SELECTED PAPER

Generation of Fundamental Frequency Contours for Thai Speech Using the Tone Nucleus Model

Oraphan Krityakien¹, Keikichi Hirose¹ and Nobuaki Minematsu²

¹ Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Phone/FAX :+81-3-5841-6767
E-mail: {oraphan, hirose}@gavo.t.u-tokyo.ac.jp

² Graduate School of Engineering
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Phone/FAX :+81-3-5841-6767
E-mail: mine@gavo.t.u-tokyo.ac.jp

Abstract

To meet classic but still intrinsic requirements of speech synthesis in tonal languages, the tone nucleus model is employed in Thai. Tone nuclei are defined for all five distinctive tones 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 the continuous speech. The objective and subjective tests confirm the efficiency and adaptability of the model in the other language than Mandarin. Compared to the F_0 contours generated by the predictors trained from the contours in the whole syllables without extracting the tone nuclei, the model generates the F_0 contours in the continuous utterances with less distortion from the target but more tone intelligibility and more naturalness. The modified method is also introduced to enhance the accuracy in the F_0 generation by the model.

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). As a tonal language, tones in Thai play a crucial role for listeners to receive the meaning of words. In acoustic aspects, fundamental frequencies (henceforth, F_0) in the speech signals and their movements in the temporal dimension, namely F_0 contour, are widely used to define and discriminate tone types. The different F_0 contour patterns characterize their underlying tone targets as depicted in Figure 1. The F_0 contour in a syllable uttered in isolation shows a very stable pattern while it changes with complex variations when uttered in continuous speech due to the co-articulation effects. However, the listeners can still 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 short interval in frame-by-frame manner which is insufficient to lexical tones in syllable-long-unit. Hence, to increase naturalness of the generated speech, the F_0 contour generation of each tone should be controlled in a longer time in-

Figure 1: F_0 contours for all five Thai tones in monosyllabic words uttered by a male native speaker [1]

terval. Previous researches in Thai language have introduced many tone modelings to fulfill the requirement: good quality with high tone intelligibility. Many have paid attention to the generation process model (known as Fujisaki model) [2], but their works are limited to analysis of limited number of natural utterances because of expensive (and rather unreliable) process of parameter extraction from the observed F_0 contours. T-Tilt model [3] is a parametric model adapting the 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 the linguistic information.

In this work, the tone nucleus model is applied to Thai language to meet the intrinsic requirements which are to generate the continuous speech with tone intelligibility and naturalness. The tone nucleus model [4] is a data driven based model which requires only compact size of the data while the HMM-based model requires much more data for the better output speech quality. All processes for F_0 contour generation by the model were presented: defining Thai tone nuclei, extracting tone nucleus from the speech utterance, predicting the model parameters, and generating the F_0 contour in continuous speech. The model performance is evaluated through objective and subjective tests by comparing the syn-

thesized 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 [4]. It was originally developed for tone type recognition and later used in speech synthesis [5, 6]. 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 [7]. On the other hand, the 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 3 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 current syllable so they are optional. In short, the 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.

2.2 Tone nucleus extraction

The tone nucleus is assumed to locate in the final vocalic part, hence, the F_0 contours to consider are extracted from vowel, 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], where manual corrections are allowed. They are mainly related to

Figure 2: Tone nuclei for all five Thai tones

maximum and minimum F_0 points in the final vocalic part. These target points are later assigned to be either onset or offset 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 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 $\log F_0$ and their $\Delta \log F_0$ 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 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_0 mean, F_0 max, F_0 min, F_0 range (displacement of maximum and minimum F_0 values) and tone template identity, which is acquired from the clustering in section 2.2. The time-related parameters were relative values to the syllable length, which was 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 35 prediction trees. These prediction trees were built by question sets corresponding to linguistic and acoustic data of current, previous and following syllables. In addition, the position type of the syllable in the sentential contexts was included in the question sets. These position types can be categorized into 6 categories according to [8], but due to the limited prosodic labeling data, the lexical

Table 1: 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

word and phrase were applied instead of prosodic ones.

To train the prediction trees, first, the training data set of each tone was separated into 2 subsets exclusively by randomly selecting 30% of the data to form a validation set. The remaining data were used to build the tree. The wagon tool [9] was deployed with stepwise option enabled to be built the tree with only features that provide good accuracy to the validation data instead of considering all features. Once all the parameters were predicted, the contour of the predicted template identity was linearly adjusted to fit the predicted minimum F_0 value, maximum F_0 value 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.

3. Experiments and Results

766 training and 30 target utterances were randomly selected from Thai tagged speech corpus for speech synthesis [10] 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 1 shows number of syllables in each tones in both data sets. The lengths of 30 target utterances varied from 8 to 58 syllables.

The tone nuclei in the training 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 2 synthesized speech data sets on the same target data set. One set contains the synthesized speeches by the F_0 contours generated by the tone nucleus model. This set was called TN approach for reference. The other was called WH approach as it consists of the synthesized speeches with the F_0 contours generated by the CARTs which were built from the F_0 contours in the whole syllable. In this approach, we used the F_0 contours in the syllables as they are without tone nucleus

Table 2: Experimental results from objective and subjective tests

Data set	RMSE (logHz)	MOS	Tone error (%)
Target Speech	–	4.714	–
WH approach	5.092	3.328	4.75
TN approach	5.074	3.442	4.87
MTN approach	4.954	3.741	3.16

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, 30 native Thais were asked to complete two tasks on a web-based system. The first task was to identify which syllable in the continuous synthesized speeches conveys incorrect tone. The total 60 synthesized sounds from both sets were randomly presented to the listeners. Thai transcripts were given along with the target speech sounds to help them judge how the given transcripts should be pronounced. In the second task, the same listeners were asked to give Mean Opinion score (MOS) from 1 (Bad) to 5 (Excellent) for naturalness evaluation of two synthesized speech sets and a target speech set. All 90 sounds also randomly appeared to them. The results in Table 2 show that the TN approach has less distortion and more naturalness but slightly less tone intelligibility than the WH approach.

3.2 Experiment 2

From section 3.1, the syllables with tone errors showed high F_0 distortion. Moreover, the predicted F_0 mean was found to be predicted more accurately than F_0 min and F_0 max. Therefore, instead of using F_0 max and F_0 min, F_0 mean and F_0 range were used to calculate the absolute value of the F_0 contour from the normalized tone template. Since many researches found that the F_0 value relates to the syllable duration, the parameter prediction was changed to dependently sequential approach instead of independently parallel approach. The time-related parameters were estimated, then added to predict the F_0 -related parameters. Besides that, the 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 our assumption that tone shape in the previous tone relates to the variation of the current tone shape, therefore, 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

Figure 3: A part of the generated contours from TN (dash line) and MTN (dot line) approaches comparing to target contours (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.

predict F_0 mean, F_0 min, F_0 max, and F_0 range. When predicting the tone template identity, the predicted nucleus duration and preceding tone template identity were also included in the question set. This modified method will be refined to as MTN approach.

It can be clearly seen from Figure 3 that the contour generated by the MTN approach are much similar to the target contour than the one generated from the TN approach. 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 results in Table 2 also reveal the improved tone generation accuracy, less F_0 distortion and more naturalness.

To evaluate the relative improvement of the MTN approach to the TN approach, the same listeners were asked to listen to the synthesized speech from the TN approach and the one from the MTN approach, respectively and give Comparison Mean Opinion Score (CMOS) on how much the sound from MTN approach is more natural than the other from the TN approach with the following rates: 2 (much better), 1 (slightly better), 0 (about the same), -1 (slightly worse) and -2 (much worse). The average naturalness comparison score of all 30 utterances are 0.449 with standard deviation of 0.715. This shows that the modified method improves the prediction results to yield more naturalness of the synthesized speech and decrease tone errors.

4. Conclusions

The tone nucleus model was adapted 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 nucleus as done in [6], the objective and subjective evaluations confirm the efficiency of the model in Thai in terms of tone intelligibility and naturalness.

References

- [1] N. Thubthong and B. Kijirikul: 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. Charnvitt: Perception of tone and vowel quantity in Thai, *Proc. of ICSLP 2002*, pp. 753–756, 2002.
- [3] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol and C. Wutiwiwatchai: Optimization of t-tilt f_0 modeling, *Proc. of Interspeech 2009*, pp. 508–511, 2009.
- [4] J. Zhang and K. Hirose: Tone nucleus modeling for Chinese lexical tone recognition, *Speech Communication*, Vol. 42, Nos. 3–4, pp. 447–466, 2004.
- [5] M. Wen, M. Wang, K. Hirose and N. Minematsu: Prosody conversion for emotional mandarin speech synthesis using the tone nucleus model, *IPSI SIG Notes*, Vol. 2011, No. 2, pp. 1–6, 2011.
- [6] Q. Sun, K. Hirose and N. Minematsu: A method for generation of mandarin f_0 contours based on tone nucleus model and superpositional model, *Speech Communication*, Vol. 54, No. 8, pp. 932–945, 2012.
- [7] N. Thubthong, B. Kijirikul and S. Luksaneeyanawin: An empirical study for constructing Thai tone models, *Proc. of the 5th Symposium on Natural Language Processing and Oriental COCODA Workshop 2002*, pp. 179–186, 2002.
- [8] Y.N. Hu, M. Chu, C. Huang and Y.N. Zhang: Exploring tonal variations via context-dependent tone models, *Proc. of Interspeech 2007*, pp. 1030–1033, 2007.
- [9] P. Taylor, R. Caley, A.W. Black and S. King: Edinburgh speech tools library–wagon (1.0.1 edition), http://www.cstr.ed.ac.uk/projects/speech_tools/, 1998.
- [10] C. Hansakunbuntheung, V. Tesprasit and V. Sornlertlamvanich: Thai tagged speech corpus for speech synthesis, *The Oriental COCODA 2003*, pp. 97–104, 2003.