Generation of Fundamental Frequency Contours for Thai Speech Synthesis using Tone Nucleus Model

Oraphan Krityakien[1], *Keikichi Hirose*[1], *Nobuaki Minematsu*[2]

[1]Graduate School of Information Science and Technology, University of Tokyo, Japan
[2]Graduate School of Engineering, University of Tokyo, Japan

{oraphan,hirose,mine}@gavo.t.u-tokyo.ac.jp

10.21437/Interspeech.2013-114

Abstract

As classic and intrinsic requirements, synthetic speech need to convey correct information with good quality of naturalness to listeners. Fundamental frequency (F_0) contours need to be controlled to meet these requirements. Additional challenges have been introduced to tonal languages because the F_0 contour reflects both intelligibility and naturalness of the speech. According to the fact that the F_0 contour in a syllable conveys information asymmetrically, Tone nucleus model has been successfully established. In this study, Tone nucleus model is applied in order to generate F_0 contours for Thai speech synthesis. This is among the first that has introduced the model to other tonal languages other than Mandarin. All tone nuclei for five distinctive tones are defined according to the underlying targets. The full process of F_0 contour generation is presented from the nucleus extraction until the F_0 contour generation for continuous speech. The efficiency and adaptability of the model in Thai language were confirmed by the objective and subjective tests. The model outperformed a baseline without applying the model. The generated F_0 contours showed less distortion, more tone intelligibility and more naturalness. The modified method is also introduced for enhancement. The results showed significant improvement on the generated F_0 contours.

Index Terms: F_0 modeling, tone nucleus model, Thai language

1. Introduction

Thai is a tonal language with five distinctive lexical tones: mid, low, falling, high and rising tones (henceforth, T0, T1, T2, T3 and T4 respectively). Tones play a crucial role in distinguishing the meaning of words with the same sequence of phonemes. In acoustic aspects, fundamental frequencies (henceforth, F_0) in speech signals and their movements in temporal dimension, namely F_0 contour, are widely used to define and discriminate tone types. These tones are characterized by their different F_0 contour patterns as illustrated in Figure 1. These contours are extracted from Thai monosyllablic words uttered by a male native Thai [1]. The F_0 contour in a syllable uttered in isolation shows a very stable pattern whereas it changes drastically with complex variations in a syllable uttered in continuous speech due to the co-articulation effects. However, the listeners can interpret these tone contour variations to the same tonal information.

Although Hidden Markov Model (HMM) provides flexibility on speech synthesis, the speech is typically considered in very shot interval in frame-by-frame manner which is insufficient to generate the lexical tones in syllable-long-unit. It generates sudden contour change, resulting in unnaturalness. Hence, to enhance naturalness of the generated speech, the F_0 contour generation of each tone should be controlled in a longer time interval.

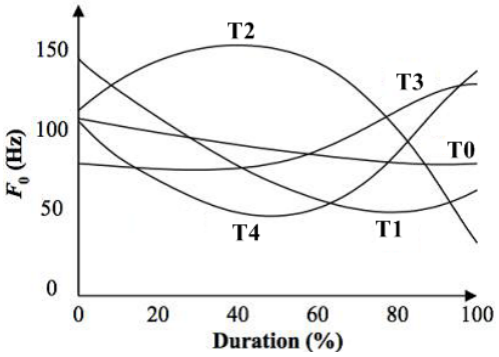

Figure 1: F_0 contours for all five Thai tones.

Previous researches in Thai language have introduced many tone modelings. Many have paid attention to the generation process model (known as Fujisaki model) [2] but most of them are limited to the analysis of a few numbers of natural utterances because of expensive and rather unreliable parameter extraction from the observed F_0 contours. T-Tilt model [3, 4] is a parametric model adapting Tilt model to analyze and synthesize F_0 contours in tonal languages. It is efficient for F_0 analysis but requires extensive work to generate. Also, it needs many parameters to model the F_0 contour and it is weak in predicting parameters from linguistic information.

In this work, Tone nucleus model is applied to Thai language to meet the requirements which are to generate the continuous speech with tone intelligibility and naturalness. Tone nucleus model [5] is a data driven based model. It has been proven that the model requires only compact size of training data, while the HMM-based model requires much more data for the better output speech quality. In this paper, all processes for F_0 contour generation by Tone nucleus model were presented: defining Thai tone nuclei, extracting tone nuclei from the speech utterances, predicting the model parameters, and generating the F_0 contours in continuous speech. The model performance is evaluated through objective and subjective tests by comparing the synthesized speech generated by the model and another synthesized speech generated without the application of the model. In addition, we discussed some modification in prediction and generation processes to get better generated contours. Improvements were observed through the objective and subjective evaluation results.

2. Method

2.1. Tone nucleus model in Thai

The model has been introduced in Mandarin to deal with F_0 contours variations in continuous speech. It was originally developed for tone type recognition [5] and later used in speech synthesis [6, 7]. The model has been built based on the fact that the F_0 contour extracted from the syllable-long-unit natural speech utterance conveys the tonal information asymmetrically. Some part conveys significant information, while the other parts are affected by the co-articulation and show less information.

The concept of the model is to eliminate the F_0 contour which brings less significant tonal information and to pay more attention on the portion which carries critical tonal information to the listeners. The later portion is called "tone nucleus" of the syllable. The model suggests that the F_0 contour should be controlled only in the tone nucleus to convey tonal information instead of directly considering that in the whole syllable which hardly brings any tonal information and produces the data sparseness problem.

Due to the carry-over effect from the preceding syllable, large F_0 perturbation is normally found at the initial consonant, thus the F_0 contour in final vocalic part which includes vowel and voiced final consonant was suggested to be used for tone modeling [8]. Likewise, the articulatory effect from the following syllable is able to be introduced to the ending of the syllable [1]. With regard to these effects, the model ideally divides the F_0 contour in a syllable into three segments: onset course, tone nucleus and offset course as shown in Figure 2. The vertical lines on each F_0 contour in Figure 2 locate the nucleus onset and offset targets. Only tone nucleus is mandatory and is less affected by the adjacent syllables; on the contrary, the onset and offset courses are F_0 transitions from/to the adjacent syllable targets to/from the targets of the current syllable, so they are optional. In short, F_0 contours in these segments are influenced by the neighboring syllables whereas the tone nucleus segment is hardly changed and draws to match the underlying pitch target.

Tone nuclei in Thai were defined according to their underlying targets as shown in Figure 1. Due to the co-articulation effect in the continuous speech, the dynamic tones (T2 and T4) are hardly found with the full-contours. For instance, the downward counterparts of T2 are frequently found without the upward counterparts. (Figure 2 shows that downward counterparts are selected as the tone nuclei.) Nevertheless, in continuous speech, there are still some contrast variations of these dynamic tone contours. These variations were also found in previous researches in Mandarin [6, 7]. In the next section, the method to cope with these variation will be discussed.

2.2. Tone nucleus extraction

In our work, the tone nucleus is assumed to locate in the final vocalic part, hence, the F_0 contours to consider are extracted from the vowel, the optional nasal and semi-vowel codas. According to the preliminary observation, some simple rules focusing on the tone targets were applied to extract Thai tone nucleus. These rules are motivated by the tone nucleus detection method developed for Mandarin speech synthesis [6]. They are mainly related to maximum and minimum F_0 points in the final vocalic part. These target points are later assigned to be either onset or offset target of the tone nucleus corresponding to its tone type. By these rules and our defined Thai tone nucleus model, the level tones (T0, T1, and T3), which Sun [6] reported

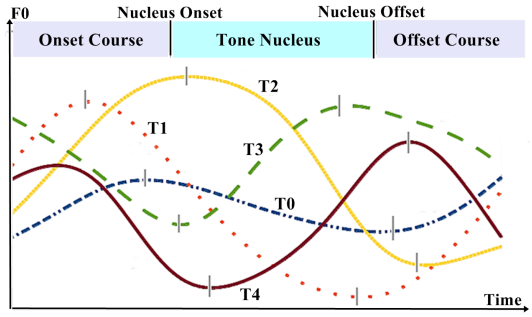

Figure 2: Tone nuclei for all five Thai tones.

that it is rather difficult to detect, can be easily detected.

The extracted tone nuclei are normalized in time dimension into evenly spaced 11 points, and in frequency dimension bounded into 0 to 1 range, where 0 and 1 correspond to minimum and maximum F_0 values. To deal with the variations of the extracted tone nuclei, the normalized $logF_0$ and their $\triangle logF_0$ values are represented by a vector. All vectors in the training data set is clustered into few groups (not more than ten) by K-means clustering with minimum correlation distance separately by tone. The contours clustered in the same group are assigned the same number to represent the tone template identity. The average vector of all 11-point-F_0 data in each group is later used as a tone template representative in the process of F_0 contour generation.

2.3. F_0 contour generation

To generate the F_0 contour in the syllable, time-related and F_0-related parameters are essential. The time-related parameters are nucleus onset and nucleus duration. The F_0 related ones are F_0min, F_0range (displacement of maximum and minimum F_0 values) and tone template identity, which is acquired from the clustering in section 2.2. In this work, the time related parameters were relative values to the syllable length, which were fixed to that of the original speech, while the F_0 related parameters were measured in log scale.

All parameters in each tone were predicted by classification and regression trees (CART) trained separately and independently. Totally there were 25 prediction trees (5 parameters $\times$ 5 tones). These prediction trees were built by question sets corresponding to linguistic and acoustic data of current, previous and following syllables as presented in Table 1. The position type of the syllable in the sentence can be categorized into 6 categories according to [9], but due to the limited prosodic labeling data, the lexical word and phrase were applied instead of prosodic ones.

To train the prediction trees, 30% of the training data set of each tone were randomly selected to form a validation set. The remaining data were used to build the tree. The wagon tool [10] was deployed with stepwise option enabled in order to build the trees with only features that provide good accuracy to the validation data instead of considering all the features. Once all the parameters were predicted, the contour of the predicted template identity was linearly adjusted to fit the predicted minimum F_0, F_0 range and time span regarding to the predicted nucleus duration. Those tone nuclei were located at the specific time points calculated from the predicted nucleus onsets, then they were concatenated by piecewise cubic hermite interpolation to form the F_0 contour of the utterance.

Table 1: Inputs to the tone nucleus parameter predictors

Inputs to the predictor	Category
Initial consonant of current syllable	38
Initial consonant of following syllable	39
Vowel of current syllable	24
Vowel of preceding syllable	25
Coda of current syllable	30
Coda of preceding syllable	31
Tone of current syllable	5
Tone of following syllable	6
Tone of preceding syllable	6
Duration of current syllable	Continuous
Part of Speech of current syllable	14
Position type of current syllable	6
Position of current syllable in word	Natural num.
Number of syllables in current word	Natural num.
Number of words in current sentences	Natural num

Table 2: Number of syllables in the training and target data sets

Set	T0	T1	T2	T3	T4
Training	7753	5112	3916	3699	2002
Target	270	199	146	127	58

3. Experiments and Results

766 training and 30 target utterances were randomly selected from Thai tagged speech corpus for speech synthesis [11] which were uttered by a Thai female professional news reporter in reading style. The selected utterances were carefully checked whether there were no mis-match between speech and linguistic labeling data. Table 2 shows number of syllables in each tones in both data sets. The length of the target utterances were varied from 8 to 58 syllables.

The tone nuclei in the training data set were detected from the utterances and then the essential parameters were extracted from them to build prediction trees. After the trees were built, the parameters in the target utterances were predicted to generate the F_0 contours. The target F_0 contours were substituted with the generated F_0 contours and the speech utterances were re-synthesized by TD-PSOLA technique.

3.1. Experiment 1

To evaluate the performance of the model, we prepared two synthesized speech data sets on the same target data set. One set contains the synthesized speeches by the F_0 contours generated from the tone nucleus model. This set was called TN approach for reference. The other was called WH approach because it consists of the synthesized speeches with the F_0 contours generated by CARTs which were built from the F_0 contours in the whole syllable. The F_0 contour was generated by the predicted F_0min, F_0range and F_0contour shape with fixed syllable length which was obtained from the target contour. In this approach, we used the F_0 contours in the syllables as they are without tone nucleus extraction to train the prediction trees.

For the objective test, the average RMSE of all utterances was considered. The RMSEs calculated from the F_0 contours of the synthesized speeches were compared to those of the target speeches utterance by utterance. In subjective tests, a total 30 native Thais aged varying from 22 to 32 years old, who can speak central Thai fluently, were asked to complete two tasks on a web-based system. None of them are reported to have

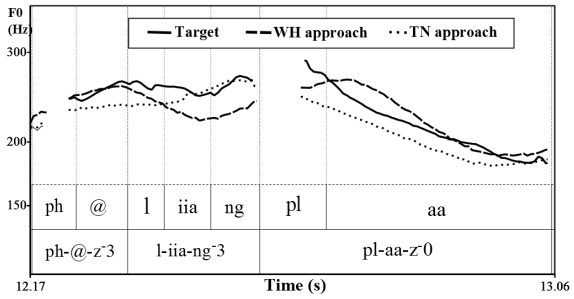

Figure 3: A part of the generated contours from WH (dash line) and TN (dot line) approaches comparing to target contour (solid line) are drawn in the top panel. The vertical lines represents the boundaries according to the phonemes and syllables in lower panels. Tones are marked at the end of each syllable annotation at the bottom panel.

Table 3: Experimental result from objective and subjective tests

Data set	RMSE [logHz]	MOS (90% CI)	Tone error [%]
Target Speech	-	4.91±0.139	-
WH approach	0.6658	3.04±0.093	4.75
TN approach	0.6583	3.22±0.046	4.87
MTN approach	0.6124	3.71±0.077	3.16

hearing problems. They could play each sound as many times as they wanted or take a rest whenever they wanted. The first task was to identify which syllable in the continuous synthesized speeches conveys incorrect tone. Thai textual transcriptions were given along with the target speech sounds to help them judge how the given texts should be pronounced. The total of 60 synthesized sounds generated by TN and WH approaches were randomly presented to the listeners. The percentage of tone error is evaluated in this task. In the second task, the same listeners were asked to give Mean Opinion Score (MOS) with 5-level-scale to evaluate the naturalness of two synthesized speech sets and a target speech set. The values 1 to 5 in the scale rank the naturalness of the utterance from bad to excellent (Bad=1, Poor=2, Fair=3, Good=4, and Excellent=5). All 90 sounds also randomly appeared to the listeners without notifying them which sound is synthetic or natural.

Figure 3 depicts an example of the F_0 contours generated by the TN and WH approaches. The result in Table 3 shows that the TN approach has less distortion and more naturalness but has slightly less tone intelligibility than the WH approach.

Table 4: Average RMSEs of 10-fold cross validation of the predicted F_0mean and F_0range by TN and MTN (modified TN) approaches

	RMSE [logHz]			
	F_0mean		F_0range	
Tone	TN	MTN	TN	MTN
T0	0.092	0.067	0.076	0.067
T1	0.097	0.082	0.086	0.067
T2	0.079	0.069	0.083	0.077
T3	0.099	0.086	0.077	0.079
T4	0.080	0.067	0.079	0.065

3.2. Experiment 2

In Experiment 1, the syllables with tone errors showed high F_0 distortion. Moreover, F_0mean was found to be predicted more accurately than F_0min. Therefore, instead of using F_0min, F_0mean was used to calculate the absolute value of the F_0 contour from the normalized tone template with F_0range. Since many researches found that the F_0 value relates to the syllable duration, the independent parallel prediction were changed to dependent sequential approach. The time-related parameters were estimated, then added to predict the F_0-related parameters. Besides that, the very low tone template identity prediction accuracy (about 45.49% in average) causes tone errors, thus this accuracy should be boosted up to generate the tone contours correctly. On the assumption that tone shape in the previous tone relates to the variation of the current tone shape, the preceding tone template identity was also added into the input question set to help predict the current tone template identity.

In summary, in the modified steps, the nucleus onset and the nucleus duration were predicted first. Next, the predicted nucleus duration were added into the question sets to predict F_0mean, F_0range, and tone template identity. Moreover, the preceding tone template identity were also included in the question set to predict the tone template identity. This modified method will be referred to as MTN approach. 10-fold cross validation on the training data set was performed to confirm the improvement of the prediction results as shown in Table 4. It is clear to see that all RMSEs decreased except in the case of T3's F_0range. However, the increase in RMSE of T3's F_0range is insignificant.

It can be clearly seen from Figure 4 that the contour generated by MTN approach are much similar to the target contour than the one generated by TN approach. A paired-samples t-test showed that RMSEs of MTN approach were significantly less than those of the TN approach, with t (29) = 2.67, and p $<$0.01. The subjective tests were conducted by 19 native Thais. The listeners were asked to identify the syllable which provides the tone error and to give MOS score in the same way as in section 3.1. The percentage of "Good" and "Excellent" evaluation results increased in MTN approach as depicted in Figure 5. The results in Table 3 also reveal the improved tone generation accuracy, less F_0 distortion and more naturalness.

To evaluate the improvement of MTN approach, the preference test was conducted. The same listeners were asked to listen to the synthesized speech from TN approach and the one from MTN approach, respectively and give Comparison Mean Opinion Score (CMOS) on how much the sound from MTN approach is more natural than the one from TN approach with the following rates: 2 (much better), 1 (slightly better), 0 (about the same), -1 (slightly worse) and -2 (much worse). Table 5 shows the percentage of evaluated results in each preference rate, 46% of all evaluations showed that MTN approach is "Better" than TN approach. The average naturalness comparison score of all 30 utterances are 0.449. Wilcoxon signed-rank test was applied to check that the average score is not zero. The test indicated that the data do not come from the distribution of zero median statistically (Z = -2.36, p $<$.05, r = 0.43) This shows that the modified method improves the prediction results to yield more naturalness of the synthesized speech and decrease tone errors.

4. Conclusion

The tone nucleus model was adopted to Thai language for F_0 contour generation with very compact parameter sets. The tone nuclei were defined for all 5 tones based on the underlying targets. The tone nuclei were all detected by the simple rules derived from the observation. CARTs were adopted in tone nucleus parameter prediction. The F_0 contours of the utterances were generated by concatenating the predicted tone nuclei and the speeches were re-synthesized by TD-PSOLA. Although in our work, the phrase components were not extracted before extracting the tone nuclei as done in [6], the objective and subjective evaluations confirm the efficiency of the model in Thai in terms of tone intelligibility and naturalness. Regarding F_0 generation, the proposed method outperformed the whole syllable approach which tends to introduce excessive but insignificant contours and causes the data sparseness. For future work, the model will be considered applying to the synthesized speech by HMM.

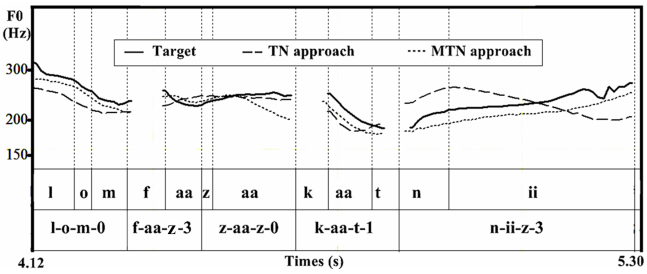

Figure 4: A part of the generated contours from TN (dash line) and MTN (dot line) approaches comparing to target contour (solid line) are drawn in the top panel. The vertical lines represents the boundaries according to the phonemes and syllables in lower panels. Tones are marked at the end of each syllable annotation in the bottom panel.

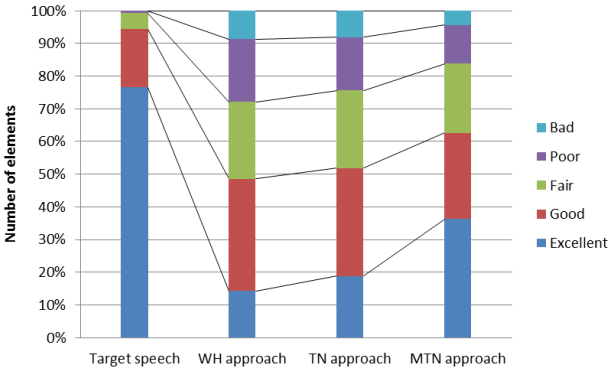

Figure 5: Distribution of the naturalness category

Table 5: Percentages of evaluated results from the preference test

Much better	Slightly better	About the same	Slightly worse	Much worse
19	27	37	14	3

5. References

[1] N. Thubthong and B. Kijsirikul, "Tone recognition of continuous thai speech under tonal assimilation and declination effects using half-tone model," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 6, pp. 815–825, 2001.

[2] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki, and P. Charnvivit, "Perception of tone and vowel quantity in thai," in *Proceedings of ICSLP2002*, 2002, pp. 753–756.

[3] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, A. Rugchatjaroen, and S. Saychum, "T-tilt: a modified tilt model for f0 analysis and synthesis in tonal languages," in *INTERSPEECH*, 2008, pp. 2270–2273.

[4] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol, and C. Wutiwiwatchai, "Optimization of t-tilt f0 modeling," in *INTERSPEECH*, 2009, pp. 508–511.

[5] J. Zhang and K. Hirose, "Tone nucleus modeling for chinese lexical tone recognition," *Speech Communication*, vol. 42, no. 3-4, pp. 447 – 466, 2004.

[6] Q. Sun, K. Hirose, and N. Minematsu, "A method for generation of mandarin f0 contours based on tone nucleus model and superpositional model," *Speech Communication*, vol. 54, no. 8, pp. 932–945, 2012.

[7] W. Miaomiao, W. Miaomiao, H. Keikichi, and M. Nobuaki, "Prosody conversion for emotional mandarin speech synthesis using the tone nucleus model," *IPSJ SIG Notes*, vol. 2011, no. 2, pp. 1–6, jul 2011. [Online]. Available: http://ci.nii.ac.jp/naid/110008584129/en/

[8] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "An empirical study for constructing thai tone models," in *Proc. the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop*, 2002, pp. 179–186.

[9] Y. Hu, M. Chu, C. Huang, and Y. Zhang, "Exploring tonal variations via context-dependent tone models," in *Proc. of Interspeech*, 2007.

[10] P. Taylor, R. Caley, and A. Black, "The edinburgh speech tools library. 1.0.1 edition," The Centre for Speech Technology Research, University of Edinburgh, 1998. [Online]. Available: http://www.cstr.ed.ac.uk/projects/speechtools.html.

[11] C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlamvanich, "Thai tagged speech corpus for speech synthesis," *The Oriental COCOSDA 2003*, pp. 97–104, 2003.