tag{3}$$

$$c_2 = f'(0) + c_1\lambda - m \tag{4}$$

$$c_3 = \left(f''(0) + 2c_2\lambda - c_1\lambda\right)/2 \tag{5}$$

qTA thus defines each pitch target with only three parameters, *m*, *b*, and λ. Of the three parameters, *m* and *b* are used to specify the form of the pitch target. For example, the Mandarin Rising and Falling tones, which differ mainly in target slope, have positive and negative *m* values, respectively; the Mandarin High and Low tones, which differ mainly in target height, have relatively high and low *b* values, respectively (Prom-on et al., 2009, 2011, 2012). Here the value of *b* is relative to a reference pitch, which can be either the speaker F_0 mean or the initial F_0 of an utterance. λ specifies how rapidly the target is approached, with a larger value indicating faster approximation. This approximation rate can define an additional property of a tone. For example, the Mandarin neutral tone is found to have a much smaller λ value than the full tones (Prom-on et al., 2012), which is consistent with the observation that the neutral tone may have a weak articulatory strength (Chen and Xu, 2006).

The ability of simple linear pitch targets to be linked to continuous F_0 contours of of distinctive tones via qTA has been demonstrated in a number of modeling studies (Prom-on et al., 2009, 2012). This, together with their descriptive economy (three parameters per syllable), makes these targets viable candidates for tonal representation. What will be further explored in the present study is that these targets are also directly learnable from raw speech data. We will test this by applying PENTAtrainer2 — a newly developed software for prosody modeling, on Thai tones produced in connected speech.

2. METHODS

2.1 Corpus

The corpus was designed to elicit F_0 contours of the five Thai lexical tones with contextual variations as well as interactions with the two contrastive vowel lengths. Each utterance consisted of four syllables, with the tones of the two middle syllables varying across all five tones and two vowel lengths. The tones of the first and last syllables were always M to minimize both carryover and anticipatory effects on the two middle syllables. Thus there were 100 combinations of tone and vowel length in total. Each utterance was repeated five times. Table 2.1 shows the sentence structure of the corpus.

Five native Standard Thai speakers, three males and two females, recorded the utterances. The corpus thus consisted of 2500 four-syllable utterances. All speakers were undergraduate students, aged 20-25, studying at King Mongkut's University of

Technology Thonburi, Bangkok, Thailand. They all grew up in the Greater Bangkok region. Speech signals were recorded at 22.05 kHz sampling rate and 16-bit resolution.

Table 2.1 Sentence structure of the corpus. The ordinal number on each column header indicates the position of the syllable on the target utterance.

1st	**2nd**	**3rd**	**4th**†
kʰun0 M "คุณ"	ʔa:0/nim0 M "อา/นิม"	la:0/loŋ0 M "ลา/หลง"	ŋa:n0 or ma:0 M "งาน/มา"
	no:j1/mam1 L "หน่อย/หม่ำ"	ʔa:n1/man1 L "อ่าน/หมั่น"	
	mɛ:2/nim2 F "แม่/นิ่ม"	wa:ŋ2/maj2 F "ว่าง/ไม่"	
	na:3/miŋ3 H "น้า/มิ้ง"	ne:n3/lom3 H "เน้น/ล้ม"	
	la:n4/jiŋ4 R "หลาน/หญิง"	ha:4/loŋ4 R "หา/หลง"	

† 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 Deriving Underlying Tonal Representations Automatically – PENTAtrainer2

The speech corpus described in the previous section was annotated and analyzed using PENTAtrainer2, a hypothesis-driven prosody-modeling tool. PENTAtrainer2 (*pen·ta·train·ner·two*) consists of a set of Praat scripts that facilitate the investigation of underlying representations of communicative functions in any language (Xu and Prom-on, 2013). Its core concept is based on the Parallel Encoding and Target Approximation (PENTA) framework (Xu, 2005). PENTAtrainer2 encapsulates the quantitative Target Approximation (qTA) model, which represents dynamic F_0 control (Prom-on et al., 2009), and simulated annealing optimization (Kirkpatrick et al., 1983), which is a stochastic learning algorithm used to globally optimize model parameters. Provided with annotated sound files, PENTAtrainer2 automatically learns the optimal parameters of all possible

functional combinations that users have annotated. After the optimization, the learned functional parameters can be used to synthesize F_0 contours according to any of the given communicative functions. Summaries of the modeling technique will be briefly discussed in the following sub-sections. The program can be freely downloaded from (http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer2/).

PENTAtrainer2 comprises of three main tools: Annotate, Learn, and Synthesize. Each tool corresponds to a main task in the prosody modeling workflow shown in Figure 2. First, the speech corpus is annotated using the Annotate tool. Communicative functions related to a corpus are annotated in separate tiers. Two tiers, tone and vowel length, are annotated for the Thai corpus in this project. Temporal boundaries in each tier are aligned consistently to the prosodic or segmental events of that tier. For the present study, because both tone and vowel length boundaries are synchronized to the syllable, the syllable boundaries are used as temporal markings for both tiers. The co-occurrences of events in the two tiers form functional combinations, which represent interactions between tiers. The annotation step is done iteratively for each sound file in the corpus. In this step, investigators can also inspect and manually rectify the vocal pulse marks used in F_0 calculation. This step requires the most human effort. Next, after all the sound files are annotated, the second step is to estimate the pitch target parameters using the Learn tool. Investigators only need to provide the Learn tool with the optimization parameters, and it will then automatically estimate the optimal parameters of the functional combinations. The third and the last step allows investigators to synthesize or predict the F_0 contours from the learned parameters (or humanly provided parameters if so desired) using the Synthesize tool. In this study, the optimized parameters we used were either speaker-dependent, i.e., learned from each individual speaker, or speaker-independent, i.e., derived by averaging the parameters of all the speakers.

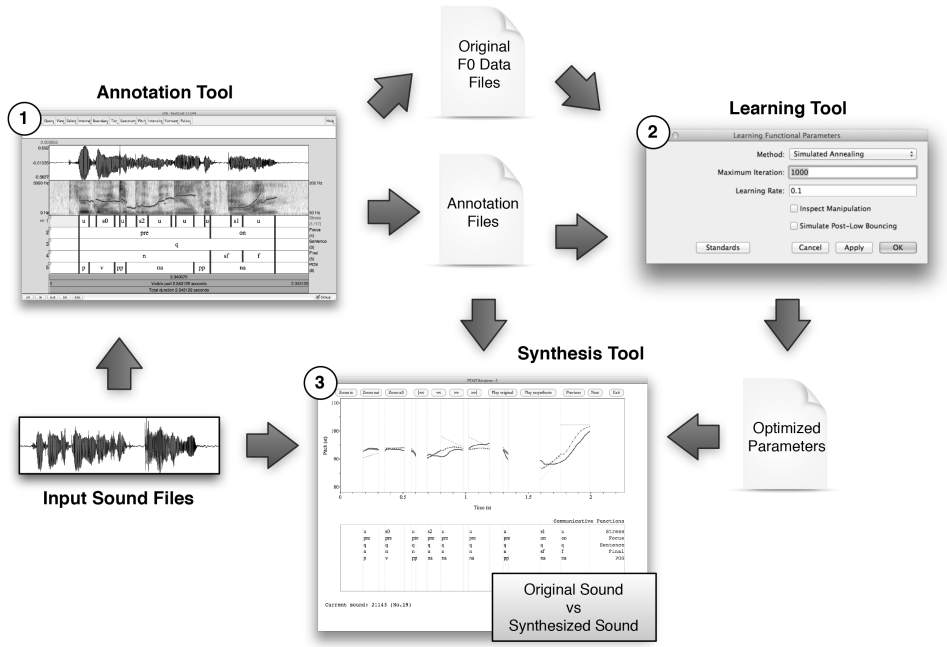

Figure 2. Prosody modeling workflow via PENTAtrainer2 (Xu and Prom-on, 2013). The numbers in circle indicate the sequence of prosody modeling tasks.

2.2.1 Parallel Annotation of Communicative Functions

Figure 3A shows a schematic of the PENTA (parallel encoding and target approximation) framework (Xu, 2005). PENTA assumes that communicative functions are encoded in parallel through function-specific pitch targets, which are articulated via the target approximation process. PENTAtrainer2 implements the parallel encoding aspect of the framework by allowing users to create tiers of factors as illustrated in Figure 3B. The two tiers shown there represent tone and vowel length factors. For other studies, these tiers can represent either well-defined communicative functions such as tone, focus, sentence type, etc., or factors that may influence prosody such as vowel length or part of speech. It is critical to note that the category names, as shown in Figure 3B, do not have any direct effects on the parameter estimation process. They serve only to group and separate individual categories within a particular function, whereas the parameter values will be learned in the learning phrase of the modeling process.

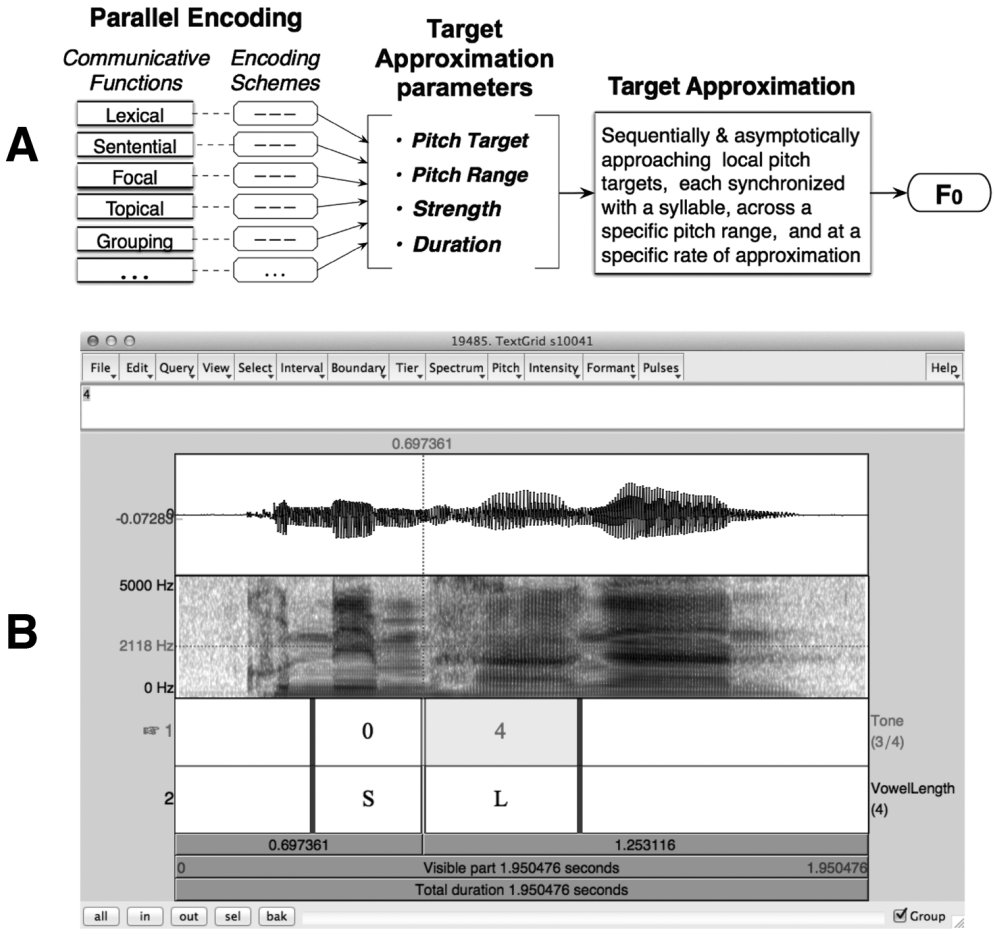

Figure 3. (A) A schematic diagram illustrating the PENTA framework. (B) The implementation of the parallel encoding scheme in the PENTAtrainer2's Annotation tool.

When parallel tiers are overlaid on top of one another, they create overlapped intervals between tiers. By projecting the temporal boundaries of all tiers to each other, we obtain the intervals that represent functional combinations. It is possible that certain functional combinations may appear more than once. For example, in this study, the functional combination 0-S was used for all syllables that have M tone with a short vowel. Summarizing unique combinations from the whole corpus would result in a set of functional combinations that fully represent the interactions between communicative

functions. Each unique functional combination has a corresponding set of pitch target parameters.

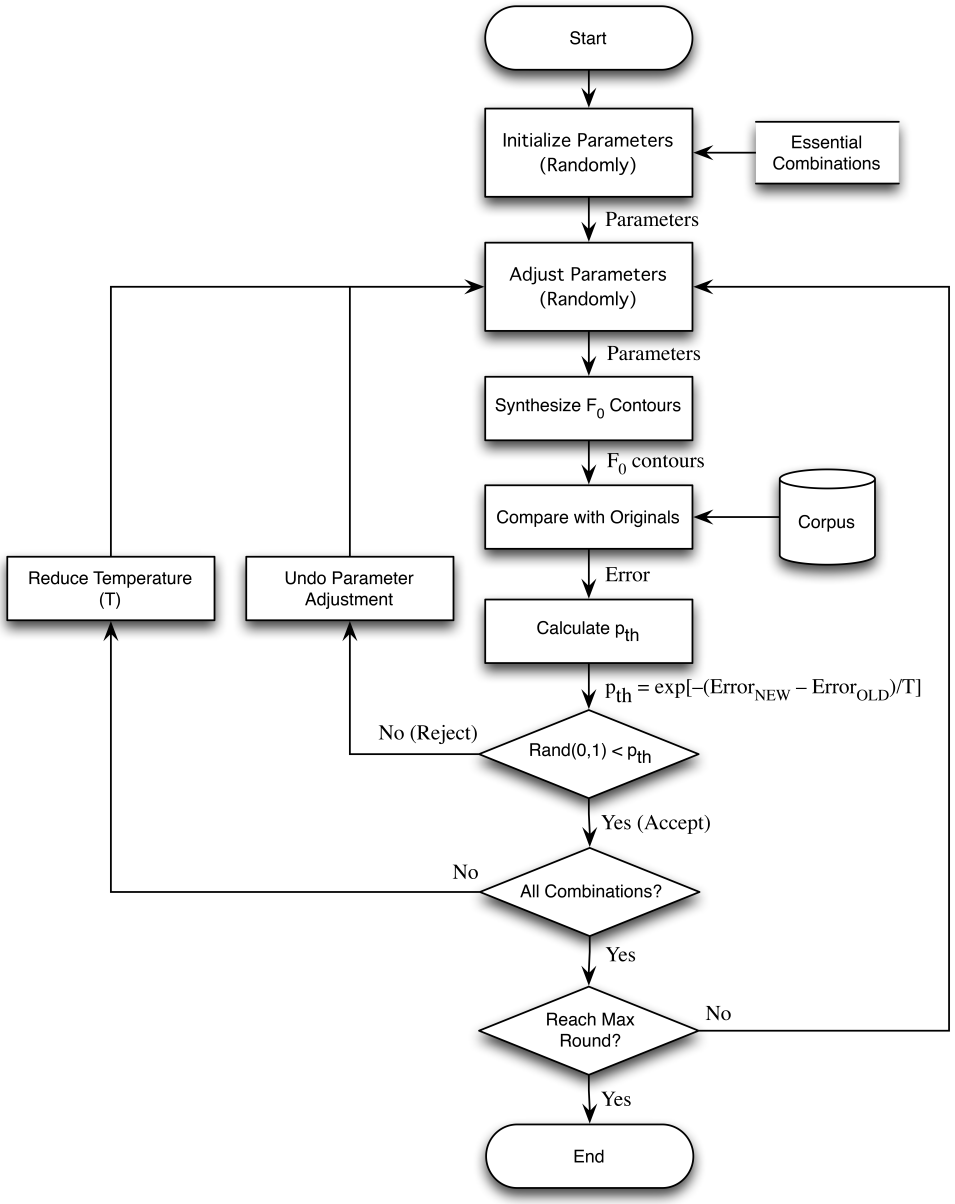

Figure 4. The flowchart of the parameter optimization process of PENTAtrainer2.

2.2.2 Optimization

In our first implementation of qTA, the derivation of the pitch targets is not a full-fledged learning process (Prom-on et al., 2009, 2011, 2012). The procedure, as enabled by PENTAtrainer1 (Xu and Prom-on, 2010-2012), estimates pitch target parameters locally for each syllable. The resulting parameters are then summarized into functional means by averaging across the parameters of all the individual targets that have the same functional combination. The disadvantage of this approach is that, because of the trade-offs between model parameters due to their non-linear interplays in the model, and the differences in the optimum conditions of parameter estimation process, certain properties of the targets, in particular their strength as represented by λ, are difficult to capture. An alternative, which more closely resembles a full-fledged learning process, is to optimize the target parameters globally, i.e., across an entire corpus. This is achieved in the newly developed PENTAtrainer2 (Xu and Prom-on, 2013). Instead of analyzing F_0 contours of each individual utterances and summarizing them afterwards, the parameters of all functional combinations are optimized simultaneously for all the utterances in a corpus, using the *simulated annealing* algorithm (Kirkpatrick et al., 1983). Figure 4 shows the flowchart of how parameters are optimized in PENTAtrainer2.

As shown in Figure 4, at the initial stage, the algorithm randomly generates the parameters of all functional categories. These parameters are then repeatedly and randomly modified and tested for acceptance or rejection. The probability of acceptance/rejection depends on the temperature parameter of the algorithm. At the initial round, the temperature is set to be high, which makes a rejection more likely. 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 order to obtain more stable solutions, we repeated the learning process 5 times for each speaker and calculated the median values of m, b, and λ across the repetitions for each functional category of each speaker.

2.3 Numerical Evaluation

We used two main measurements, root-mean-square error (RMSE) and Pearson's correlation coefficient (henceforth, correlation), to assess the accuracy of synthesis generated with the optimized parameters. RMSE indicates the average distance between

the synthesized and original F_0 contours. Correlation indicates difference in shape between the synthesized and original F_0 contours. These two matrices have been widely used in previous prosody modeling research (Dusterhoff et al., 1999; Jilka et al., 1999; Mixdorff and Jokisch, 2003; Prom-on et al., 2009, 2011, 2012; Taylor, 2000; Xu and Prom-on, 2013). They can be mathematically shown in the following equations.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_0(t_i) - y(t_i)\right)^2} \tag{6}$$

$$\text{Correlation} = \frac{N\sum_{i=1}^{N} y(t_i) f_0(t_i) - \sum_{i=1}^{N} y(t_i)\sum_{i=1}^{N} f_0(t_i)}{\sqrt{N\sum_{i=1}^{N}\left(y(t_i)\right)^2 - \left(\sum_{i=1}^{N}\left(y(t_i)\right)\right)^2}\sqrt{N\sum_{i=1}^{N}\left(f_0(t_i)\right)^2 - \left(\sum_{i=1}^{N}\left(f_0(t_i)\right)\right)^2}} \tag{7}$$

where $y(t_i)$ denotes the original F_0 value at time t_i and N is the total number of sample points of that utterance.

Another way of assessing the modeling results is through the examination of the distributions of the learned pitch target parameters. The parameter analysis was done with a two-way analysis of variance (ANOVA) using SPSS. Post-hoc analysis of each functional category was also performed using Scheffé's 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 Learnability and Synthesis Accuracy

A critical issue in computational modeling is to avoid local minima during automatic optimization. A "greedy" algorithm would quickly settle down on an immediate improvement and give up considering possible further improvements in later trials. As a result, the algorithm would easily fall into a local minimum and remain there. A less greedy algorithm is more "open-minded" and thus would have a better chance of climbing out of a local minimum, but it often needs more learning time than a greedy algorithm. Simulated annealing, as described in 2.2.2, combines of the advantages of both types of algorithms by adjusting the acceptance threshold, as controlled by *temperature*, over time. This allows the learning process to easily get out of local minima in the earlier stages but become more efficient in the later stages as the chance of encountering large errors is reduced. This can be seen in Figure 5, which shows what happened during one run in

terms of the overall errors in RMSE between the original and synthesized F_0. The overall errors gradually reduced over iterations as the functional parameters are adjusted during the optimization process.

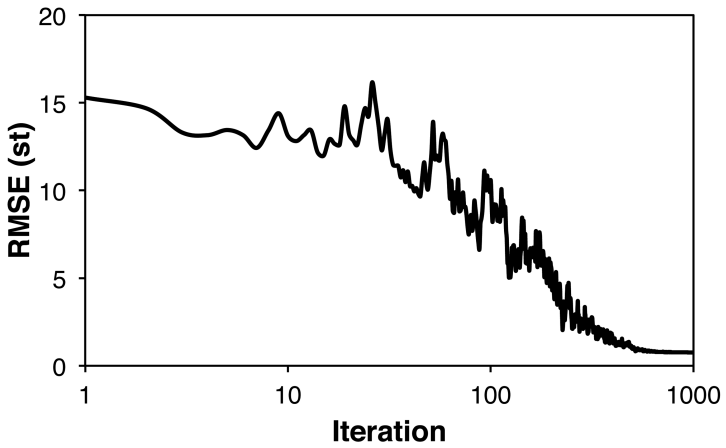

Figure 5. An example of the per-utterance error reduction in the learning process. The data are from one simulation run.

Table 2 shows means and standard errors of RMSE and correlation when parameters were summarized for either speaker-dependent or speaker-independent conditions. Speaker-independent parameters were derived from averaging across speaker-dependent parameters of all speakers. The speaker-dependent parameters (which were also functionally separated) yielded significantly better synthesis accuracies than speaker-independent 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. Both RMSE and correlation were at acceptable levels compared to previous works (Prom-on et al., 2009, 2011, 2012), indicating that the optimized functional parameters correctly represented F_0 movements resulting from the interactions between tone and vowel length. This accuracy assessment is a very important process before proceeding to parameter analysis as it indicates whether or not the given annotations were able to accurately predict the F_0 contours.

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

Condition	Number of Parameter sets	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

3.2 Parameter Analysis

Table 3 shows optimized 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 is found in *m* or λ, *b* is significantly different between vowel lengths ($F(1,49)=5.37$, $p=0.026$). In particular, M tone is significantly higher in short vowels than in long vowels ($F(1,49) = 5.37$, $p = 0.026$). This could be a result of undershoot of M tone in short vowels. There are no interactions between tone and vowel length, indicating the independence of the mechanisms controlling the two factors.

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

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

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

Post-hoc analysis of the *m* values reveals interesting patterns in categorical arrangements of tones. Static tones are generally not significantly different from one another, although there is 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 significantly differs 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, is not

different from that R tone (H-R: p=0.293), but significantly different from F tone (H-F: p<0.001). These results are consistent with the static-dynamic dichotomy in the traditional classification of Thai tones (Abramson, 1962).

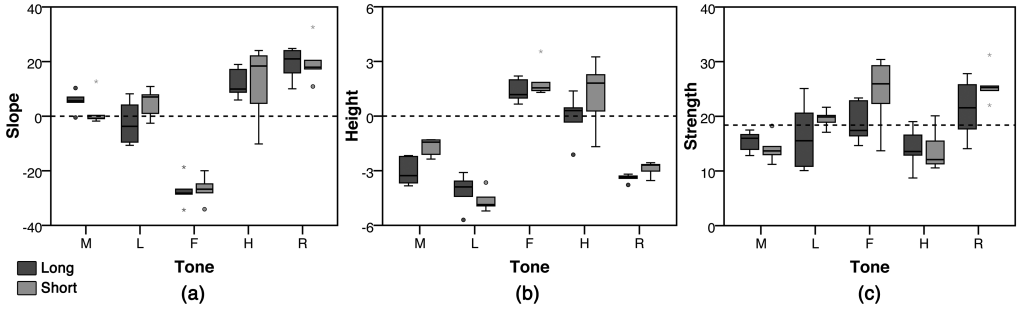

Figure 6. Parameter distributions of each tone compared to reference values (dashed line) which are zero in the cases of m and b and total mean in the case of λ. Dark gray color represents parameters from long vowels while light gray color represents parameters from short vowels.

Comparing the parameter distributions of each tone to the reference values (0 for m and b, total mean for λ) reveals more distinctive properties of each tone as shown in Figure 6. The slopes of the dynamic tones, including F and R, are significantly lower and higher than zero, respectively, regardless of 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 that slope is the distinctive property of the dynamic tones. On the other hand, m of L tone is 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 is significantly higher than zero 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 b, only H is 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. They also indicate the effectiveness of PENTAtrainer2 in learning the abstract representation of communicative functions.

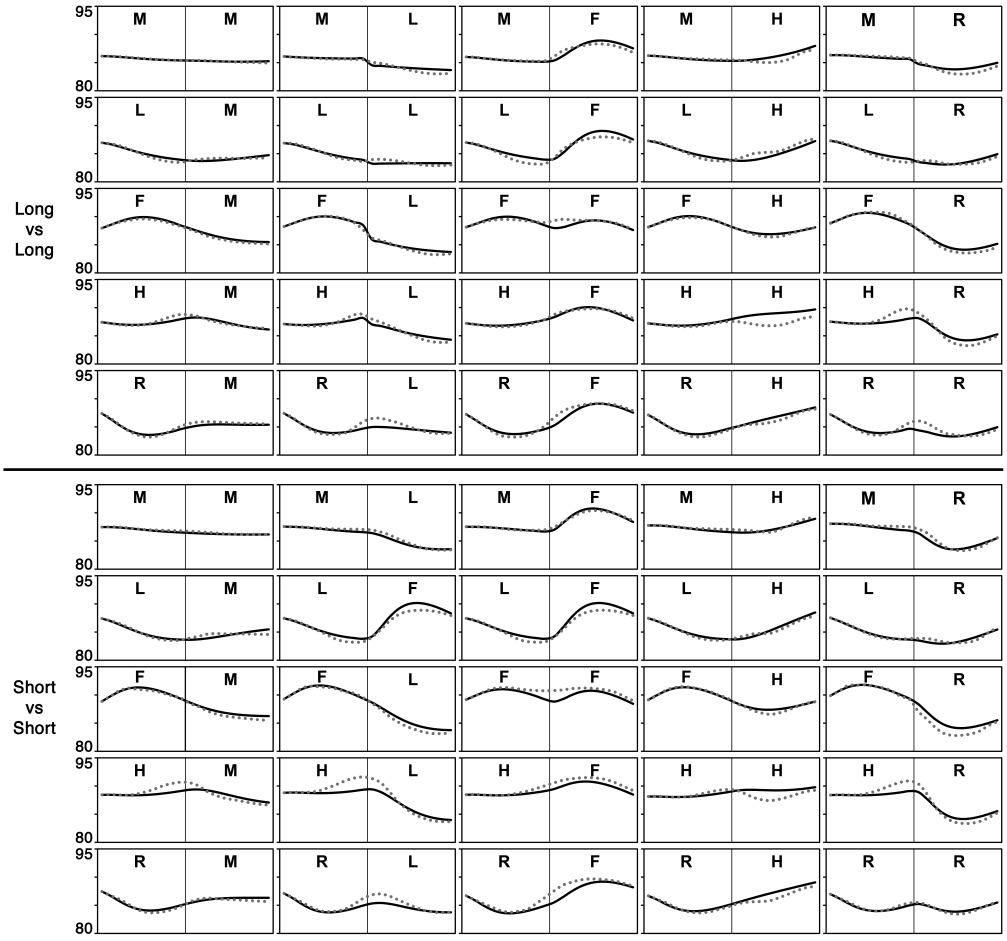

Figure 7. Mean time-normalized F_0 contours averaged in semitone across 4 vowel length conditions, 5 repetitions, and 5 speakers. The Y-axis displays F_0 values in semitone. In each panel, the solid black line is the mean original F_0 contour while the dotted gray line is the mean synthesized F_0 contour from PENTAtrainer2 using the parameters in Table 3.

3.3 Graphical Inspection

Figure 7 shows comparisons between the original (solid black line) and synthetic (gray dotted line) mean time-normalized F_0 contours. The synthetic contours are those generated with the parameters in Table 3. The overall close fit between the two indicates that with the learned categorical targets PENTAtrainer2 can generate most of the contextual tonal variations. Carryover effects can be clearly observed on the right part (second syllable) of each panel. For example, the initial level and slope of M tone of the second syllable (first column, right part) depends on the preceding tones: mid-static when following another M, low-static when following L, high-falling when following F, high-rising when following H, and low-rising when following R.

What cannot yet be simulated, however, is anticipatory dissimilation (Potisuk et al., 1997), which is not currently modeled in qTA. This can be seen in Figure 7 when H tone is followed by tones that approach a relatively low F_0 such as M, L or R. The original contours of H tone are higher than the synthesized ones. Nevertheless, the synthesized contours fit H tone in other contexts quite well. It should be noted that the anticipatory dissimilation is different from the tone sandhi in Mandarin since the affected H tone is still the same tone. This phenomenon clearly indicates that there may be a mechanism other than the normal F_0 control that further influences the surface F_0. This can be studied by analyzing the F_0 contour deviations from the normal contextual variation. Such a strategy for identifying extra factors influencing the F_0 variation has been previously employed in studying Mandarin post-low bouncing (Prom-on et al., 2012; Xu and Prom-on, 2013). In those studies, the F_0 deviation from the normal target approximation was modeled by a simple linear relation based on a balance-perturbation account of the post-low bouncing phenomenon.

4. DISCUSSIONS

The modeling experiment in this study was aimed to explore if it is possible to computationally establish underlying tonal representations that have the properties of a) categorical distinction, b) descriptive economy, c) capability to predict fine F_0 details, d) learnability from raw speech data and e) generalizability across languages. The results show that simple linear pitch targets, when implemented in the qTA model, clearly exhibit at least the first four of these properties.

With regard to property (a), the distinctiveness of the learned pitch targets can be seen in the parameter distributions shown in Figure 6. Despite the variance of each target parameter, the combination of *m*, *b* show clear separations with minimum overlap. A critical contributor to the good separation of the tones is the ability of qTA to model contextual variations. For example, the surface F_0 contour of the M tone in the second syllable, especially in terms of its slope, varies extensively depending on the preceding context, but all the variants can be generated by qTA with a single target. This is thanks to the transient response in target approximation that effectively simulates the carryover effect of the preceding tone. The ease with which qTA simulates large contextual variability has also been shown in the modeling of the neutral tone in Mandarin (Prom-on et al., 2011, 2012).

As for descriptive economy, each tonal target is defined by only three parameters, *m, b* and λ. Even when both tone and vowel length are considered, the total number of parameters reached only 30. In comparison, in Kochanski and Shih (2003), each tone in Mandarin is specified by no less than 8 parameters: 5 pitch values for the tone template, plus word strength, position of the template relative to the syllable and length of the template relative to the syllable. In Bailly and Holm (2005) and Fujisaki et al. (2005), each syllable is specified by a minimum of 5 parameters. The capability of the simple linear pitch targets to predict fine F_0 details when implemented in qTA is clearly seen in Figure 7, where we can see that the categorically predicted F_0 contours closely fit those of the natural speech. The goodness of fit is also shown in the RMSE and correlation values in Table 2, which compare favorably with previously reported modeling results, especially considering the fact that fewer parameters are used in the present study than any of the earlier studies (Prom-on et al., 2009, 2012).

One of the new advantages of the present approach is that the highly distinctive and economical tonal targets, capable of predicting close-to-natural F_0 contours, can be automatically learned with PENTAtrainer2 with a single optimization procedure. This has never been achieved before in modeling studies. In both Kochanski and Shih (2003) and Bailly and Holm (2005), syllable-sized pitch contour templates need to be extracted first, before extracting parameters for the modification of the templates through either strength manipulation or superposition of functional contours. In the implementation of the Fujisaki model, fully predictive F_0 contour generation from tonal categories have never been attempted (Fujisaki et al., 2005; Gu et al., 2006). This learnability property of the

linear tonal targets has profound theoretical implications, as it demonstrates that, instead of relying on universal (and identical across languages) tonal features as assumed in tonal phonology, contrastive categorical tonal targets can be learned directly from natural speech. Thus children may not need to refer to universal tone features (Zsiga and Nitisaroj, 2007) or "simple level tones" (Clements et al., 2011) in order to acquire the tones of a language.

Similarly, with regard to the issue of whether Thai tones should be represented by unified or compositional units, the present results have shown that it is possible to use unitary underlying pitch targets for all the Thai tones. As seen from Figure 6, the same set of targets are equality applicable to both short (hence monomoraic) and long (hence bimoraic) vowels, with the only exception of certain cases of M tone which is probably due to excessive undershoot. Note in particular how the variations in peak alignment in different tonal contexts are closely simulated, although there is no explicit parametric representation of the peak location as done in Zsiga and Nitisaroj (2007). While these results by themselves do not directly invalidate the compositional tone hypothesis, they have shown the possibility of using predictive modeling to link hypothetical underlying representations to fine-detailed surface F_0 contours. To demonstrate that compositional tonal representations are as equally adequate, it would be necessary to make the hypothesis also computationally falsifiable through quantitative modeling. Note that to build such a model, at least four specifications would have to be made: a) specific underlying heights of H and L, b) timing of the within-syllable boundary between the H and L that make up a falling or rising tone, and c) speed of transition between the two adjacent tones, d) manner of the transition (i.e., linear or curve-linear), and e) how exactly contextual variations are represented. At least in terms of descript economy, this is already more specifications than three-parameters-per-syllabic tonal targets in the qTA model.

Finally, with regard to the generalizability property of underlying tonal representations, there is at first a need to rethink what it actually means. If it means that there exist abstract tone or tonal feature such as H and L across languages, we have to resolve the issue of how they can be mapped onto the different F_0 heights and shapes in different languages. For example, the H and L tone in Thai have very different F_0 values found in the present study from those of the H and L tones in Mandarin as found in Prom-on et al. (2009). It would be inadequate to attribute these differences to phonetic realization rules, because doing so would be to theoretically postulate that tonal

representations have little to do with phonetic reality. If, on the other hand, we take generalizability to mean that there is a universal mechanism with which F_0 can be used to encode lexical contrasts, the fact that qTA is applicable to both tone languages like Thai and Mandarin and non-tone languages like English (Prom-on et al., 2009; present study) can be interpreted as evidence we have identified a mechanism that is fairly generalizable. The full range of its generalizability, of course, will await further testing on other languages.

5. CONCLUSION

The results of this study have shown that it is possible to use the newly developed PENTAtrainer2 to automatically find syllable-sized simple linear pitch targets for Thai tones from raw speech data. With these targets F_0 contours that closely resemble those of natural speech in fine detail can be generated by PENTAtrainer2. Based on these results we have argued that these tonal targets are viable candidates as underlying tonal representations, as they have shown at least four of the desirable properties: categorical distinction, descriptive economy, capability to predict fine F_0 details and learnability from raw speech data. We have also argued that the mechanism of encoding functional contrasts with syllable-sized linear pitch targets could be universally shared across languages, whether or not they have lexical tones, as shown by our other modeling studies (Prom-on et al., 2009, 2011). In this way, at least, simple linear pitch targets can also be considered as having the property of being generalizable across languages.

As unitary units that specify fully continuous parameter space, pitch targets are significantly different from previously proposed qualitative tonal representations, in the form of either tone features or autosegments. But given that pitch targets are both fully contrastive and highly economical, it is hard to argue that, because they are quantitatively specific, they cannot serve to make qualitative contrasts, as is implied by arguments for a clear divide between phonological and phonetic representations (Pierrehumbert, 1990). Instead, what the present findings suggest is that model-based phonetic representations can not only adequately serve contrastive functions, but also at the same time resolve the problem of how to link abstract underlying representation to fine-detailed surface acoustic form. In addition, as also shown by the present results, these linear pitch targets are learnable directly from raw speech data, thus demonstrating their potential as units of speech acquisition.

The present results have also demonstrated the effectiveness of PENTAtrainer2 as a tool for conducting research that is of both theoretical and practical relevance. The practical value of the tool is obvious, as it has the potential of being applied in speech technology. But its theoretical relevance is even more interesting. For the first time, together with our other recent studies (Prom-on et al., 2009, 2011, 2012), we have seen the possibility of using computational modeling to test hypotheses which have so far been addressed only by theoretical argumentation based on symbolic operations or empirical testing supported by statistical analysis. With computational modeling, we can also test a theory by letting it predict full surface phonetic details. This would raise the bar for theory testing and accelerate theoretical development in linguistic research.

ACKNOWLEDGEMENTS

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

REFERENCES

ABRAMSON, A. S. 1962. The vowels and tones of Standard Thai: acoustical measurements and experiments. *International Journal of American Linguistics*, 28(2): pt.3.

ABRAMSON, A. S. 1975. The tones of Central Thai: some perceptual experiments. In *Studies in Tai linguistics*, edited by J. G. Harris and J. Chamberlain, 1-16, Bangkok: Central Institute of English Language.

ABRAMSON, A. S. 1978. Static and dynamic acoustic cues in distinctive tones. *Language and Speech*, 21(4): 319-325.

ABRAMSON, A. S. 1979. Lexical tone and sentence prosody in Thai. *Proceedings of the 9th International Congress of Phonetic Sciences*, Copenhagen, Denmark, 380-387.

BAILLY, G., HOLM, B. 2005. SFC: A trainable prosodic model. *Speech Communication*, 46(3-4): 348-364.

CLEMENTS, N., MICHAUD, A., and PATIN, C. 2011. Do we need tone features? In Tones and Features, edited by E. Hume, J. Goldsmith, and L. Wetzels, 3-24, Berlin: De Gruyter Mouton.

DUANMU, S., 1990. A formal study of syllable, tone, stress and domain in Chinese languages. *PhD thesis*, MIT.

DUANMU, S. 1994. Against contour tone units. *Linguistic Inquiry*, 25(4): 555-608.

DUSTERHOFF, K. E., BLACK, A. W., and TAYLOR, P. Using decision tree within the tilt intonation model to predict F0 contours. *Proceedings of EUROSPEECH'99*, Budapest, 1627-1630.

FROMKIN, V. A. 1972. Tone features and tone rules. *Studies in African Linguistics*, 3(1): 47-76.

FUJISAKI, H., WANG, C., OHNO, S., GU, W. 2005. Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech Communication*, 47: 59-70.

GANDOUR, J. 1977. On the Interaction between tone and vowel length: evidence from Thai dialects. *Phonetica*, 34(1): 54-65.

GANDOUR, J. 1999. Effects of speaking rate on Thai tones. *Phonetica*, 56(3-4):123-134.

GANDOUR, J. T., and FROMKIN, V. A. 1979. On the phonological representation of contour tones. *Linguistics of the Tibeto-Burman Area*, 4(1): 73-74.

GANDOUR, J., POTISUK, S., PONGLORPISIT, S., and DECHONGKIT, S. 1991. Inter- and intraspeaker variability in fundamental frequency of Thai tones. *Speech Communications*, 10(4): 355-372.

GOLDSMITH, J. 1976. Autosegmental phonology. *PhD thesis*, MIT.

GOLDSMITH, J. 1990. Autosegmental and Metrical Phonology. Oxford: Basil Blackwell.

GU, W., HIROSE, K., FUJISAKI, H., 2006. Modeling the effects of emphasis and question on fundamental frequency contours of Cantonese utterances. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1155-1170.

JILKA, M., MÖHLER, G., and DOGIL, G. 1999. Rules for the generation of ToBI-based American English intonation. *Speech Communications*, 28: 83–108.

KEATING, P. A. 1988. Underspecification in phonetics. *Phonology*, 5: 275-292.

KIRKPATRICK, S., GELATT, C. D., and VECCHI, M. P. 1983. Optimization by simulated annealing. *Science*, 220(4598): 671-680.

KUHL, P. K., 1989. On babies, birds, modules, and mechanisms: A comparative approach to the acquisition of vocal communication. In *The comparative psychology of audition: Perceiving complex sounds*, edited by R. J. Dooling and S. H. Hulse, 379-419 Hillsdale, NJ: Erlbaum.

LUKSANEEYANAWIN, S. 1998. Intonation system in Thai. In *Internation Systems: A survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 376-395, Cambridge: Cambridge University Press.

MIXDORFF, H., and JOKISCH, O. 2002. Evaluating the quality of an integrated model of German prosody. *International Journal of Speech Technology*, 6: 45-55.

MORÉN, B., and ZSIGA, E. 2006. The lexical and post-lexical phonology of Thai tones. *Natural Language and Linguistic Theory*, 24(1): 113-178.

PIERREHUMBERT, J. 1990. Phonological and phonetic representation. *Journal of Phonetics*, 18: 375-394.

POTISUK, S., GANDOUR, J., and HARPER, M. P. 1997. Contextual variations in trisyllabic sequences of Thai tones. *Phonetica*, 54(1): 22-42.

PROM-ON, S., LIU, F., and XU, Y. 2011. Functional modeling of tone, focus, and sentence type in Mandarin Chinese. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, 1638-1641.

PROM-ON, S., LIU, F., and XU, Y. 2012. Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America*, 132: 421-432.

PROM-ON, S., XU, Y., and THIPAKORN, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125: 405-424.

SELKIRK, E. 1990. On the nature of prosodic constituency: comments on Beckman and Edwards's paper, in: J. Kingston and M. E. Beckman (Eds.), Papers in Laboratory Phonology 1 — Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge: 179-200.

TAYLOR, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107: 1697–1714.

THEPBORIRUK, K. 2010. Bangkok Thai tones revisited. *Journal of the Southeast Asian Linguistics Society*, 3(1): 86-105.

WANG, W. S.-Y. 1967. Phonological features of tone. *International Journal of American Linguistics*, 33(2): 93-105.

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

XU, Y. 2004. Understanding tone from the perspective of production and perception. *Language and Linguistics*, 5: 757-797.

XU, Y. 2005. Speech melody as articulatory implemented communicative functions. *Speech Communications*, 46(3-4): 220-251.

XU, Y., and PROM-ON, S. 2013. Modeling speech prosody via automatic analysis-by-synthesis: From surface acoustics to invariant underlying representations. Manuscript submitted for publication.

XU, Y., and WANG, Q. E. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communications*, 33: 319-337.

YIP, M. 1980. The tonal phonology of Chinese. *PhD thesis*, MIT.

YIP, M. 2001. Tonal features, tonal inventories and phonetic targets. *UCL Working Papers in Linguistics*, 13: 303-329.

ZSIGA, E., and NITISAROJ, R. 2007. Tone features, tone perception, and peak alignment in Thai. *Language and Speech*, 50(3): 343-383.

ZSIGA, E. 2008. Modeling diachronic change in the Thai tonal space. *University of Pennsylvania Working Papers in Linguistics*, 14(1): article 30.

通过量化建模确立声调的底层形式：泰语的个案研究

Santitham Prom-on and Yi Xu

题要

本研究测试一种确立声调底层形式的计算方法。我们所探讨的底层形式是由简单线性函数表达的理想音高目标。用这种目标可以由算法生成近似于自然语句的 F_0 曲线。音高目标的估算是采用 PENTAtrainer2— 一个由假设驱动的新型韵律建模工具。该工具综合了交际功能标注、量化目标趋近模型和全局随机优化搜索。本项目测试用 PENTAtrainer2 学习并模拟泰语的声调。我们把 PENTAtrainer2 用于一个多发音

人的、有交际功能标注的泰语语料库。学习到的音高目标表现出清楚的声调分类，并且无论目标值是来自于多人的平均还是单个的发音人，合成出的 F_0 曲线都与自然曲线高度相似。实验结果表明，可以通过量化建模确立高度经济的声调底层形式：每个声调只有一个目标，每个目标只有三个参数。这些目标按泰语声调分类的区别明确，并能通过合成还原出微小的语音细节。本研究同时也验证了 PENTAtrainer2 作为一种韵律研究工具的有效性，以及量化建模作为一种新的语言科学基础研究手段的潜力。

关键词　Thai Tones, Representation, Pitch Target, Modeling, Target Approximation