Generation of fundamental frequency contours for Thai speech synthesis using tone nucleus model

声調核モデルによるタイ語音声合成の基本周波数パターン生成

by

48-116420
Ms. Krityakien Oraphan

Supervisor: Professor Hirose Keikichi

MASTER THESIS

Presented
to
the Graduate School of Information Science and Technology
The University of Tokyo

in Partial Fulfillment of the Requirements
for the Degree of
Master of Information Science and Technology

Department of Information and Communication Engineering
The University of Tokyo

February 2013

Acknowledgements

This thesis would not be possible if I did not have invaluable support from these following people. Firstly, I would like to express my gratitude and pay a respect to my advisor, Prof. Hirose Keikichi, for his precious time, unconditional advices, worth discussion, and supports throughout my study. Without him, studying in Japan would be very tougher. He has given the great opportunity to me for coming to study in the University of Tokyo without any difficulties. He is very nice, open-minded and humorous. I really appreciate his kind advices. Moreover, I would like to give gratefulness to Prof. Minematsu Nobuaki, who always gives very useful comments, interesting advices and valuable questions to my research in every meeting. He dedicates his personal life to the research work. His efforts, attentions and guidance are really appreciated. I have gained various knowledge and methods which I hardly find in my home country.

Secondly, I would like to pay my thankfulness to Panasonic Scholarship co., Ltd and Panasonic Management (Thailand) co., Ltd to provide the good opportunity for me with great financial support. Without them, the study would not be smooth as this is. I can pay all my effort to the study without worrying about the financial status.

In this study, I had to deal with challenges in exploring new area of knowledge. A lot of time and effort have been paid on the experiments both successful ones and unsatisfied ones. I would like to pass my thankfulness to the people who did help me for their collaboration in data allocation and experiments. With regard to the speech corpus used in the study, I would like to gratefully refer to Human Language Technology Laboratory (HLT) at National Electronics and Computer Technology Center (NECTEC) in Thailand who promptly provided the Thai Speech Corpus for Synthesis (TSynC-1) with free of charge. I also would like to give a big thank to Greg Hamerly, an associate professor of computer science at Baylor University, the united states of America, who provided PG-means source code without restriction although it is finally not included in this study. I would like to thank all experiment participants who are in Japan and Thailand to help do the subjective evaluation tests.

In addition, I would like to thank all former and current members in Hirose and Minematsu laboratory especially, to Mr. Saito Daisuke for his enthusiastic help, Mr. Masayuki Suzuki for his lively questions, Ms. Miaomiao Wang and Ms. Miaomiao Wen for their online supportive advices, Mr. Greg Short for his advices, and also the other members for all supports, discussions and warm friendships.

Finally, I would like to express my gratitude to my family who always gives unconditional supports throughout my study life. I also would like to thank my friends both in Japan and Thailand who always give encouragement, listen deeply and share precious experiences.

Abstract

In this information decades, speech media is one of the new coming interfaces between human and machines. Applications with this interface help users to access information while they can continue their front tasks. Not only speech recognition but speech synthesis has been also introduced and embedded in such applications. However, the users prefer the synthetic speech with intelligibility and naturalness regardless of how many other abilities the application provides. The speech synthesis for tonal languages is much more challenge than that for non-tonal languages, because both intonation and tones need to be concerned.

Fundamental frequency is one of acoustic features relating to the intonation and tones. Existing F0 models for Thai language are expensive to complete the F0 generation from their parameters and suffer when the size of the available data to build the model is small.

With many advantages of the tone nucleus model which has been originated in Mandarin, we have pioneered adapting this model in Thai language to meet the classic but still intrinsic requirements of speech synthesis in continuous speech. Tone nuclei are analytically defined for all five distinctive Thai tones according to their underlying targets. The full process of the F0 contour generation is presented from the tone nucleus extraction, parameter extraction, parameter prediction, until the F0 contour generation for the continuous speech.

Again, the model is successfully proven to be adapted in the other language than Mandarin through objective and subjective tests. The tests confirmed the efficiency and adaptability of the model. Compared to the F0 contours generated by the predictors trained from the contours in the whole syllables without extracting the tone nuclei, the model generated the F0 contours in continuous utterances with less distortion but more tone intelligibility and naturalness. Proposed methodology in parameter prediction and the F0 contour generation processes improved the quality of the synthetic speeches by reducing the distortion and increasing the tone intelligibility and naturalness significantly.

Table of Contents

List of Figures

List of Tables

1 Introduction

As part of the information era, the information accessibility is necessary and widely concerned. Speech technologies have been introduced to help human being access the information more conveniently, such as many applications on cellphones also provide a speech interface to users. By its advantage, this kind of interface can help the users access the information without leaving their main activities. Originally speech is one of the most fundamental media between human and human. Currently, it also becomes one of the media between human and machines to exchange information. Speech synthesis technology can make the machines speak out by generating sound signals to represent human speech. Generally the synthetic speech needs to meet 2 intrinsic requirements: it needs to deliver correct contents and it needs to be as similar and natural as human speech. Many users of such applications can accept the robotic-sounding synthesized speech but they still prefer natural speech. Most users are extremely intolerant to unnaturalness and finally they will refuse to use those non-natural-sounding systems regardless of what other benefits they can obtain. Therefore, the speech synthesis system needs to meet the intelligibility and human acceptance on naturalness of the synthetic speech.

In this chapter, we introduce the motivating importance and the difficulties of F0 contour generation in continuous speech synthesis for tonal languages. The main specifications of the F0 modeling are then addressed. We then introduce the general goal and approach of this thesis. Finally, we give an overview of each chapter in the thesis and its organization.

1.1 MOTIVATION AND BACKGROUNDS

Intonation is one of main factors to control the naturalness of the synthetic speech. Human can receive the intonation and prosodic information by pitch in the sounds they heard. In acoustic aspect, the fundamental frequency buried in the speech signal is the physical quantity we use to represent the perceptual pitch. The fundamental frequency is referred as “F0” among speech researchers.

In tonal languages[1], F0’s movements in the temporal dimension, namely F0 contour, are widely used to define and discriminate tone types. The F0 contour in a syllable uttered in isolation shows a very stable pattern while it changes drastically with complex variations when uttered in spontaneous and continuous speech due to various factors. However, the listeners can still perceptually realize these tone contour variations to the same tonal information. The challenge of the speech synthesis system in such languages is that not only the high intonation level should be concerned but also the lower level itself, which is the local pitch contour representing tone, have to be taken into consideration intrinsically to guarantee that the system meets the fundamental requirements.

Due to the complexity of the human speech, it has been suggested to analyze the F0 contour by modeling, which employs parametric representations of the attributes that are

[1] In tonal languages, each tone bearing unit has an inherent pitch contour. Words which have the same phoneme sequence but different tones are treated as different words in the tonal languages.

believed to be important in the particular task. Consequently, the F0 modeling has been significantly introduced for improving the naturalness of the synthetic speech in speech synthesis system especially, text-to-speech synthesis system and later for wider area of applications, e.g., emotional conversion, voice conversion, intonation synthesis, etc. There are varied approaches of many modelings which have been proposed. Many factors are significant to build and choose the model. They are, for instances, considering unit size (e.g., syllable or phrase), characteristic of the languages (e.g., tonal language or non-tonal language), etc. However, the most important concern when designing or selecting the model is the purpose of the usage. If our goal is to only represent and substitute the F0 contour by a set of parameters, we can straightforwardly use some polynomial functions to fit the contour in order to reduce very huge amount of raw data. But if we need to use the model for speech generation, the model has to be able to practically predict the parameters and these parameters should be linguistically meaningful.

In this study, the F0 contour generation for Thai language is focus. Thai is one of the tonal languages. However, as a local language which is used by the limited number of speakers, researches and developments related to the F0 contour generation in Thai language have been carried out not so many comparing to the other popular tonal languages e.g., Mandarin and Cantonese. The process of the speech synthesis system for Thai language could be generally similar to those for non-tonal languages, e.g., English, but it may be implemented in different ways to deal with the special features related to tones.

Hidden Markov Model (HMM), a statistical model, is widely used in speech synthesis because of its flexibility on various types of synthesized speeches. However, the lexical tones in Thai are considered to be distinguishable in a syllable-long-unit. This leads to some unnatural pitch pattern when that contour has been generated by the HMM based model because the model typically considers F0 values in very short interval of frame-by-frame. The generated F0 is dependent only among the nearby frames. Therefore, it has been suggested that to increase naturalness of the generated speech in the tonal languages, the F0 contour generation of each tone should be operated in a longer time interval.

Previous researches about tones in Thai language have introduced many tone modelings to fulfill the requirement on how to generate the speech with good quality and high tone intelligibility. Many [1] [2]have paid attention to the generation process model (known as Fujisaki model [3]), but most of them are still in the analysis phrase because of the expensive and unreliable parameter extraction from the observed speech utterance. T-Tilt model [4] [5] is also a parametric model adapting the Tilt model [6] to analysis and synthesize F0 contour in tonal languages. It is efficient for F0 analysis but requires extensive work in generation. Also, it needs many parameters to model the F0 contour and it is weak in predicting parameters from the linguistic information.

The tone nucleus model [7], which has been originated to apply in Mandarin, is a data driven based model. It suggests that for the F0 contour in each syllable, there is only a part that significantly conveys the tonal information while the others are just transitory contours which generally contribute sparseness and noise data. By considering only this part, F0 contour which is less affected by its adjacent syllables can be obtained to further apply in various tasks. The validity of the model has been presented by many applications for Mandarin language: tone recognition [7] [8], F0 contour generation [9], and prosody conversion [10]. Furthermore, the model requires only compact size of the data while the HMM-based model requires much more data for the better output speech quality.

1.2 THESIS GOALS

By the advantages of the tone nucleus model, we have been motivated to apply the model to Thai language to meet the intrinsic requirements of continuous speech synthesis in tonal languages. To the best of our knowledge and literature surveys, this work is among the first that has applied the tone nucleus model as the F0 modeling in other tonal language rather than Mandarin.

Based on the above general goal, we can specify what we expect to accomplish as following tasks:

- Defining the tone nucleus model which is applicable to Thai language practically.
- Using the tone nucleus model to generate the F0 contours in continuous speech with tone intelligibility and naturalness.

1.3 ORGANIZATION OF THE THESIS

This study mainly focuses on the data-driven based F0 contour generation for Thai language using the tone nucleus model. All processes are presented from the very beginning until the speech synthesis, sequentially.

The remainder of the thesis is organized into 5 chapters. First, the literature reviews corresponding to the F0 modeling in tonal languages is introduced in Chapter 2. The background knowledge about nature of Thai tone was introduced in Chapter 3 to provide cues in designing tone nuclei for Thai language. Thai tone nuclei characteristics are also described in the same chapter. Next, to generate the F0 contour for continuous speech using the model, the model parameters were defined for further production step in Chapter 4. We showed the parameter extraction process, prediction process and generation, respectively. And then, we revealed some modification in prediction and generation processes to get the better generated contours in Chapter 5. Last but not least, in Chapter 6, conclusion and future work are provided to the readers who are interested in this study.

Chapter 1: Introduction

This chapter introduces the research background, the motivation and accomplished goals of this thesis. To give the overview of the thesis, the short summary of the content in each chapter is provided.

Chapter 2: Literature reviews

This chapter presents various F0 modelings which are basically related to the generating the F0 contour for tonal languages. They are 5 models: unit-selection approach model, Generation process model, T-Tilt model, Hidden Markov model and Tone nucleus model. These models were analyzed by their advantages and disadvantages.

Chapter 3: Tone nucleus model in Thai language

To define how tone nuclei in Thai language look like, background knowledge about nature of Thai language, and Mandarin tones are introduced. We newly introduced the characteristics of each Thai tone nucleus. According to the analysis by modeling, the experiment was conducted to confirm the validity of the defined tone nuclei in Thai.

Chapter 4: F0 contour generation by the tone nucleus model

We developed a system to automatically detect the tone nucleus by a set of rules adopting the defined tone nuclei in chapter3. As a predictive model, the set of parameters were defined and extracted. The process on how we predicted the model parameters is later presented. The methodology to generate the pitch contour from the predicted parameters is provided. Then, the evaluations of the synthetic speeches were conducted through the objective and subjective tests by comparing to the contours generated by the prediction tree built from the parameters that are extracted from the whole syllable contours. At the end of the chapter, the discussion corresponding to the experiment results and conclusion are presented.

Chapter 5: Modification of the tone nucleus model parameter prediction and the F0 contour generation

To improve the quality of the generated F0 contours, we analyzed the parameter prediction and F0 contour generation processes. The proposed methods are provided accordingly. The objective and subjective tests were conducted. The results are shown at the end of the chapter.

Chapter 6: Conclusions and future work

In this final chapter, we conclude what we have studied on the validity of the tone nucleus model in Thai language. And then, lastly, future work of this thesis is introduced to the readers who desire to extend our study**.**

2 Literature reviews

In tonal languages[2], F0 contours play very important roles not only to convey the intonation of the speech but also to contribute the lexical meaning to the syllables. However, the F0 contour in a syllable uttered in isolation shows a very stable pattern while it changes drastically with complex variations when uttered in spontaneous and continuous speech due to various factors. Because of the complicated F0 contour characteristic, storing all data from the observed F0 contours finally reaches the limit. The parameterization is an alternative to obtain and store only some essential information which can re-generate the contours by those pieces of information. Consequently, the F0 modeling has been introduced in order to improve the naturalness of the synthetic speech in speech synthesis system. There are many approaches to build the F0 modeling. However, the most important concern when designing or selecting the model is the purpose of the usage. In our case, we aim to generate F0 contours with high quality and naturalness.

In this chapter, various F0 modelings are introduced. They are 5 models: unit-selection approach model, Generation process model, T-Tilt model, Hidden Markov model and Tone nucleus model. The motivation and conclusion why we selected to apply the tone nucleus model in this study are presented at the end of the chapter.

2.1 UNIT-SELECTION APPROACH MODEL

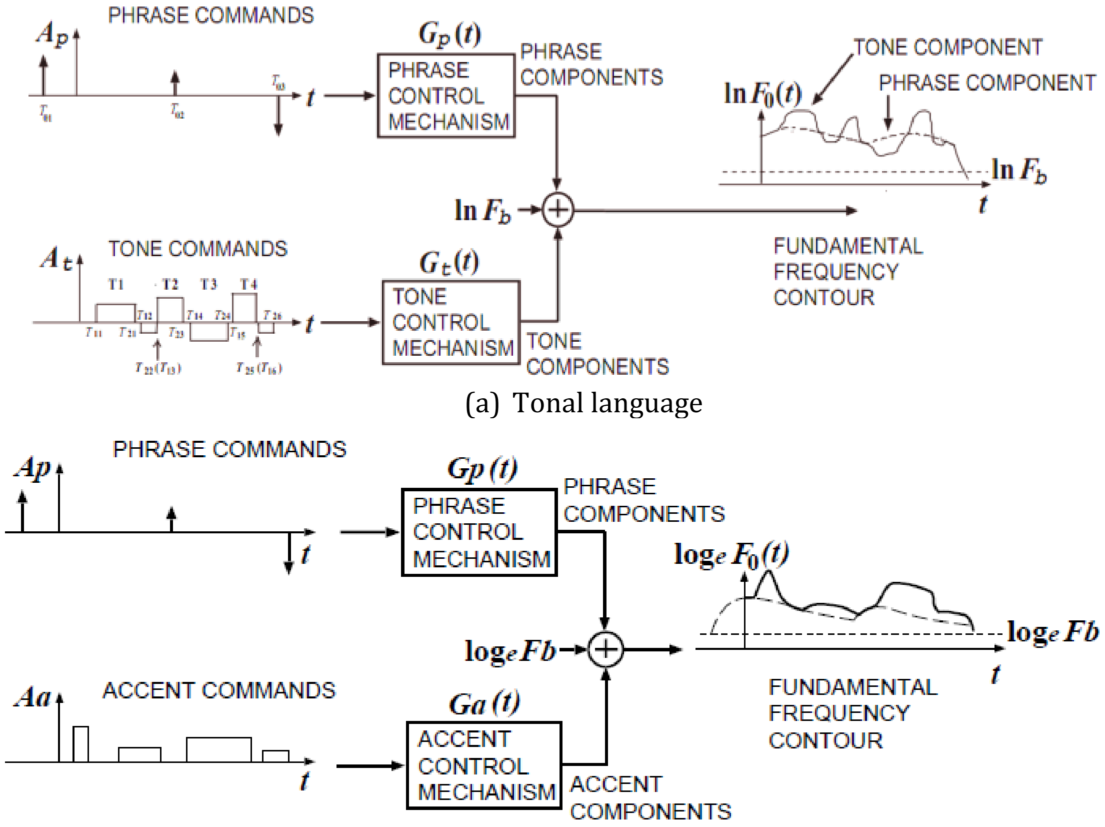
The unit selection approach has been first introduced when power and storage capacity of computers and the efficient searching techniques were sufficient enough. This approach uses a concept of concatenating synthesis. Through a selection method, the most suitable chain of units is determined for any given target sentence. If the suitable units are available in the recording database, no or just little signal processing is needed, and the resulting synthesized speech can sound so natural that it is difficult to distinguish the synthetic speech from the real human speech. This means that if the exact target unit exists in the database, we can get the very natural synthetic speech; on the other hand, if there is not a good-enough-unit to be matched, the result is very poor. Hence, the weakness lies on its inflexibility.

2.2 GENERATION PROCESS MODEL (FUJISAKI MODEL)

Generation process model or Fujisaki model [3] is a mathematical model consisting of a set of physiologically and physically meaningful parameters which describes the process generating F0 contours by vocal folds in a reasonably simplified form. It describes F0 contours in the logarithmic scale as the linear summation of global phrase components, local accent components and a base line level shown in Fig 1. The phrase command and a base F0 level line produce a baseline component while the accent command is added up from the phrase command. Fig 1 also shows one of differences of the models between tonal languages (a) e.g., Mandarin, Thai and non-tonal languages (b) e.g., Japanese. In tonal languages, local command is

[2] In tonal languages, each tone bearing unit has an inherent pitch contour. Words which have the same phoneme sequence but different tones are treated as different words in the tonal languages.

called "tone command" which can be a combination of positive and negative polarities in a syllable-long unit, whereas, non-tonal languages especially in Japanese has only positive command polarity and it is called "accent command".



(a) Tonal language

(b) Non-tonal language

FIG 1: FUNCTIONAL MODEL FOR GENERATING F0 CONTOURS IN FUJISAKI MODEL

In this model, an F0 contour can be expressed by

$$\ln F0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} \sum_{k=1}^{K(j)} A_{a,jk} \{ G_{a,jk}(t - T_{1jk}) - G_{a,jk}(t - T_{2jk}) \}$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-i\alpha t} & if\ \ t \geq 0; \\ 0 & if\ \ t < 0; \end{cases}$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)]\, e^{-\beta t}, \gamma] & if\ \ t \geq 0; \\ 0 & if\ \ t < 0; \end{cases}$$

where $G_p(t)$ represents the impulse response function of the phrase control mechanism (phrase commands)
$G_a(t)$ represents the step response function of the accent control mechanism (tone or accent commands)
F_b is a baseline value of the fundamental frequency
I is the number of phrase commands
J is the number of accent/tone commands
$K(j)$ is the number of components in the jth accent/tone command
A_{pi} is a magnitude of the ith phrase command

$A_{a,jk}$ is an amplitude of the component k^th^ in the j^th^ accent/tone command
T_{0i} is timing of the i^th^ phrase command
T_{1jk} is onset of the component k^th^ in the j^th^ accent/tone command
T_{2jk} is offset of the component k^th^ in the j^th^ accent command
α is a natural angular frequency of the phrase control mechanism
β is a natural angular frequency of the accent control mechanism
γ is a relative ceiling level of accent/tone components

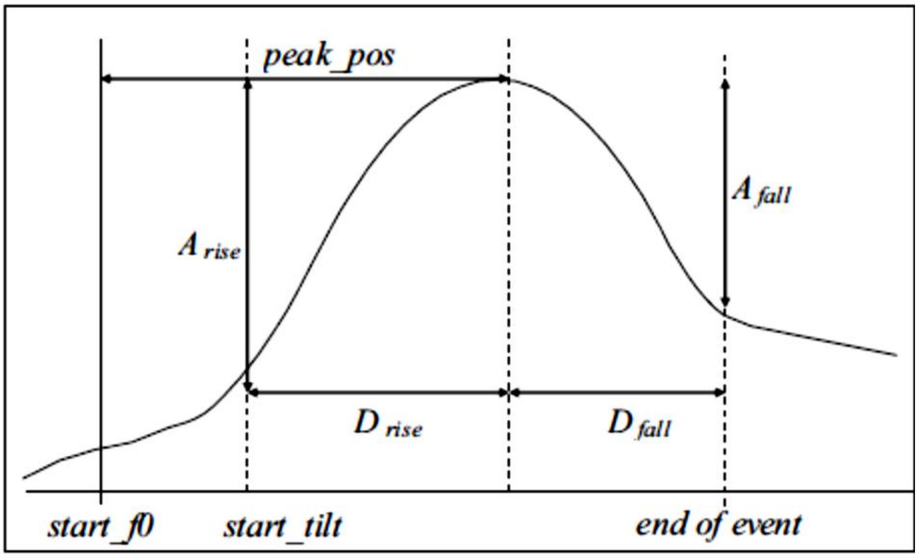
In tonal languages, β and γ can vary with the polarity, onset and offset of the tone command. Also the baseline frequency is speaker dependent and it can vary slightly from an utterance to another even those are spoken by the same speaker. In addition, the values of each parameter are varied and depend on the environment of analysis. Also it need proper configuration by researchers with knowledge.

The tone components in tonal languages are not always positive but can be positive or negative according to the F0 contours of the local tones. Consequently, the syllables which contain at least one tone commands need more parameters to generate the contours. In Thai language, many studies analyzed the parameters of this model in various aspects but a few studies [11] used this model to generate F0 contours with restriction because of the very expensive time consuming in extracting the parameters.

2.3 TONAL TILT MODEL (T-TILT MODEL)

Tonal Tilt model (T-Tilt model) is a low level F0 model adapted from the conventional Tilt model [6] in order to analysis F0 contours in tonal languages. The Tilt model which has been successfully designed for intonation modeling was extended to cover syllable-based F0 contour. The T-Tilt model consists of various parameters to forming the F0 contour of a syllable. Listed below and in Fig 2, the parameters are described.

- *start_f0* : the F0 at the starting point of the syllable
- *start_*tilt : the starting time of the Tilt in the syllable
- *event_ amp* : the summation of absolute rising (A_{rise}) and falling (A_{fall}) amplitudes (negative for the valley F0 shape)
- *event_dur* : the summation of rising (D_{rise}) and falling duration (D_{fall})
- *tTilt_amp* : the difference of rising and falling amplitudes divided by their summation
- *peak_pos* : the duration distance between the starting point of the syllable to the peak of the Tilt
- *shape_type* : types of F0 shape (8 types by considering the F0 curve which every tone can have this kind of shape)
- *tTilt_dur* : the difference of rising and falling duration divided by their summation



-

FIG 2: PARAMETERS IN T-TILT MODEL

The model uses following equations to create a hill-shape-F0 contour at time *t*.

$$f_o(t) = A_{abs} + A - 2A\left(\frac{t}{D}\right)^2 \qquad 0 < t < \frac{D}{2}$$

$$f_o(t) = A_{abs} + 2A\left(1 - \frac{t}{D}\right)^2 \qquad \frac{D}{2} < t < D$$

The below equations are used to create F0 contour at time *t* with valley shape.

$$f_o(t) = A_{abs} + A - A\left(\frac{t}{D}\right)^2 \qquad 0 < t < D$$

$$f_o(t) = A_{abs} + A\left(1 - \frac{t}{D}\right)^2 \qquad 0 < t < D$$

where A_{abs} is an absolute F0 value at the starting point of the rising or falling curve,
D is rising or falling duration.
A is rising or falling amplitude.

Once the T-Tilt parameter is predicted, the synthesis process is to convert those parameters as following.

$$A_{rise} = \frac{event\_amp(1 + tTilt\_amp)}{2}$$

$$A_{fall} = \frac{event\_amp(1 - tTilt\_amp)}{2}$$

$$D_{rise} = peak\_pos$$

$$D_{fall} = syllable\_duration - \ peak\_pos$$

There are a few studies [4] [5] which applied this model to analyze the F0 patterns in Mandarin and Thai languages and conducted F0 contours generation for Thai. The overall quantitative and qualitative results showed that this model was effective in analysis but still needed improvement in synthesis because the model mainly focuses on the automation of the curve fitting. This leads to impractical F0 contour production from the linguistic features.

2.4 HIDDEN MARKOV MODEL

Hidden Markov model (HMM) is a statistical finite state model which is widely used because of its flexibility in generating various types of outputs. It can generate outputs which do not exactly exist in the training data. The merit of the model is it is general easy to train and allow more complexity with high tolerance to noise.

HMMs have a discrete state space and when we move from one state to another during the generation, the observation probabilities change accordingly. However, the model has hidden states which are actually discrete. Oppositely, the F0 contour is normally continuous. Consequently, sudden change in generated contour is normally introduced when it is moved from one state to another. Besides, the model treats the data in discrete manner, it divides the continuous data such as the F0 contour into tiny data unit, e.g., frame which normally covers only about 10ms. The F0 contours in the tonal languages that have syllable-long-unit to be a tone-bearing-unit, get effect of the sudden change. From the characteristic of the intonation and the F0 tone contour, it is recommended to consider a unit whose duration is appropriately long enough. However, over-smoothed generated contour is also one of the weaknesses found in the model when the model is overly cautious when generating the contour. This also brings unnaturalness to the synthetic speech.

In addition, to generate the F0 contours with high naturalness, it is required a lot of allocated data which should be plenty enough to build the trees, especially in the tonal language, because in such language there are various tone contexts and the complex F0 contour variations. More number of tones means much more training data are required.

2.5 TONE NUCLEUS MODEL

The model has been originated in Mandarin to deal with F0 contours variations in continuous speech [7]. It was originally developed for tone type recognition and later used successfully in speech technology in various applications e.g., the speech synthesis [9] and emotional speech synthesis [10].

2.5.1 THE CORE CONCEPT OF THE TONE NUCLEUS MODEL

The concept of the model is to eliminate the part of the F0 contour which brings less significant tonal information and pay more attention on the portion which carries more critical tonal information to the listeners. The portion which is considered to bring more critical tonal information is named as **tone nucleus**[3]. Each syllable contains up to one tone nucleus. The F0 contour in the whole syllable might include the portion that hardly brings any tonal information but produces the data sparseness problem. Hence, the model suggests controlling only the tone nucleus to convey tonal information instead of directly considering the whole syllable.

[3] Tone nucleus, here, is defined differently to the syllable nucleus.

There are some researches about Thai language that recommend considering only some part of the syllable rather than the whole part in the syllable because of the co-articulation phenomenon. Due to the carry-over effect from the preceding syllable, large F0 perturbation is normally found at the initial consonant, thus the F0 contour in final vocalic part which includes vowel and voiced final consonant was suggested to be used for tone modeling [12]. On the other hand, the effect from the following syllable is able to introduce to the ending of the syllable [13]. With regard to these effects, the model ideally divides the F0 contour in a syllable into 3 segments: onset course, tone nucleus and offset course as illustrated in Fig 3. These segments are defined in detail as follows: [7]

- **Onset course** is a F0 transitions from the preceding syllable target to the onset target of the considering syllable. This segment covers the initial consonant and the transition period of the vowel due to the physiological constraint of human beings.
- **Tone nucleus** is a F0 transitions which has less effect by the adjacent syllables. *Onset target* is located at the beginning of the tone nucleus while *offset target* is at the other end. The F0 contour in this segment keeps the underlying targets of the tone unless it is modified by the higher effects e.g., emphasis, focus, neutralization, and etc. This segment is found in the final vocalic part of the syllable (vowel and vocalic final consonant).
- **Offset course** is a F0 transitions from the offset of the considering syllable to the target of the following syllable. This segment covers the final consonant and the transition period of the vowel to that final consonant.

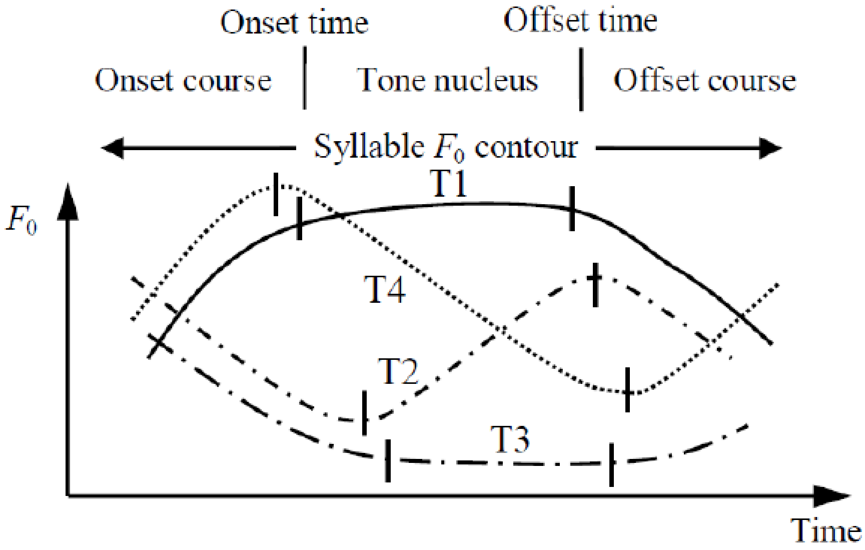


FIG 3: TONE NUCLEI OF FOUR MANDARIN LEXICAL TONES

In Fig 3, the vertical lines on each F0 contour locate the nucleus onset and offset targets. Only the tone nucleus segment is mandatory as its function to convey the critical tonal information, on the contrary, the onset and offset courses are optional. Some syllables do not hold such courses but still convey the tonal information correctly to the listeners. Therefore, it is suggested to take only the tone nucleus into the consideration and ignore the onset and offset courses when analyzing or generating the F0 contour.

2.5.2 THE TONE NUCLEUS MODEL AND TONE RECOGNITION IN MANDARIN

In Mandarin tone recognition, the voiced/unvoiced of the initial consonant brings difficulty in tone recognition. To raise an example, two contours are both consisted of 2 syllables with same tone sequences (T2 follows with T2). One contour (a) has voiced initial consonant in the second syllable where the other (b) has unvoiced initial consonant in its second syllable. Their contours are illustrated in Fig 4 whose the vertical lines are syllable boundaries [7]. The F0 contour in the second syllable of (a) has dipping shape along ABC curve. This contour is wrongly detected to be T3 whereas in (b) it is correctly detected to be T2. When apply the tone nucleus model, the tone nuclei in both cases were detected as BC and give the correct tone recognition results. This example shows that tone nucleus model is very useful because the F0 contour patterns in the tone nuclei are quite stable and consistent while the variations are easily found when considering the F0 contour of the whole syllable.

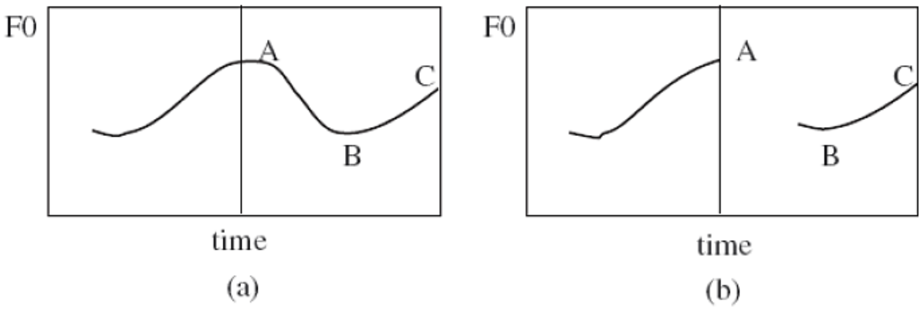


FIG 4: CONTOURS OF THE SAME TONE SEQUENCE BUT DIFFERENT VOICED/UNVOICED INITIAL CONSONANTS OF THE SECOND SYLLABLES

2.5.3 THE TONE NUCLEUS MODEL AND F0 CONTOUR GENERATION IN MANDARIN

In Mandarin speech synthesis, the tone nucleus model was also applied with the generation process model [3] to generate the F0 contour of the utterance [9]. The tone nucleus model was mainly applied in tone contour generation. The superposition of the generated tone contours and the generated phrase contour were performed to generate the output utterance contour. The advantage of adopting the generation process model is that the influence of declination on the extracted tone nuclei is smaller. In generation process, the tone nucleus parameters were predicted by binary decision trees while the phrase components were generated by rules. The synthesized speeches were evaluated to have better quality than the HMM-based method on the same number of training data.

In addition, by controlling the phrase component in the generation process model, the F0 contours with word emphasis were possibly generated with very varied emphasized levels. This shows the possibility to apply the tone nucleus model to control and generate F0 contours especially in the limited number of training data.

Furthermore, the model is applied in emotional conversion [10]. Only the model parameters were changed to convert F0 contour of the neutral speech (emotionless) to the F0 contours with various prosodic styles e.g., angry, happy, and sad. Among the F0 contours of the various speaking styles, the shape of the F0 contours in the tone nuclei of the same syllables are, again, stable and less variation. It is found that the emotional stylization makes an effect on the average pitch and pitch range of the tone nuclei. For instance, the F0 contours of the syllable marked as “da4” in neutral speaking style and angry style in Fig 5 have the similar downward shape of the extracted tone nuclei (F0 contour between the red marks in the syllables) but the angry one have more wider range of F0 value than the neutral one. As a result, instead of directly convert the whole syllable F0 contour; the tone nucleus corresponding parameters were proposed to be converted from the neutral contour into the stylized contour. The other acoustic parameters, which are spectrum and phone duration, were also converted in the research. The result reveals that the tone nucleus model based emotional speech conversion is applicable in specific emotions.

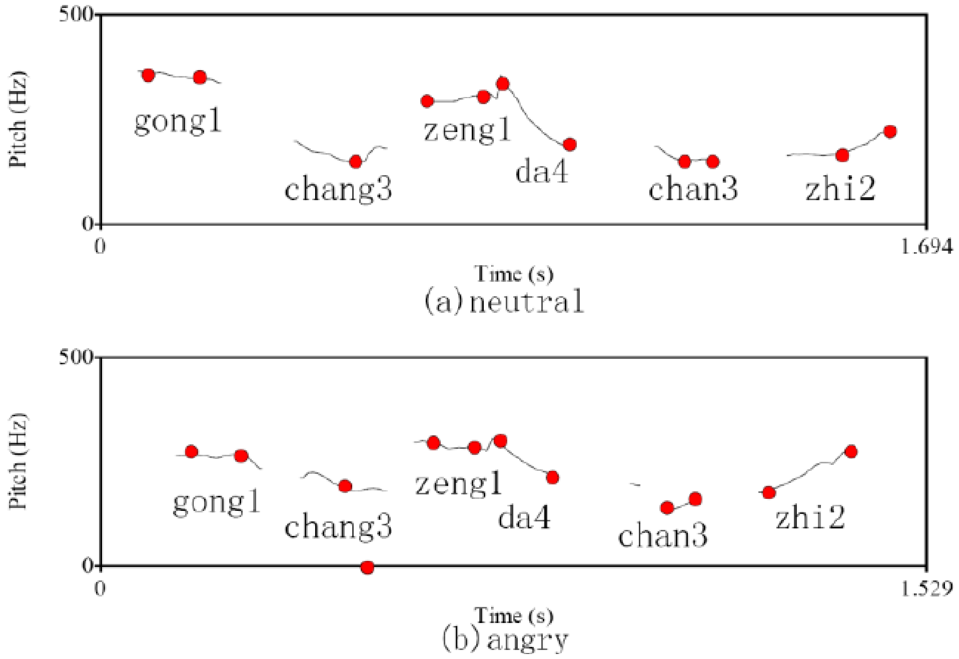


FIG 5: F0 CONTOURS OF THE SPEECH UTTERANCES WITH NEUTRAL (A) AND ANGRY (B) EMOTIONS

To sum up, F0 contours in the tone nucleus segments are more consistent and drawn to match the underlying pitch targets than those in the onset and offset courses. Therefore, the tone nucleus model can be utilized to deal with the F0 contours recognition/generation in various speech applications in Mandarin.

2.6 CONCLUSION

Five F0 contour generating models were introduced. They are unit-selection approach model, Generation process model, Tone Tilt model, Hidden Markov model, and finally the tone nucleus model.

The first approach, unit-selection based approach is not flexible enough to use in the general application, it is practical in small scale of requirements e.g., applications with specific domains which number of the required speech sounds is limited.

For the generation process model, it is very powerful to generate the contours from the available parameters and applicable to various applications. Moreover, its parameters are meaningful to the underlying physiological mechanism. However, the extracted parameters are not reliable because there are some parameters which need the prerequisite knowledge to configure. Also, the parameter extraction is very time-consumed process if there is no provided automatic procedure. Besides, the number of parameters representing all components in an F0 contour utterance in the tonal languages is significantly increased comparing to the non-tonal languages which might degrade the robustness of the model.

The T-Tilt model is a good model for F0 curve fitting because it can operate this task automatically but it needs more extensive work to generate the F0 contour. Moreover, the parameters of the model are also difficult to be predicted from the linguistic data.

For the HMM based model, although the synthetic speeches by this model sound quite natural, but they have strange prosody and they are consisted of mismatched lexical tones. It is because the prosodic features normally cover a wider time span (syllable, sentence or phrase) than segmental features which are usually in frame-by-frame manner. Besides, a lot of training data are essential to construct the model with high intelligibility.

The tone nucleus model is required less data to build, but the knowledge of the particular language is intrinsic. Compared to the other models, the number of the parameters is compact. Besides, the concept of the model relates to the underlying targets. The model concentrates on the important part and ignores the part which is tentative to be noise which brings sparseness. The validity of the model has been proven by the various applications, both recognition and synthesis in Mandarin. Based on the knowledge of the natural characteristic of the language, the algorithm detecting the tone nucleus and procedure in generating the F0 contour is flexible. Moreover, the model can be merged easily to the other models.

3 Tone nucleus model in Thai language

Thai language, the official language of Thailand, is one of tonal languages in which the fundamental frequency (F0) contour or pitch plays important phonemic roles. In speech production, F0 is a frequency of human vocal cords' vibrations whereas in speech perception, human beings can perceive this frequency as pitch. In the tonal languages, F0 does not only express the intonation as in non-tonal languages but it also distinguishes the meaning of lexical words with the same sequence of phonemes.

The tone nucleus model has been initially introduced in Mandarin for tone recognition. Later researches were obviously shown the successful validation of the model to other applications, for instance, F0 contour generation, and emotional conversion with compact set of parameters. This motivates us to adapt the model to generation F0 contour for Thai language. To the best of our knowledge, our research is among the first that applies this model into Thai. However, to achieve our goal, which is to synthesize speech with high quality and intelligibility, we need to define what the tone nuclei in Thai language looks like.

In this chapter, background knowledge about the nature of Thai language is advised to help us understand how to define the tone nuclei in Thai. With the analysis-by-modeling methodology, the preliminary experiment is presented later in the chapter to confirm the validity of the defined tone nuclei in Thai.

3.1 Backgrounds on the standard Thai language

3.1.1 Thai lexical tones

There are five contrasting lexical tones: mid, low, falling, high and rising tones (in short, T0, T1, T2, T3 and T4 respectively). The lexical tone is a crucial feature in Thai because it affects word meanings by distinguishing words that have the same sequence of phonemes. Table 1 depicts an example of the variation of tones over the words with the same sequence of phonemes (/$k^ha:$/) along with their corresponding meanings and Thai textual scripts.

Table 1: Variations of all Thai tones on /$k^ha:$/

IPA	Tone	Thai script	Meaning
$k^ha:$	Mid (T0)	คา	n. “Type of grass” adj. “stuck”
$k^h\grave{a}:$	Low (T1)	ข่า	n. “galangal root”
$k^h\hat{a}:$	Falling (T2)	ข้า	n. “I (tradition word)”
		ค่า	n. “price, value, cost”
		ฆ่า	v. “to kill”
$k^h\acute{a}:$	High (T3)	ค้า	n. “commerce” v. “to sell”
$k^h\breve{a}:$	Rising (T4)	ขา	n. “leg”

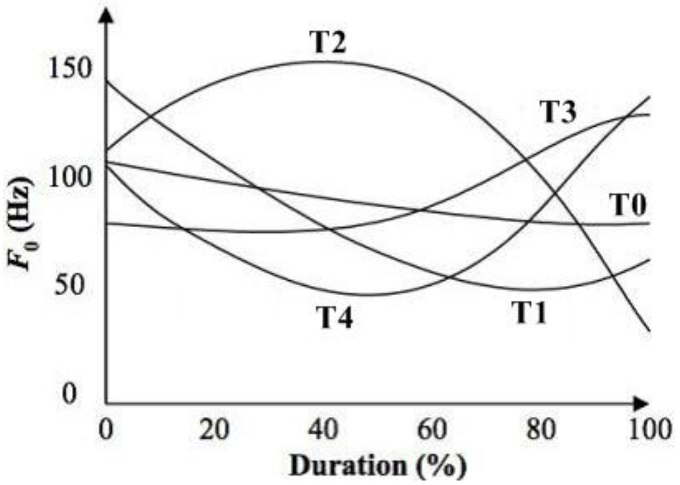


FIG 6 : F0 CONTOURS OF FIVE THAI TONES

All of the tones are characterized principally by their different F0 contour patterns as illustrated in Fig 6. These contours are obtained by referring to the F0 contours extracted from Thai monosyllabic words uttered by a male native Thai [13]. In our work, we assumed these contours in Fig 6 are the underlying targets of all five Thai tones.

Each syllable in Thai is associated with one of five tones. Traditionally, five Thai tones are categorized into 2 types: level tones and dynamic tones (contour tones). Mid, low and high tones are classified as level tones whose pitch contours have relatively static movement and less change. In contrary, dynamic tone is a tone which the movement of pitch is much change. Falling tone has convex pitch shape while the rising tone has concave pitch shape. Therefore, they are considered into dynamic tones by their characteristic. As demonstrated in Fig 7, the conceptual pitch of each tone can obviously discriminate Thai lexical tones.

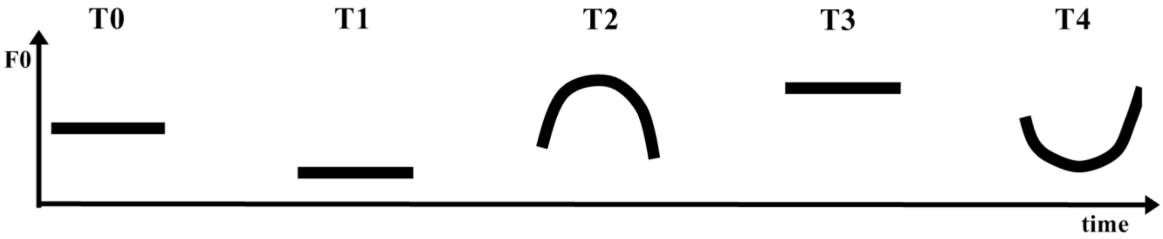


FIG 7 : CONCEPTUAL PITCHES OF FIVE THAI LEXICAL TONES

However, some researchers found that shape of some tones in Thai language can be changed during decades. The study in [14] shows that tones produced by subjects whose age is over 80 years old are different from those produced by younger speakers. The author also claimed that shapes of some tones have been changed and tone grouping scheme which had been proposed by [15] need revising. In addition, [16] studied the tone characteristics change in the new generation of speakers' utterances and also found that some tones e.g., high tone changes from level tone to be more dynamic tone with larger excursion.

Consequently, it has been found and concluded that the two main features used to distinguish and describe these five tones from each other are the F0 excursion and the F0 shape.

3.1.2 THAI CONSONANTS AND VOWELS

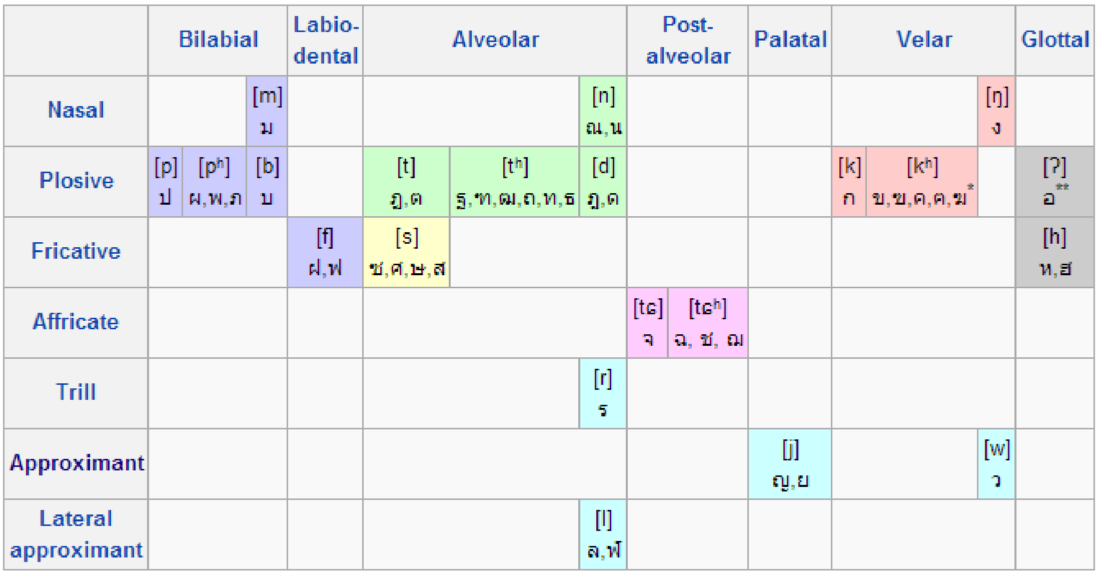
Thai textual syllable structure may be composed of initial consonants (C_i), a vowel (V), final consonants (C_f) and a tone (T) as shown in Fig 8. In Thai, consonants of both initial consonants and final consonants can be grouped roughly in voice and voiceless whereas vowels are always voice.

T
Ci(Ci) V(V) Cf

FIG 8: THAI TONAL SYLLABLE STRUCTURE [17]

The initial consonants can be single consonant and cluster consonants. The cluster consonants are combined from two single consonants with some restrictions that only one of /*k*, k^h, *p*, p^h, *t*/ consonants can be the first element of the cluster initial constants. The second element in the cluster consonants is restricted to be only /*l*/, /*r*/ and /*w*/. However, the combination of these elements is also restricted. In total, they are 21 original Thai single sounds depicted in Table 2 [18] and 11 original Thai cluster sounds (/*kr*/ (กร), /*kl*/ (กล), /*kw*/ (กว), /$k^h r$/ (ขร,คร), /$k^h l$/ (ขล,คล), /$k^h w$/ (ขว,คว), /*pr*/ (ปร), /*pl*/ (ปล), /$p^h r$/ (พร), /$p^h l$/ (ผล,พล), and /*tr*/ (ตร).

TABLE 2: IPA OF ORIGINAL THAI SINGLE INITIAL CONSONANTS

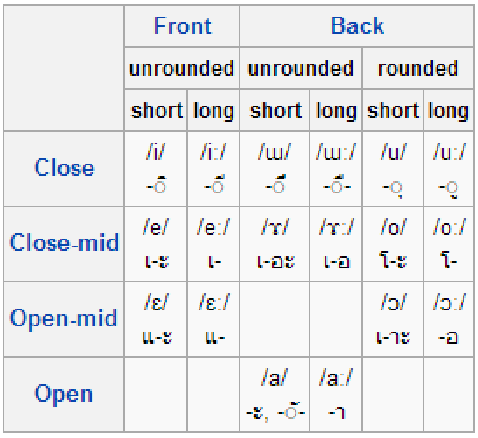


	Bilabial			Labio-dental	Alveolar			Post-alveolar		Palatal	Velar			Glottal
Nasal			[m] ม				[n] ณ,น						[ŋ] ง	
Plosive	[p] ป	$[p^h]$ ผ,พ,ภ	[b] บ		[t] ฏ,ต	$[t^h]$ ฐ,ฑ,ฒ,ถ,ท,ธ	[d] ฎ,ด				[k] ก	$[k^h]$ ข,ฃ,ค,ฅ,ฆ*		[ʔ] อ**
Fricative				[f] ฝ,ฟ	[s] ซ,ศ,ษ,ส									[h] ห,ฮ
Affricate								[tɕ] จ	$[tɕ^h]$ ฉ, ช, ฌ					
Trill							[r] ร							
Approximant										[j] ญ,ย			[w] ว	
Lateral approximant							[l] ล,ฬ							

Basic vowels in Thai can be considered in 9 short-long pairs as depicted in Table 3 by the shapes of lip and tongue. These basic vowels can be combined into 6 diphthongs. The first sub-vowel in diphthongs can be one of /*i, iː, u, uː, ɯ, ɯː*/ and combines with /*a*/ to construct a diphthongs as listed in Table 4.

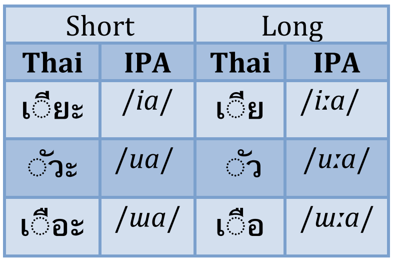
For the final consonant, there are only 9 sounds for original Thai final consonants as demonstrated in Table 5. However, recently, many loanwords have been increasing significantly in Thai language. Much more sounds which do not exist in traditional Thai are added such as /*ft*/ from 'Microsoft'.

TABLE 3: IPA OF THAI MONOPHTHONGS



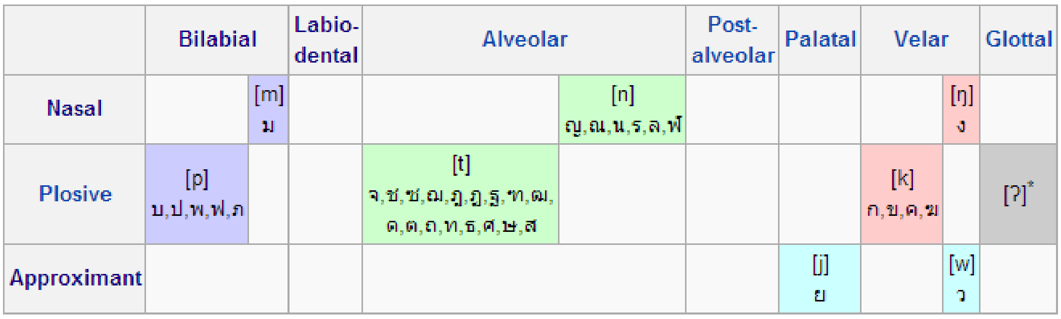
	Front		Back			
	unrounded		unrounded		rounded	
	short	long	short	long	short	long
Close	/i/ -◌ิ	/iː/ -◌ี	/ɯ/ -◌ึ	/ɯː/ -◌ื-	/u/ -◌ุ	/uː/ -◌ู
Close-mid	/e/ เ-ะ	/eː/ เ-	/ɤ/ เ-อะ	/ɤː/ เ-อ	/o/ โ-ะ	/oː/ โ-
Open-mid	/ɛ/ แ-ะ	/ɛː/ แ-			/ɔ/ เ-าะ	/ɔː/ -อ
Open			/a/ -ะ, -◌ั-	/aː/ -า		

TABLE 4: IPA OF THAI DIPHTHONGS



Short		Long	
Thai	**IPA**	**Thai**	**IPA**
เ◌ียะ	/ia/	เ◌ีย	/iːa/
◌ัวะ	/ua/	◌ัว	/uːa/
เ◌ือะ	/ɯa/	เ◌ือ	/ɯːa/

TABLE 5: IPA OF ORIGINAL THAI FINAL CONSONANTS



	Bilabial		Labio-dental	Alveolar		Post-alveolar	Palatal	Velar		Glottal
Nasal		[m] ม			[n] ญ,ณ,น,ร,ล,ฬ				[ŋ] ง	
Plosive	[p] บ,ป,พ,ฟ,ภ			[t] จ,ช,ซ,ฌ,ฎ,ฏ,ฐ,ฑ,ฒ, ด,ต,ถ,ท,ธ,ศ,ษ,ส				[k] ก,ข,ค,ฆ		[ʔ]*
Approximant							[j] ย		[w] ว	

3.1.3 *TONE BEARING UNIT IN THAI*

In phonology unit, Thai tones are normally described in a syllable-long unit although some researches [19] claimed that mora is a suitable tone bearing unit. However, in this work, we consider syllable to be a tone bearing unit according to the concept of the tone nucleus model and the textual structure. Moreover, we consider that the syllable unit is more suitable because the F0 contour in such unit show unique and discriminative characteristic among each other while in mora-unit approach, at most 2 morae are required to describe and discriminate the tone characteristics.

3.2 BACKGROUNDS ON MANDARIN TONES

Mandarin is another well-known tonal language. There are 4 lexical tones and a neutral tone. The lexical tones are high-level (T1), high-rising (T2), low-dipping (T3) and high-falling (T4). Their F0 contours are characterized as illustrated in Fig 9 (adopted from [20]).

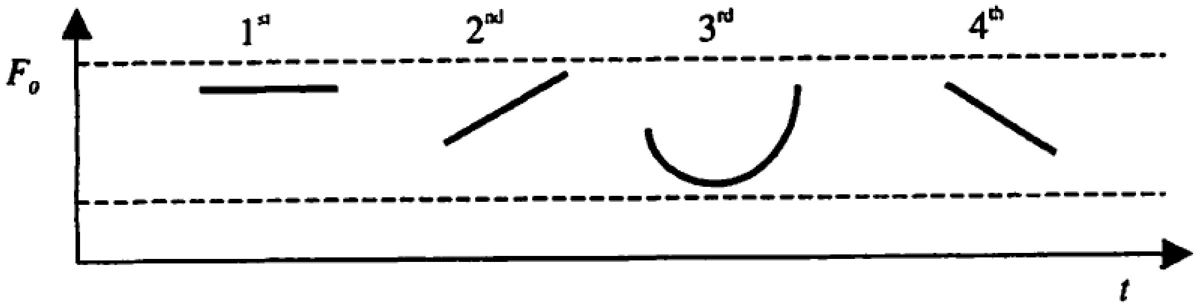


FIG 9: CONCEPTUAL F0 CONTOURS OF FOUR MANDARIN LEXICAL TONES

The neutral tone (T0) has no specific tone pattern and the syllable with this kind of tone is sometimes considered to be lack of tone. Such syllables are normally found at the end of the words or phrases. It is also considered as unstressed syllable and normally pronounced with light and short. Previous researches showed that its pitch depends on the preceding tones.

3.3 CHALLENGES OF TONES IN CONTINUOUS SPEECH

The F0 contour in a syllable uttered in isolation shows a very stable pattern while it is subject to change with complex variations when uttered in continuous speech due to several factors. Principally, these factors are intonation and tonal co-articulation. The intonation effect makes the F0 contour of the utterances decline gradually [21]. The pattern of the F0 contour in a syllable affected by the existence of neighboring syllables is referred to as tonal co-articulation.

3.3.1 DECLINATION EFFECTS

One of common agreements among many researchers is that a gradual declination of a F0 contour across a phrase has been observed. This phenomenon normally effects in long-term downward trend of the pitch height across a typical reading-style, semantic neutral, and declarative utterances. However, if this effect has been overdone, it can express boredom or uncomprehending reading style. Declination can be reset at the utterance or phrase boundaries.

Many researchers have treated this declination as an automatic physiological effect arising from changes in sub-glottal pressure during the course of an utterance, for instance, the Fujisaki model [3] interprets this phenomenon into a rise-fall baseline called "phrase component" in the model.

Referring to Fig 10, the F0 contours of the succeeding syllables are lower than the preceding syllables for all five Thai tones [22]. This leads to the difficulties in both speech recognition and synthesis. In [23], it is also said that due to the declination effect, the mid tone was wrongly recognized to be the low tone because of the change of average F0 level. Likewise, to gain more naturalness of the synthetic speech, this phenomenon needs treating with care especially the prosody synthesis e.g., the conversational speech.

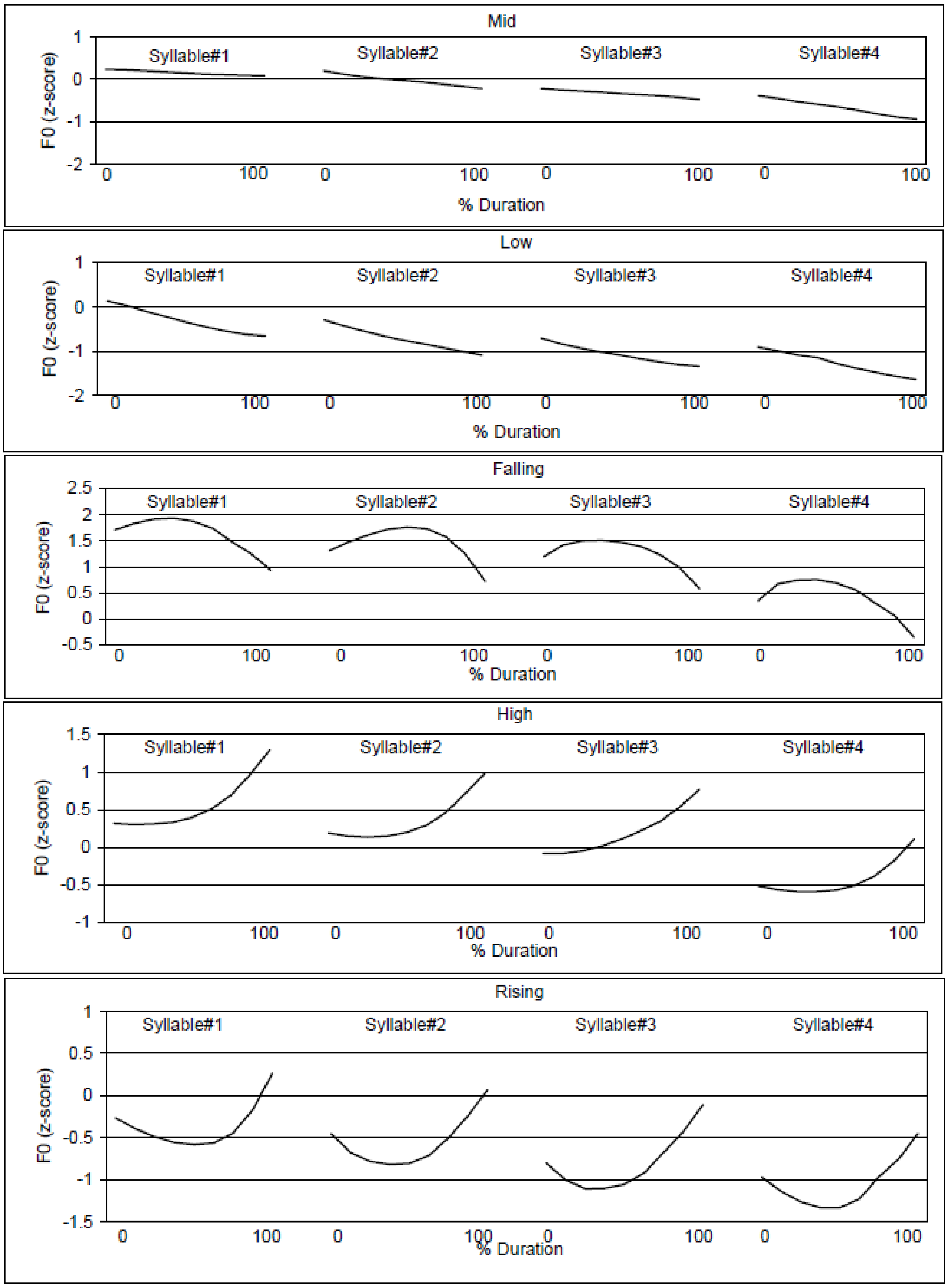


Fig 10: Average F0 contours of five Thai tones at different locations in sentences. From top to bottom: Mid, Low, Falling, High and Rising tones, respectively

3.3.2 Tonal co-articulation effects

In continuous speech, neighboring phonemes and other acoustic features usually interferes the characteristics of the considering unit due to the articulatory constraint of human-beings. Many researches [21] [24] [25] realized some changes in height and slope of the F0 contour as the interferences from the preceding and following syllable tones. Hence, the early portion of the syllable F0 contour tends to vary according to the carryover effect of the preceding tones. This kind of effect is called "carryover co-articulation". In the same way, F0 contour at the end of the syllable is easily changed according to the following syllable. This phenomenon is named as "anticipatory".

The co-articulation effects the F0 pattern change asymmetrically. From a previous research about tone co-articulation in Thai language, although both syllables are stressed, it has been reported that the tonal co-articulation influences the F0 pattern shapes in both syllables. Referring to the study, the F0 patterns in Thai tones were more subjected to change by carryover effect than the anticipatory. Moreover, it was suggested that these effect can be affect only a part of the F0 contour and it is not necessary to influence the whole F0 contour of considering syllable.

In Fig 11 shows interference of following tone which is high tone (*/ch-e-k^-3/*) to the considering tone, the low tone (*/c-aa-k^-1/*). The high tone in the following syllable causes the end of F0 contour of the considering syllable go upward despite keeping downward. This changes the F0 contour pattern in the low tone to be much more similar to the rising tone (T4).

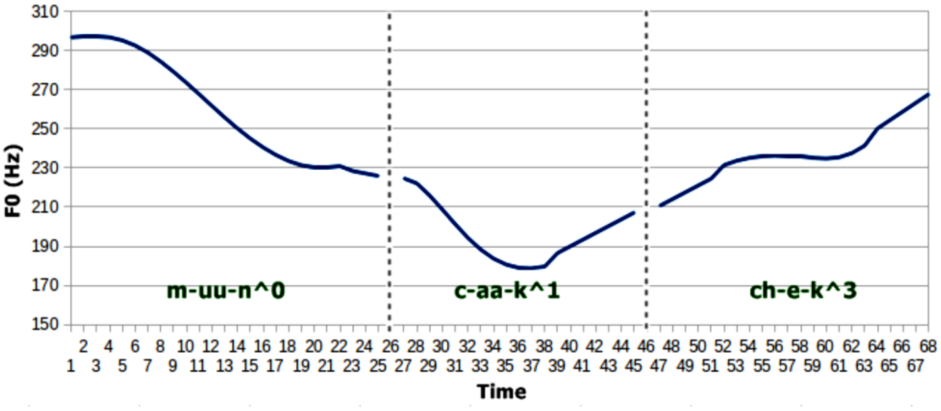


FIG 11: INTERFERENCE OF NEIGHBORING TONES

3.4 DEFINE THAI TONE NUCLEI

In our work, the tone nucleus model has been applied in Thai language for the first time. We need to define what tone nucleus of each Thai tone looks like. From our observation on the characteristic of Thai tones and previous researches across two languages: Thai and Mandarin, the tone nuclei for all five Thai tones were defined mainly based on the underlying target.

3.4.1 DEFINE FROM MANDARIN TONE NUCLEI

The conceptual pitch characteristics shown in Fig 7 and Fig 9 reveal that the tones in Mandarin and Thai are very similar. Some share common characteristics, for instance, the pitch shapes of falling tone (T3) in Thai and high-falling tone (T4) in Mandarin look partly alike. Low-dipping (T3) in Mandarin and low tone (T1) in Thai are perceptually close. As a result, to define the tone nuclei in Thai, we assume that defining the tone nucleus in Thai can significantly refer to the defined tone nucleus in Mandarin.

The research in [26] also confirmed this assumption. They applied one-to-one-functions to map Mandarin tones to Thai tones in order to utilize a Mandarin corpus to synthesize Thai speech utterances. The mapping functions described in Table 6 were obtained by observing the phonological structures suggested in [27]. However, some tones are partially adaptable. As also mentioned in the research, this mapping table should be used with proper consideration, for instances, neutral tone (T0) in Mandarin is different from mid tone (T0) in Thai in some sense and they only share some overlapped characteristics.

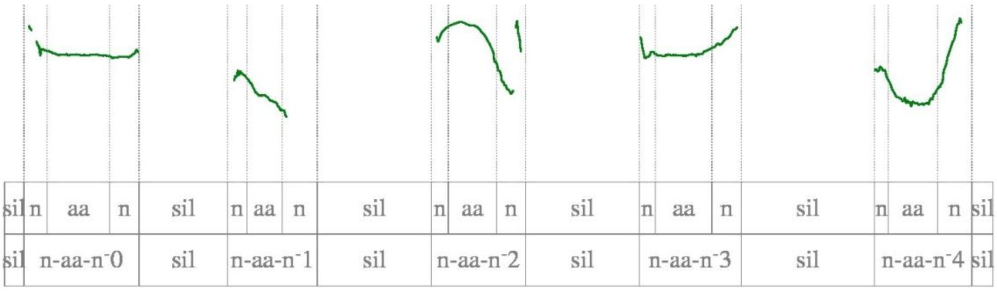
TABLE 6 : MANDARIN AND THAI TONES MAPPING

Thai	Mandarin
Mid (T0)	Neutral (T0)
Low (T1)	Low-dipping (T3)
Falling (T2)	High-falling (T4)
High (T3)	High level (T1)
Rising (T4)	High rising (T2)

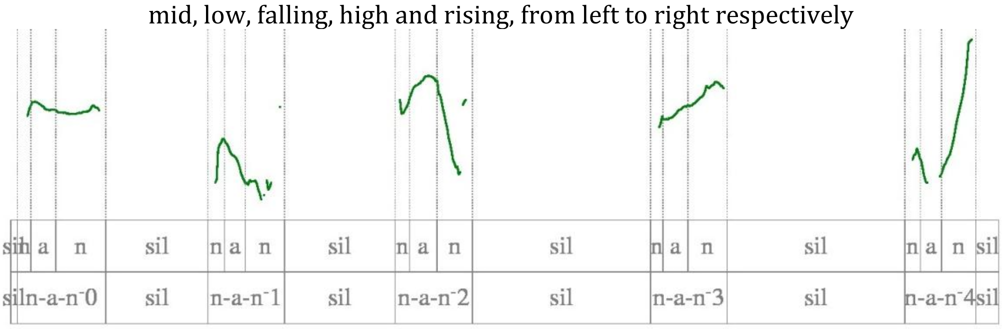
3.4.2 DEFINE FROM THE OBSERVATION

The other way is to observe the F0 contours of the natural utterances. There are 2 sets of natural utterances in our observation. One is consisted of the syllables uttered by the author in citation form; the other is consisted of the continuous utterances uttered in reading style.

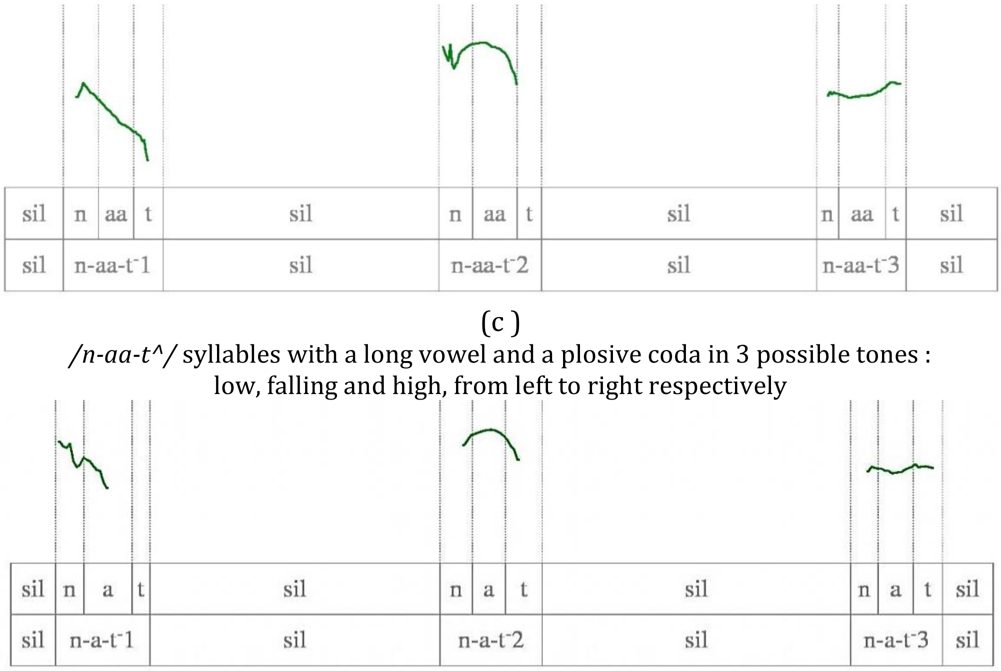
The F0 contours extracted from the first set are shown in Fig 12. Regardless with the vowel duration and the type of the final consonants, they all comply with the underlying target illustrated in Fig 6. Again, we can notice that F0 contours of high tone (T3) do not form in level; they rise up at the end of the syllable and they begin at the same level as mid tone (T0).



(a)
/n-aa-n^/ syllables with a long vowel and a nasal coda in all 5 Thai tones:
mid, low, falling, high and rising, from left to right respectively



(b)
/n-a-n^/ syllables with a short vowel and a nasal coda in all 5 Thai tones:
mid, low, falling, high and rising, from left to right respectively



(c)

/n-aa-t^/ syllables with a long vowel and a plosive coda in 3 possible tones : low, falling and high, from left to right respectively

(d)

/n-a-t^/ syllables with a short vowel and a plosive coda in 3 possible tones : low, falling and high, from left to right respectively

FIG 12: F0 CONTOURS OF THE SYLLABLES UTTERED IN CITATION FORM.

As aforementioned study, the F0 contours in the first set differ from those in running speech in the other set due to the co-articulation effect in the continuous speech. Therefore, the full counterpart contours of the dynamic tones according to Fig 6 are hardly found in the continuous speech, for instance, the convex shape for falling tone (T2) or full concave shape for rising tone (T4). However, they can be found in the prominent syllables of the running speech, otherwise they can be simplified into some counterpart of the F0 contours e.g., only upward for falling tone is able to be observed in the running speech [28] .

Referring to our running speech utterances in the other set were picked randomly from a Thai tagged speech corpus [29]. The syllable "*h-aa-z^-2*" in Fig 13 has falling tone and it is a middle syllable of a prosodic word. Instead of the full convex shape, the falling tone contour is reduced into only an upward rise without the final fall. Also, the "*h-aa-n^-4*" syllable in Fig 14 has rising tone but at the beginning of the syllable is affected by the preceding syllables, so we can observe the downward F0 contour with a very little upward shape or similarly a level contour at the very end of the contour. Therefore, the full counterparts of the dynamic tones are hardly found in the spontaneously running speech.

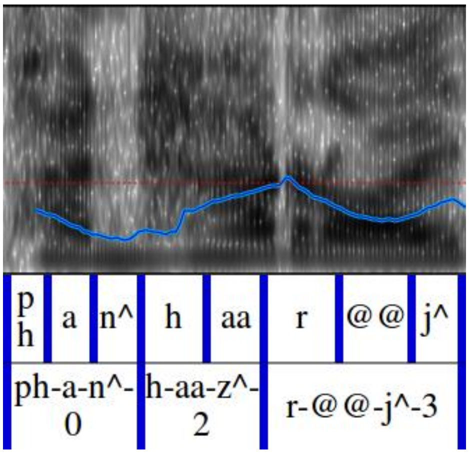


FIG 13: UPWARD F0 CONTOUR FOUND IN A FALLING TONE SYLLABLE (*H-AA-Z^-2*)

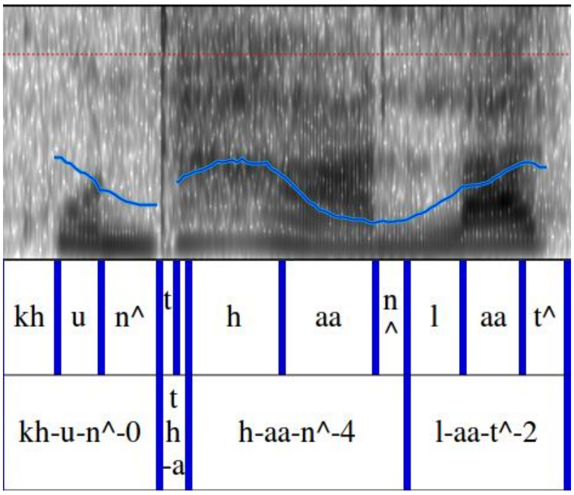


FIG 14: DOWNWARD F0 CONTOUR FOUND IN A RISING TONE SYLLABLE (*H-AA-N^-4*)

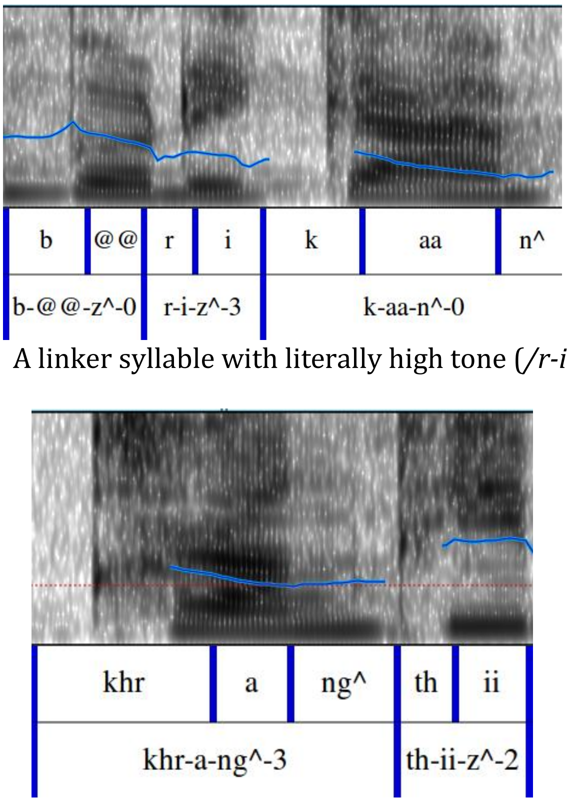
In addition, not only the dynamic tones, the high tone (T3), which cannot be well-defined between the static tone at high pitch and the dynamic tone with downward in the front-half of the syllable and the upward trajectory in the final-half. We observed the F0 contours in the running speech dataset and found that many syllables with high tone have downward trajectories.

There are 2 main possibilities. The first one is that those syllables are linker syllables[4]. Their pitches are usually neutralized to be the mid tone (T0). Fig 15 (a) shows the F0 contour of a lexical word whose the middle syllable is a linker syllable (/*r-i-z^-3*/). Its F0 contour is obviously neutralized to be downward which contradicts to the traditional defined shape.

The other is that the syllables are characterized as a dynamic tone. If we presume that the high tone is a dynamic tone. This might be the same phenomena as in the case of the dynamic tones that has been called "dynamic overshoot" or "peak delay" [30]. The contour is

[4] Linker syllable is a syllable which appears systematically and literally as high tone but is pronounced with weak stress and usually prefixed to the normal syllable. [46]

reduced or the target is postponed to the following syllable. Noticeable evidence in Fig 15 (b) explains this apparently. The syllable with high tone (*/khr-a-ng^-3/*) at the beginning of the phrase has a downward F0 contour which is opposite to the conceptual pitch, although, it is followed by a syllable with falling tone (*/th-ii-z^-2/*) whose the F0 contour characteristic tends to support its preceding syllable to raise the contour up.



(a) A linker syllable with literally high tone (*/r-i-z^-3/*)

(b) A non-linker syllable with literally high tone (*/khr-a-ng^-3/*)

FIG 15: DOWNWARD F0 CONTOURS FOUND IN A HIGH TONE SYLLABLES

3.4.3 SPECIFICATION OF THAI TONE NUCLEI

From a study on Thai tone perception of native Thais [31], both pitch target and contour are significant features to discriminate Thai tones. Only target or contour shape is not sufficient in tonal perception. In our study, we assume that to reach the target point, due to the physiological constraint, the F0 value gradually changes across the time axis until the target is reached. Some claimed that the peak or the valley of the F0 contour in the syllable does not provide the real target which the speaker intends to reach because of the delay of the biophysical mechanism. However, they can be the approximate targets. As a result, the maximal and minimal F0 values in the F0 contour are assumed to be the pitch targets in the syllables.

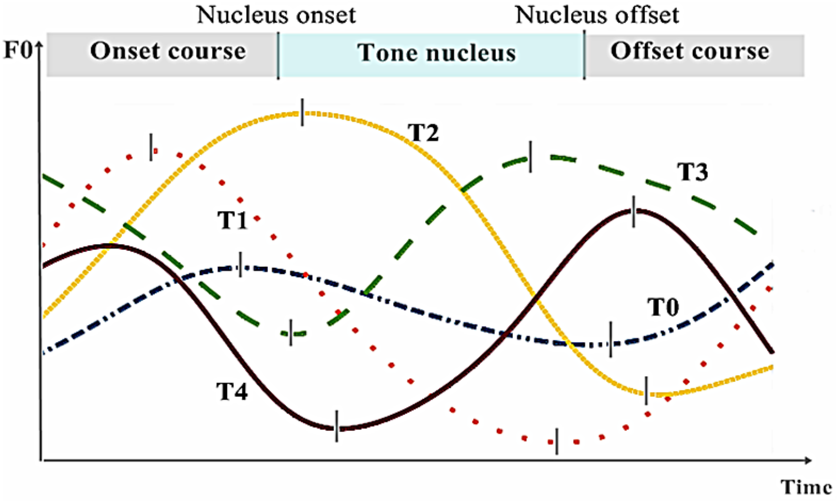


FIG 16: TONE NUCLEI OF FIVE THAI TONES

Our aim is to apply the tone nucleus model for F0 contour generation in the running synthetic speech. Therefore, tone nuclei of the level tones: mid (T0), low (T1), and high (T3) are mainly defined from the observed continuous speech whereas tone nuclei of the dynamic tones: falling (T2) and rising (T4) are additionally defined based on T4 and T2 tone nuclei in Mandarin, respectively. Nevertheless, the tone nuclei of every tone are reserved to follow the underlying pitch targets. Fig 16 shows the tone nuclei defined from our observations. The vertical lines on each F0 contours represent the tone nucleus onset and tone nucleus offset targets.

As a result, we can specify the characteristic of the tone nuclei as in Table 7. Each characteristic in the table is listed according to how it matches the underlying conceptual pitch characteristic from above to below; for instance, high tone (T3), we prioritized upward than downward, so let's say, giving an F0 contour of a high tone syllable with dipping shape (downward followed by upward), the upward part will be assigned to be the tone nucleus because the upward has higher priority than downward.

TABLE 7: F0 CONTOUR CHARACTERISTICS OF THE TONE NUCLEI FOR FIVE THAI TONES

Tone	Tone nuclei characteristics
Mid (T0)	• Level • Gradually downward
Low (T1)	• Downward
Falling (T2)	• Convex (falling) • Upward • Downward
High (T3)	• Upward • Downward
Rising (T4)	• Concave (dipping) • Downward • Upward

Although, the defined characteristics of the tone nuclei in Table 7 shows some contrast variations e.g., in falling, high and rising tones. These variations were also found in several previous researches in Mandarin [9] [10] [32].

3.5 EXPERIMENT AND RESULT

Our experiment was drawn on the basis of the definition of the tone nucleus. Because the tone nucleus is the most significant part that conveys the tonal information, if we adjust the other part but fix the tone nucleus, the modified F0 contour is expected to give the same tonal information. To evaluate the defined tone nuclei in Thai, we applied analysis-by-synthesis technique. From the F0 contour in the natural continuous speech signal, we made some changes on the transitory contour parts and kept the other parts that are assumed to be the tone nuclei fixed.

In our work, the syllable was a considerable unit. The F0 contours were firstly segmented into syllables with initial consonant, vowel, and final consonant boundaries. The tone nucleus was assumed to locate in the final vocalic part; hence, the F0 contour in each syllable was considered over vowels, nasal consonants *(/m/*, */n/* and */ŋ/*) and semivowel final consonants (*/j/* and */w/*). We focused on the onset and the offset targets of the tone nucleus in each syllable by searching the longest contour that matches the characteristic defined in Table 7 with priority concerns. Basically we can search for the maximal and minimal F0 values in the considerable locus, and then expand the searching contour to find the other possible shapes. The onset and the offset targets are later mapped to the detected boundary points corresponding to the defined tone nucleus characteristic of the particular tone.

To raise some examples, Fig 17 shows the detected tone nucleus onset and tone nucleus offset targets in big red circles. The vertical solid and dot lines show syllable boundaries and phone boundaries, respectively. Look at the most left syllable with rising tone (*/s-aa-m^-4/*) shown in the figure. The minimal point can be detected firstly. Because of the rising tone, there are 3 possible shapes representing the tone nucleus: concave, only downward and only upward. If we expand the searching area to the left and the right of the detected minimal point, we will finally obtain the concave shape. Likewise, the tone nucleus of the other dynamic tone (falling, T2) can be also detected. For the high tone syllable (*/phr-@@-m^-3/*), as described earlier, the upward part is detected and it is assigned to be the tone nucleus rather than the downward one.

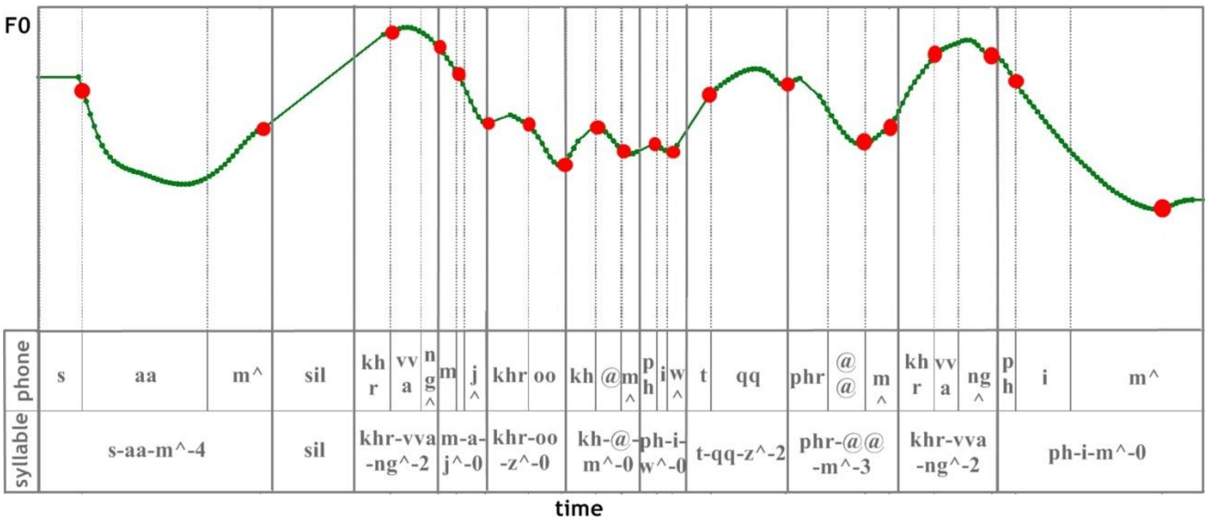


FIG 17: AN F0 CONTOUR OF AN UTTERANCE WITH DETECTED TONE NUCLEI ACCORDING TO THE SPECIFIED TONE NUCLEUS CHARACTERISTICS

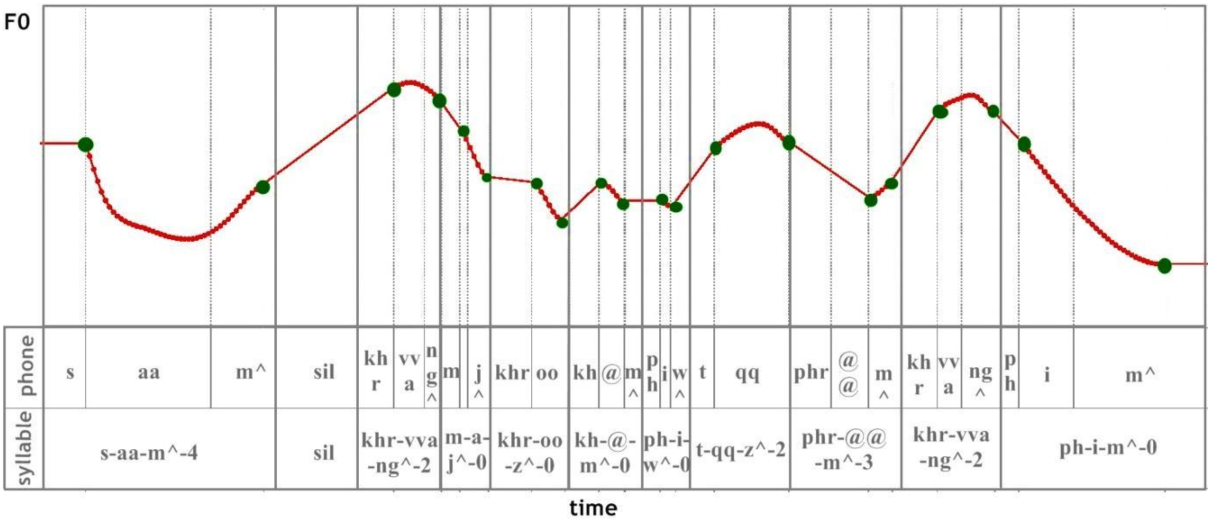


FIG 18: AN EDITED F0 CONTOUR WHOSE TRAJECTORIES BETWEEN TONE NUCLEI ARE REPLACED BY LINEAR INTERPOLATION

After the tone nucleus detecting task, each end of the detected tone nucleus is assigned to be the tone nucleus onset or the tone nucleus offset targets. Those detected tone nuclei were fixed as they are and the transitory parts (the onset and offset courses) were changed by linear interpolation. In short, the detected tone nuclei were concatenated by linear interpolation from the offset target of the previous tone nucleus to the onset target of the following tone nucleus as shown in Fig 18. The tone nucleus targets are illustrated as green big circles.

Lastly, the original F0 contour was substituted by this edited contour and the speech utterance was re-synthesized by TD-PSOLA technique. The preliminarily perceptual test was conducted to evaluate the specified tone nucleus characteristics. Thai natives were assigned to listen to the synthesized speeches and the original speeches. The participants were asked to compare each speech pair whether they convey the same tonal information. The result of the perceptual test indicates that the tonal information the participants perceived is as same as the original one without information distortion.

3.6 CONCLUSION

We aim to define the tone nucleus characteristic of each Thai tone in order to apply the model in the F0 contour generation for Thai continuous speech. Based on the observations and existing knowledge, the tone nuclei of five Thai tones were described on the unified syllable unit. They conceptually match the underlying targets with acceptable variations on falling, high and rising tones. The result of the perceptual test confirms that the defined tone nuclei convey the same tonal information to the native listeners.

4 F0 contour generation by the tone nucleus model

In F0 contour generation approach, the goodness of the model is not only to fit the F0 contour and represent it by a set of parameters but also those parameters of the model should be predictable. In addition, some critical concerns about the modeling should be carefully taken into account. The model should give linguistically meaningful parameters.

Many studies has been using parameters (both, polynomial coefficients and non-polynomial ones) to represent the F0 contour and only utilized them for analysis, not yet in the new F0 contours prediction. Some models [4] [5] suffer in prediction from the linguistic information. The generative model [3] can well represent the underlying physiological mechanism but it is not yet efficient enough for Thai language because the parameter extraction process is very expensive and the automatic parameter extraction has been still under researching stage.

Our study aims to generate the F0 contour of the utterance for Thai language with high quality and tone intelligibility. We have attempted to adapt the tone nucleus model and the defined Thai tone nuclei in section 3.4 in order to generate new F0 contours to assess whether the model is predictively applicable to generate the F0 contours for Thai language.

In this chapter, we developed a system to detect the tone nucleus by a set of rules in chapter 4.1. These rules were basically defined by the analysis in Chapter 3. The model parameters are specified and extracted in section 4.2. In section 4.3 shows how we predicted the model parameters to generate the new F0 contours. The methodology how the predicted results were used to form the F0 contour is presented in section 4.4. Then, the evaluation of the parameter model set on generating new F0 contour utterances were conducted through the objective and subjective tests as described in section 4.6. At the end of the chapter, the discussion corresponding to the experiment results and conclusion are presented.

4.1 Tone nucleus detection

It is necessary to extract the F0 values from the speech signals to perform further tasks. In our work, the F0 values were extracted by Praat speech processing tool [33], [34] with autocorrelation algorithm under the condition in Table 8.

Table 8: Condition of F0 extraction

Configuration	Value
Analysis window	Gaussian window
Time step	10 ms
Pitch floor	120 Hz
Pitch ceiling	600 Hz
Data depth	16 bits/sample
Sampling rate	44.1kHz

It has been a general agreement that the tonal information is not evenly distributed throughout the F0 contour in the syllable. Some segment carries more tonal information than others. Therefore, we need to extract the tone nucleus from the F0 contour of the whole syllable. According to section 3.5 , the tone nucleus was assumed to locate in the final vocalic part; hence, the F0 contour in each syllable was focused only in some parts: the vowel, the optional nasal consonants and semi-vowel final consonants. According to the preliminary observation, some simple rules focusing on the tone targets were applied to detect Thai tone nucleus. These rules are also motivated from the tone nucleus detection method developed for Mandarin speech synthesis [9].

In case of the dynamic tones, the full shapes can be detected by their partial counterparts. The convex trajectory in falling tone can be detected by its reduced forms which share the maximal F0 point. Those are upward and downward parts. Likewise, the concave curve in rising tone can be also reducibly detected by its minimal F0 point and the tone nucleus is represented by one of the partially counterparts. Consequently, being listed below, these rules are mainly related to the optimal F0 points in the final vocalic part.

- For mid (T0) tone, the time index of the minimal F0 is treated as a target point. The tone nucleus covers the segment from the vowel onset to this target point.
    - In case of the minimal F0 is at the vowel onset, the whole contour in the final vocalic part is considered to be a tone nucleus. Set the vowel onset as the tone nucleus onset and the end of the final vocalic part as the tone nucleus offset.
- For low tone (T1), firstly set the minimal F0 as the tone nucleus offset, and then find the maximal F0 of the sub-segment between the vowel onset and the formerly detected tone nucleus offset. The tone nucleus covers the segment from the maximal F0 point to the minimal F0 point.
- For falling tone (T2), the minimal F0 and maximal F0 are treated as target points. The tone nucleus covers the segment between these target points.
- For high tone (T3), the maximal F0 is treated as a target point. The minimal F0 of the sub-segment between the vowel onset and the formerly detected target point is later detected. The latter detected point is treated as another target point. The tone nucleus covers the segment between these target points.
- For rising tone (T4), the minimal F0 and maximal F0 are treated as target points. The tone nucleus covers the segment between these target points.

The detection of the tone nucleus was performed basically by these specified rules. These target points were later assigned to be either onset or offset of the tone nucleus corresponding to its tone type as shown concretely in Table 9.

TABLE 9: TONE NUCLEUS ONSET AND OFFSET TARGETS OF THAI TONE NUCLEI

Tone	Curve characteristic	Tone nucleus Onset	Tone nucleus offset
Mid (T0)	downward	At vowel onset	At F0 Min
Low (T1)	downward	At F0 Max	At F0 Min
Falling (T2)	upward	At F0 Min	At F0 Max
	downward	At F0 Max	At F0 Min
High (T3)	upward	At F0 Min	At F0 Max
	downward	At F0 Max	At F0 Min
Rising (T4)	downward	At F0 Min	At F0 Max
	upward	At F0 Max	At F0 Min

By these rules and our defined Thai tone nucleus model, the level tones can be easily detected. The study of the tone nucleus model [9] reports that it is rather difficult to detect the level tones because the F0 contours of these tones do not give a very stable level line. With the advantage of the Fujisaki model, although the F0 value of the phrase component were excluded from the original contours, the declination effect can still remain in the extracted tone component depending to the tone component extraction process. The F0 contours in the tone components show the non-level contours in the level tone syllables. The well-defined threshold was needed to judge whether that contour can be categorized into the level tones. Differently, in our study, we assigned the downward trajectory to describe the characteristic of the level tones due to the effect of the declination which mention earlier in section 3.3.1 Fig 19 shows the detected tone nuclei (marked with big red circle) by applying these specific rules.

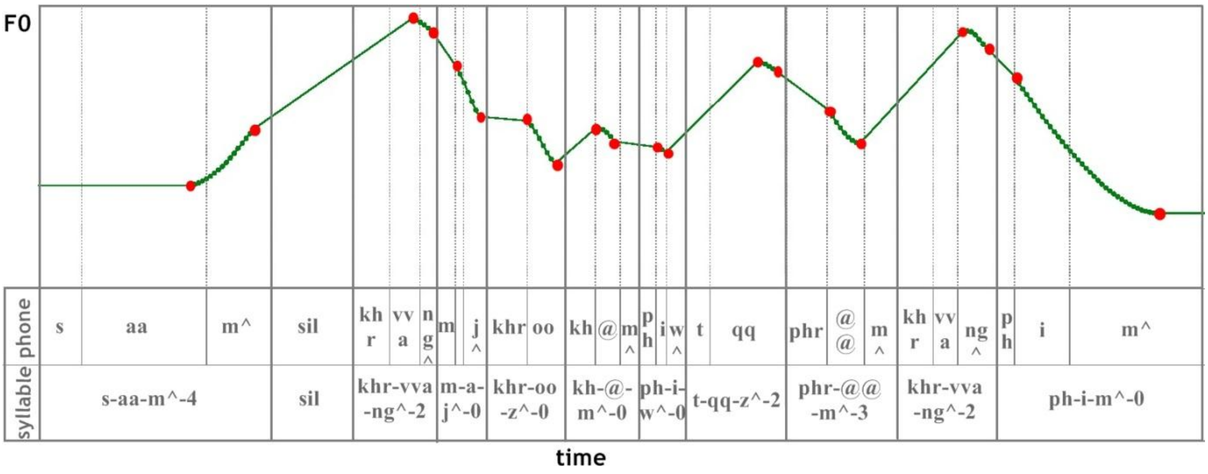


FIG 19: AN F0 CONTOUR OF AN UTTERANCE WITH DETECTED TONE NUCLEI BY THE SPECIFIC RULES

4.2 TONE NUCLEUS PARAMETERS EXTRACTION

In the F0 generation process, these following parameters are essential: TNonset, TNoffset, F0min, F0max, and TNshape.

- TNonset – a time related parameter indicates where the beginning of the tone nucleus is.
- TNoffset – a time related parameter indicates where the end of the tone nucleus is.
- F0min – a fundamental frequency related parameter indicates the minimum value of the F0 in the detected tone nucleus.
- F0max – a fundamental frequency related parameter indicates the maximum value of the F0 in the detected tone nucleus.
- TNshape – a fundamental frequency related parameter indicates how the tone nucleus looks like.

However, the inconsistency may occur during the parameter prediction in the next step, for instance, from the prediction, the TNoffset parameter which indicates the absolute time point in the utterance may be predicted to be less than the TNonset parameter. Also, to describe how the tone nucleus looks like, we used nominal parameters to represent TNshape. Therefore, the inconsistency brings some restrictions to be defined, and then the parameters were adjusted accordingly.

There are 2 types of parameters: time related and F0 related parameters. The time related parameters are TNonset and TNdur. The F0 related ones are F0min, F0range and TNshape. In this work, the time related parameters were relative values to the syllable length, which was fixed to that of the original speech, while the F0 related parameters were measured in log scale. The adjusted parameters used in prediction as follows are listed:

- **TNonset** – a time related parameter indicates where the beginning of the tone nucleus is. This parameter is a relative value to the syllable duration[5]. It is calculated by syllable duration and the syllable onset.

$$TNonset\ (rel.) = \frac{TNonset\ (abs.) - Syllable\ onset}{Syllable\ duration}$$

- **TNdur** – a time related parameter indicates how long the tone nucleus is. By this parameter, TNoffset can be obtained. Also, this parameter is a relative value to the syllable duration.

$$TN\ dur\ (abs.) = TNoffset(abs.) - \ TNonset\ (abs.)$$

$$TNdur\ (rel.) = \frac{TNdur\ (abs.)}{syllable\ duration}$$

- **F0min** – a fundamental frequency related parameter indicates the minimum value of the F0 in the detected tone nucleus. This is directly used from the observed value.

- **F0range** – a fundamental frequency related parameter indicates the excursion of the extracted F0 minimum and maximum values. By this parameter, F0max can be calculated.

$$F0range = F0max + F0min$$

- **TN template identity** – a fundamental frequency related parameter indicates how the tone nucleus looks like. This parameter is nominal. We used natural number to represent the shape. These natural numbers are acquired from the clustering the contours of the extracted tone nuclei.

The extracted tone nuclei are normalized in time dimension into 11 points evenly spaced, and in frequency dimension bounded into 0 to 1 range, where 0 and 1 correspond to minimum and maximum F0 values. To deal with the variations of the extracted tone nuclei, the normalized logF0 and their ΔlogF0 values are represented by a vector. Let O_i represent a vector of the extracted tone nucleus in the i^{th} syllable and the ΔlogF0 can be calculated from the considering syllable compared to the following syllable as shown in below equations.

[5] Syllable duration is measured from syllable onset to the syllable offset. In this case, we refer the syllable onset and offset labeled in the existing corpus. In practice, these values can be obtained from the syllable length prediction or the phoneme prediction.

$$O_i = \begin{pmatrix} logF0_1 \\ \vdots \\ logF0_{11} \\ \Delta logF0_1 \\ \vdots \\ \Delta logF0_{10} \end{pmatrix}$$

$$\Delta logF0_j = logF0_j - logF0_{j+1} \qquad where\ j\ \in \{1,..,11\}$$

All vectors of the extracted tone nuclei are clustered into few groups 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. Each mean vector of all 11-point-F0 data in each group was later used as a tone template representative (T_k) in the F0 contour generating process later in section 4.4. The mean vector of each group is calculated by below equation. Fig 20 shows, as an example, all of the representatives of the tone nucleus shape separated by tone, when K is fixed to be 4.

$$T_k = \frac{1}{n_k} \sum_{i}^{n_k} \begin{pmatrix} logF0_{i1} \\ \vdots \\ logF0_{i11} \end{pmatrix}$$

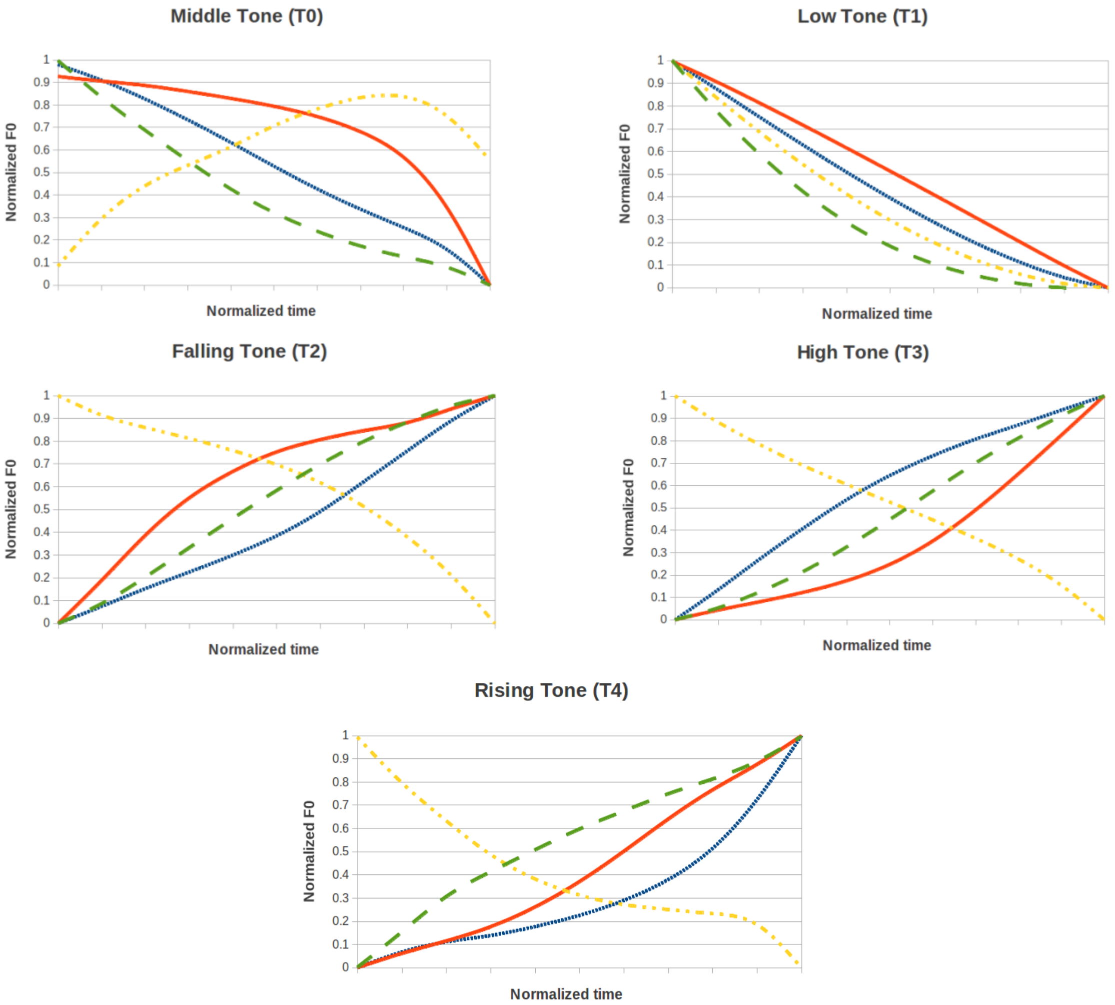


FIG 20: EXAMPLES OF TONE NUCLEUS SHAPE REPRESENTATIVES OF ALL FIVE TONES

4.3 TONE NUCLEUS PARAMETERS PREDICTION

After defining the tone nucleus parameters and extracting them, we need to predict these parameters in order to generate the F0 contour for unseen utterance (the target utterance). From the section 4.2, the parameters to be predicted are following ones:

Time related parameters : TNonset , TNdur

Frequency related parameters : F0min, F0range, TN template identity

All five parameters in each tone were predicted by classification and regression trees (CART) [35] trained separately and independently as shown in Fig 21. Totally there were 25 prediction trees (5 parameters × 5 tones).

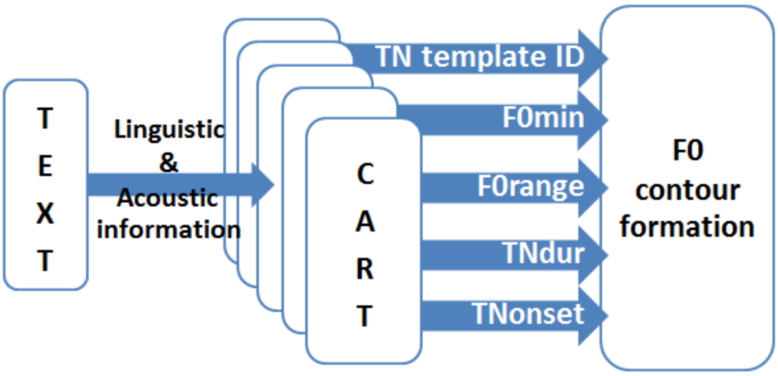


FIG 21: DIAGRAM OF THE PARAMETER PREDICTION

These prediction trees were built by question sets corresponding to linguistic and acoustic data of current, previous and following syllables as described in Table 10.

TABLE 10: INPUTS TO THE TONE NUCLEUS PARAMETER PREDICTORS

Inputs to the predictor	Category
Linguistic data	
Initial consonant of current syllable	38
Initial consonant of following syllable	39
Vowel of current syllable	24
Vowel of preceding syllable	25
Final consonant of current syllable	31
Final consonant of preceding syllable	32
Tone of current syllable	5
Tone of following syllable	6
Tone of preceding syllable	6
Part of Speech of current syllable	12
Position type of current syllable	6
Position of current syllable in word	Natural number
Number of syllables in current word	Natural number
Number of words in current sentences	Natural number
Acoustic data	
Duration of current syllable	Continuous

For linguistic data, these data can be normally obtained from the available labeled information in corpora. The corpus we used in this study [29] has both Thai original phonemes and additional foreign phonemes from loanwords, for instance, in original Thai, there is no */fr/* sound but some English loan words such as "free" is commonly used in daily life recently. These phonemes are mostly added to the sets of initial and final consonants, but hardly added to the vowel set.

As a result, the initial consonants used in the input list, here, are included original Thai and loaned consonants. In addition to Table 2, there are 6 more sounds from loanwords depicted in Table 11 (adopted from [29]). The vowel are 18 monophthongs (9 pairs of short-long vowels) and 6 diphthongs (3 pairs of short-long vowels) shown in Table 3 and Table 4 in section 3.1.2, respectively. The final consonants give 9 original Thai sounds as shown in Table 5 and additional 22 foreign sounds which come from other languages such as English words as shown in Table 12. Totally, there are 31 sounds for final consonants.

TABLE 11: FOREIGN PHONEMES OF INITIAL CONSONANTS FOUND IN THAI LOANWORDS

Place of articulation	Manner of articulation	Foreign Phonemes
Bilabial	Plosive + Trill	br
	Plosive + Lateral	bl
Alveolar	Plosive + Trill	dr , thr
Labiodental	Fricative + Trill	fr
	Fricative + Lateral	Fl

TABLE 12: FOREIGN PHONEMES OF FINAL CONSONANTS FOUND IN THAI LOANWORDS [29]

Manner of articulation		Foreign Phonemes
Stop +	Fricative	ks, ts, ps
Fricative +	Stop	st, sk
	Fricative	fs, th
Fricative	-	s, f
Nasal +	Fricative	ms, ns ,ngs
	Affricative	nch
Lateral +	Fricative	ls, lf
	Affricative	lch
Glide +	Fricative	js, ws, jf, wf
Glide	-	l
Affricate	-	ch

The part of speech (POS) is also provided in the corpus with regards to [36]. There are 46 totally categories. However, to reduce the sparseness in the limited number of selected training data, we applied only main categories of this feature into 12 categories as listed in Table 13.

TABLE 13: DESCRIPTION OF THE PART OF SPEECH TYPE

Type	Description
N	Noun
P	Pronoun
V	Verb
X	Pre-verb auxiliary and post-verb auxiliary
D	Definite and indefinite determiner
A	Adverb
C	Classifier
J	Conjunction
R	Preposition
I	Interjection
F	Nominal and Adverbial Prefix
E	Ending of the sentence

To cope with the declination effect, the position of the considering syllable was taken into account. The position type of the syllable in the sentence can be categorized into 6 categories according to [37]. This position type describes the position of the considering syllable corresponding to a higher level-unit, e.g., lexical word, prosodic word, phrase, breath group, etc. This position type has been originally designed to consider the locus in level of prosodic words, and phrases, but due to the limited prosodic labeling data of the existing corpus, the lexical word and phrase were applied instead of the prosodic ones.

TABLE 14: DESCRIPTION OF THE SYLLABLE POSITION CONTEXT TYPE

Type	Description
1	First syllable of a breath group
2	First syllable of a word but not first one of a breath group
3	Syllable within a word
4	Last syllable in a word but not the last one of a breath group
5	Last syllable of a breath group
6	Monosyllabic word

In building the decision trees process, each training tree was considered independently. In each tone, the training data set was first separated into 2 subsets exclusively by randomly selecting 30% of the data to form a validation set. The remaining data were used to train the trees. The wagon tool [38] was deployed with stepwise option enabled to build the trees. By enabling this option, only features that provide good accuracy to the validation data are considered forming the tree instead of all the input features. With greedy algorithm, firstly, the stepwise option builds a tree by searching for the best individual attribute which gives the best accuracy on the test set. In each step, the next best attribute which give the best accuracy among the remaining attributes is added incrementally until the accuracy is not improved or some stop criteria is met. We attempted to build 2 trees on the same parameter with this option enabled and without this option enabled; the result showed that the prediction tree built with this option enabled gave better accuracy on the validation set.

4.4 F0 CONTOUR GENERATION

Once all the parameters were predicted, the tone nucleus contour was constructed. The contour in the predicted TN template identity was linearly adjusted to fit the predicted F0min, calculated F0max and time span regarding to the predicted nucleus duration. The F0max can be calculated from the predicted F0min and F0range which depicted in section 4.2. Those constructed tone nuclei were later located at the specific time points calculated from the predicted nucleus onsets. And then, they were concatenated by piecewise cubic hermite interpolation[6] to form the F0 contour of the utterance.

4.5 EXPERIMENTAL DATA PREPARATION

766 training and 30 target utterances were randomly selected from Thai tagged speech corpus for speech synthesis [29] which was uttered by a Thai female professional news reporter in narrative style. The selected utterances were carefully checked whether there was no mismatch between speeches and provided linguistic labeling data. Table 15 shows number of syllables of each tone in both data sets statistically. The length of the 30 target utterances were varied from 8-58 syllables.

TABLE 15: NUMBER OF SYLLABLES IN THE TRAINING AND TARGET DATA SETS

Data set	T0	T1	T2	T3	T4
Training	7,753	5,112	3,916	3,699	2,002
Target	270	199	146	127	58

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 F0 contours. The target F0 contours were substituted with the generated F0 contours and the speech utterances were re-synthesized by Time Domain Pitch Synchronous OverLab-Add technique (TD-PSOLA) technique provided in Praat. TD-PSOLA is a non-parametric based pitch modification technique without performing any explicit source/filter separation.

To evaluate the performance of the model, we prepared 2 synthesized speech data sets on the same target data set. One set contains 30 speech utterances synthesized with the F0 contours generated from the tone nucleus model. This set was called "**TN approach**" for reference. The other was called "**WH approach**". We used the same method as in TN approach in prediction and F0 contour generation. The only difference is that no tone nucleus was detected. We used the F0 contours in the syllables as they are to train the prediction trees. The syllable F0 contours were clustered. The related parameters were extracted. As it is named, this data set consists of the speech utterances synthesized with the F0 contours generated by the prediction trees which were built from the F0 contours in the whole syllable. We compared the results between these two data sets through the objective and subjective tests which are explained later in section 4.6.

[6] For more detail, please visit http://www.mathworks.com/help/matlab/ref/pchip.html

4.6 EXPERIMENTS AND RESULTS

We need to evaluate the synthesized speeches objectively and subjectively whether they meet our research goals, which are the naturalness and the tone intelligibility. The objective evaluation is to measure goodness of fit between the synthetic speech and the target ones. RMSE (Root mean square error) and correlation coefficient are effective indicators and popular among researchers. In our objective test, we also used these metrics to evaluate the synthetic speeches. The detail is presented in section 4.6.1.

However, it is not easy to define suitable and meaningful objective distance measures which correspond to the perception. In addition, difficulty is brought when comparing objective values using different metrical distances, hence, many researches have evaluated the performances of their systems by other standard subjective tests, e.g., MOS (Mean Opinion Score).

In our subjective evaluation tests, two perceptual tests were performed separately on a web-based system. The first one is to check the tone intelligibility of the synthetic speeches by having the listeners check if the tone of given utterance sound matches the given textual script or not. The other is to have the listeners indicate how natural the synthetic speeches are perceived. The detail of the first one is discussed in section 4.6.2 whereas the latter one will be discussed later in section 4.6.3.

4.6.1 OBJECTIVE EVALUATIONS

RMSE (Root mean square error) and Pearson's product-moment correlation, in short correlation coefficient, between the generated data and the target one are employed as indicators. In this study, the average RMSE of all utterances was considered objectively.

RMSE shows how much the generated F0 contour is deviant from the target one. The RMSEs were calculated from the F0 contours of the synthesized speeches comparing to the one of the target speeches utterance by utterance.

$$RMSE_u = \sqrt{\frac{\sum_i^N (P_i - T_i)^2}{N}}$$

where P_i is the predicted F0 contour at i^{th} time point in the u^{th} utterance.

T_i is the target F0 contour at i^{th} time point in the u^{th} utterance.

N is the total number of time indices in the u^{th} utterance.

The correlation coefficient tells us how strong the linear association between a pair of variables is. The closer to 1 or -1, the stronger the relationship is. We can calculate it from below equation.

$$Corr_u = \frac{\sum_i^N (P_i - \mu_P)(T_i - \mu_T)}{\sqrt{\sum_i^N (P_i - \mu_p)^2 \sum_i^N (T_i - \mu_T)^2}}$$

where P_i is the predicted F0 contour at i^{th} time point in the u^{th} utterance

μ_P is the average F0 of predicted contour of the u^{th} utterance.

T_i is the target F0 contour at i^{th} time point in the u^{th} utterance

μ_T is the average F0 of target contour of the u^{th} utterance.

N is the total number of time indices in the u^{th} utterance

The chart in Fig 22 shows RMSE of each utterance in both approaches. 16 utterance contours generated by TN approach have less distortion than those generated by WH approach. Table 16 shows the averages and standard deviations of correlation coefficient and RMSE from the testing data set in both WH and TN approaches. Also, in average, TN approach generated less error than WH approach.

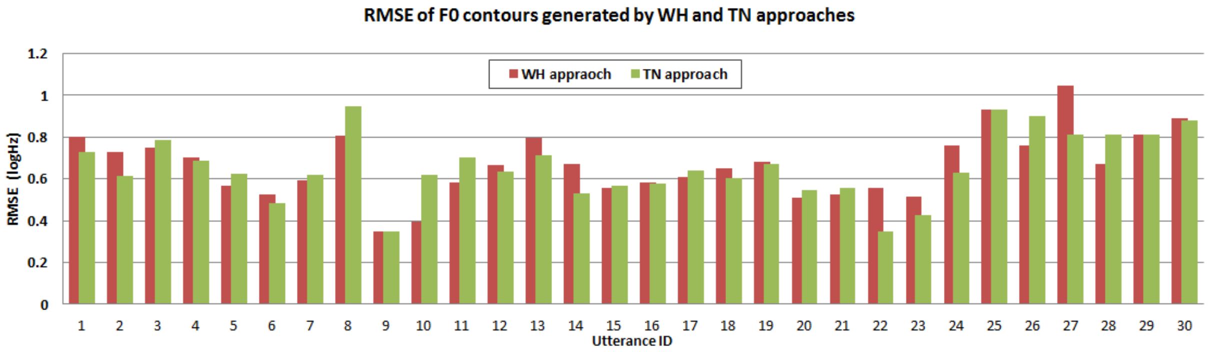


FIG 22: RMSE OF EACH F0 CONTOUR UTTERANCE GENERATED BY WH AND TN APPROACHES

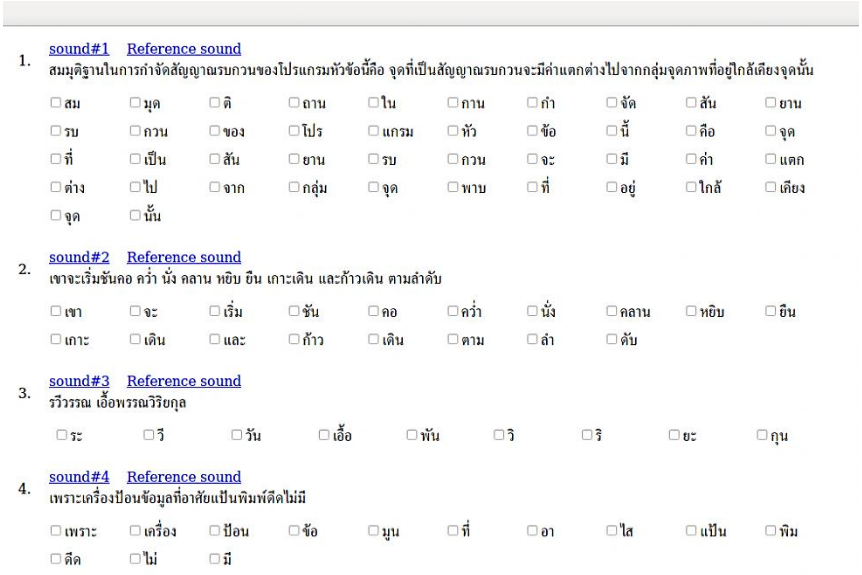
TABLE 16: RMSE AND CORRELATION RESULTS OF WH AND TN APPROACHES

Dataset	Evaluation	
	RMSE (log Hz)	Correlation
WH approach	0.6658±0.1547	0.9597±0.0276
TN approach	0.6583±0.1554	0.9615±0.0225

4.6.2 *TONE INTELLIGIBILITY EVALUATION*

In this section, we aimed to evaluate the synthetic speech utterance with F0 contour generated by the tone nucleus model in the aspect of tone intelligibility. This test was performed on a syllable basis rather than in utterance basis by assessing number of syllables which deliver incorrect tones.

A total of 30 native Thais aged varying from 22 to 32 years old, who can speak central Thai fluently, took part in the experiment. None of them are reported to have speech or hearing problems. They were requested to complete a given task on a web-based system[7]. The task was to identify which syllable in the continuous synthesized sounds conveys incorrect tone. Fig 23 shows the interface of the evaluation form. They could play each sound as many times as they want by clicking the corresponding link. They could also take a rest whenever they wanted to. In each sound, Thai textual transcriptions were given along with the target speech sounds to help them judge how the given texts should be pronounced. The listeners have to mark the checkbox if they find the tone of that syllable does not match the given text. The total 60 synthesized sounds from TN approach set and WH approach set (30 utterances from each set) were randomly presented to the listeners.



1. sound#1 Reference sound
สมมุติฐานในการกำจัดสัญญาณรบกวนของโปรแกรมหัวข้อนี้คือ จุดที่เป็นสัญญาณรบกวนจะมีค่าแตกต่างไปจากกลุ่มจุดภาพที่อยู่ใกล้เคียงจุดนั้น

□ สม □ มุด □ ติ □ ถาน □ ใน □ กาน □ กำ □ จัด □ สัน □ ยาน
□ รบ □ กวน □ ของ □ โปร □ แกรม □ หัว □ ข้อ □ นี้ □ คือ □ จุด
□ ที่ □ เป็น □ สัน □ ยาน □ รบ □ กวน □ จะ □ มี □ ค่า □ แตก
□ ต่าง □ ไป □ จาก □ กลุ่ม □ จุด □ พาบ □ ที่ □ อยู่ □ ใกล้ □ เคียง
□ จุด □ นั้น

2. sound#2 Reference sound
เขาจะเริ่มชันคอ คว่ำ นั่ง คลาน หยิบ ยืน เกาะเดิน และก้าวเดิน ตามลำดับ

□ เขา □ จะ □ เริ่ม □ ชัน □ คอ □ คว่ำ □ นั่ง □ คลาน □ หยิบ □ ยืน
□ เกาะ □ เดิน □ และ □ ก้าว □ เดิน □ ตาม □ ลำ □ ดับ

3. sound#3 Reference sound
รวีวรรณ เอื้อพรรณวิริยกุล

□ ระ □ วี □ วัน □ เอื้อ □ พัน □ วิ □ ริ □ ยะ □ กุน

4. sound#4 Reference sound
เพราะเครื่องป้อนข้อมูลที่อาศัยแป้นพิมพ์ดีดไม่มี

□ เพราะ □ เครื่อง □ ป้อน □ ข้อ □ มูน □ ที่ □ อา □ ไส □ แป้น □ พิม
□ ดีด □ ไม่ □ มี

Fig 23: Screen for the tone error evaluation task

The tone error is, here, presented in percentage of the total number of the evaluated syllables. It can be calculated as below equation.

$$Tone\ error = \frac{N_{err}}{N_s \times N_p} \times 100$$

where N_{err} is number of syllable which evaluated to convey tone incorrectly.

N_s is number of syllable in the testing data set.

N_p is number of participant of the evaluation test. In this case, N_p equals 30.

Table 17 shows the average tone error percentage of WH and TN approach.

[7] The evaluation system can be visited at
http://www.gavo.t.u-tokyo.ac.jp/~oraphan/listen_120712/120712te_eval.html

TABLE 17: TONE ERROR PERCENTAGES OF WH AND TN APPROACHES

Data set	Tone Error (%)
WH approach	4.75
TN approach	4.87

4.6.3 NATURALNESS EVALUATION

In this section, the quality of the speech synthesized based on the tone nucleus model are aimed to be evaluated subjectively in the aspect of naturalness. The participants are the same listeners in the experiment in section 4.6.2. They were requested to complete another task on the same web-based system[8] as in the previous section but different interface as illustrated in Fig 24. We used the MOS (Mean Opinion Score) with 5-level-scale. The values 1 to 5 in the scale rank the naturalness of the speech utterance from bad to excellent. (Bad = 1, Poor = 2, Fair = 3, Good = 4, and Excellent = 5) In this task, we had the listener evaluate not only the synthetic sounds but also the target sounds. Totally we had 3 sets of sounds to be evaluated: WH, TN and target data set. All 90 sounds (30 sounds per set) were randomly presented to the listeners without notifying them which sound is synthetic or natural.

MOS is an ordinal score to measure the opinion of the users in category and it cannot be treated numerically. Moreover, it is hardly interpreted the result meaningfully when the average of the score is not a natural number. To avoid the meaningless result, it is recommended to use median or the mid-points to be a descriptive statistic for this kind of data [39]. Thus, the median of each utterance were calculated and the contingency table was created as show in Table 18. Table 19 shows the descriptive statistical data e.g., means, medians, modes and standard deviations of naturalness of all 3 datasets. If each utterance has equal weight and we considered all the raw data, the number of the elements in each category can be obtained. Fig 25 shows the comparison of all MOS evaluated from the participants to the synthetic sounds in WH and TN approaches.

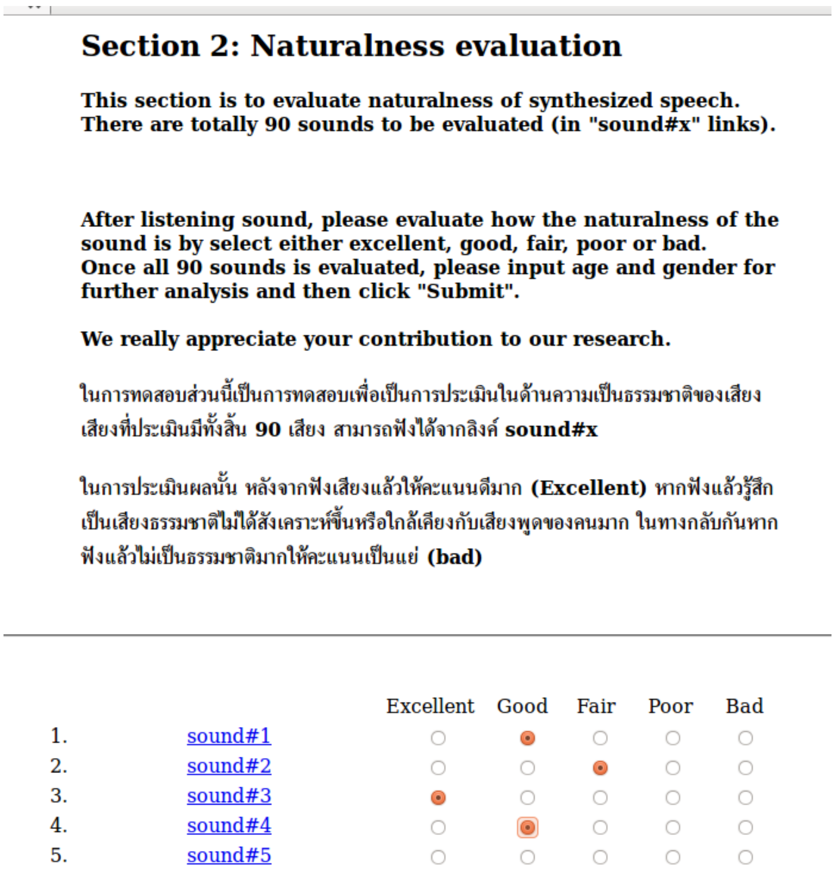
TABLE 18: CONTINGENCY TABLE OF MEDIAN MOS OF ALL UTTERANCES IN EACH DATASETS

Dataset	Excellent	Good	Fair	Poor	Bad
Target speech	30	0	0	0	0
WH approach	0	15	13	2	0
TN approach	1	16	11	2	0

TABLE 19: DESCRIPTIVE STATISTICS OF MOS

Dataset	Mean	Median	Mode	S.D	Number of samples
Target speech	5.0	5.0	5	0	30
WH approach	3.4	3.5	4	0.62	30
TN approach	3.5	4.0	4	0.67	30

[8] The evaluation system can be visited at http://www.gavo.t.u-tokyo.ac.jp/~oraphan/listen_120712/120712mos_eval1.html



Section 2: Naturalness evaluation

**This section is to evaluate naturalness of synthesized speech.
There are totally 90 sounds to be evaluated (in "sound#x" links).**

**After listening sound, please evaluate how the naturalness of the sound is by select either excellent, good, fair, poor or bad.
Once all 90 sounds is evaluated, please input age and gender for further analysis and then click "Submit".**

We really appreciate your contribution to our research.

ในการทดสอบส่วนนี้เป็นการทดสอบเพื่อเป็นการประเมินในด้านความเป็นธรรมชาติของเสียง เสียงที่ประเมินมีทั้งสิ้น **90** เสียง สามารถฟังได้จากลิงค์ **sound#x**

ในการประเมินผลนั้น หลังจากฟังเสียงแล้วให้คะแนนดีมาก **(Excellent)** หากฟังแล้วรู้สึกเป็นเสียงธรรมชาติไม่ได้สังเคราะห์ขึ้นหรือใกล้เคียงกับเสียงพูดของคนมาก ในทางกลับกันหากฟังแล้วไม่เป็นธรรมชาติมากให้คะแนนเป็นแย่ **(bad)**

		Excellent	Good	Fair	Poor	Bad
1.	sound#1	○	●	○	○	○
2.	sound#2	○	○	●	○	○
3.	sound#3	●	○	○	○	○
4.	sound#4	○	●	○	○	○
5.	sound#5	○	○	○	○	○

FIG 24: SCREEN FOR THE NATURALNESS EVALUATION TASK

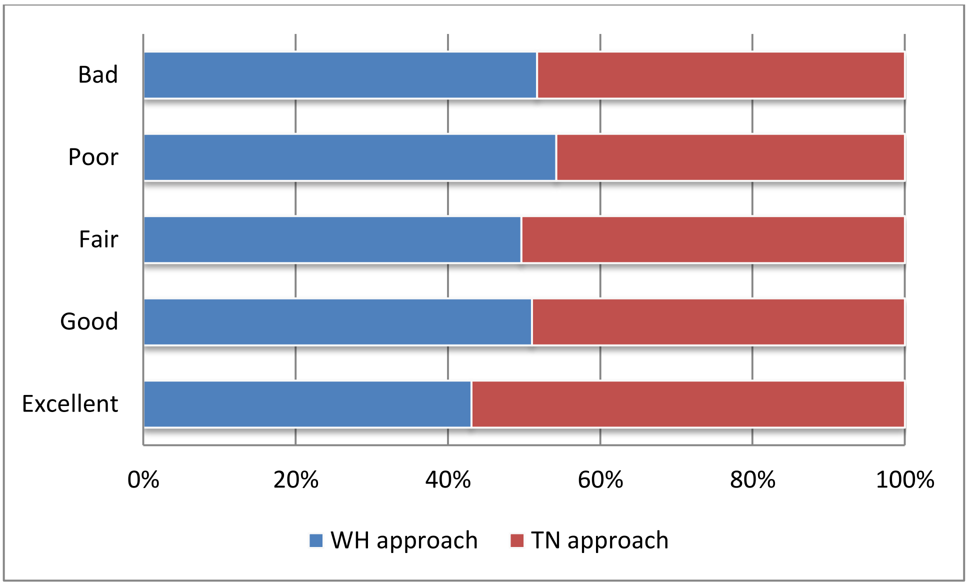


FIG 25: COMPARISON CHART OF MOS BETWEEN WH AND TN APPROACHES

4.7 DISCUSSION AND CONCLUSION

From our observation, the simple rules to detect the tone nucleus were determined. Five parameters were extracted from each detected tone nucleus. These parameters are TNonset, TNdur, F0min, F0range, and TN template identity. The classification and regression trees were employed to predict the tone nucleus parameters. Given the predicted parameters the F0 contour of the tone nucleus were generated. These tone nuclei were concatenated by cubic interpolation to form the F0 contour of the utterance. The F0 contours in the target utterances were substituted by the generated ones. And then, the target speech sounds were re-synthesized by TD-PSOLA.

By comparison to the whole syllable approach, we performed the evaluation through the objective and subjective tests. All evaluation results can be summarized as in Table 20. The result addresses that the tone nucleus model generates F0 contours that give less distortion about 1.12% than the model that considers the F0 contour in the whole syllable.

TABLE 20: OBJECTIVE AND SUBJECTIVE RESULTS OF THE EXPERIMENTS

Data set	RMSE (log Hz)	MOS (median)	Tone error (%)
Target speech	-	5.0	-
WH approach	0.6658	3.5	4.75
TN approach	0.6583	4.0	4.87

The tone nucleus model gives more naturalness than the other model that considers the F0 contour in the whole syllable. From Table 18, number of utterances which evaluated to be very natural (excellent) and natural (good) in the TN approach is more than that in the WH approach. Also from Table 19, although the mean of the MOS score that is not a number corresponding directly to the defined category is hard to interpret, but both mean and median of the WH approach is less than the TN approach. The number of evaluation for each category in MOS on WH and TN approaches in Fig 25 reveal apparently that the tone nucleus model can generate F0 contour more naturalness. The number of elements in Excellent on TN approach is more than that on the WH approach. On contrary, the number of elements in Bad on TN approach is less than that on the WH approach.

However, slightly less tone intelligibility is found in the generated F0 contour with the model. In addition, in the aspect of the naturalness, there is still a wide gap between the synthetic speech and the target natural speech.

Moreover, it is worth noting that the participants which are native speakers in a tonal language are very sensitive to the tone-error-syllable especially the syllables which locate at the end of the phrase or the syllables preceding short pauses. From our observation, we performed the naturalness evaluation on the utterance basis, but the participants tended to degrade the whole utterance even when the distortion appears on only few syllables. Also, the utterances containing the tone-error-syllable got evaluated to be unnatural.

5 Modification of the tone nucleus model parameter prediction and the F0 contour generation

From section 4.6.2, the syllables which were claimed to contain tone errors had high F0 distortion from the target speech. We investigated the predicted parameters and found that they were estimated inaccurately. Since many researchers have found that the F0 value relates to the syllable duration, the parameters prediction process was re-visited and changed in order to improve the prediction accuracy. Moreover, the feature list for training the decision trees is modified accordingly.

In this chapter, the problems are addressed and the idea how we tune the parameter prediction process is also presented in section 5.1. The modified prediction process reveals the better results through the objective and subjective tests in section 5.3. Lastly, discussion and conclusion are elaborated in section 5.4.

5.1 PROBLEMS AND PROPOSED METHODS

Due to the results from the section 4.6, the syllables containing tone error degraded the naturalness of the utterance. The syllables, which were judged to have tone errors by more than 50% of the native listeners, are listed in Table 21 with their statistical percentages, which indicate how many evaluators judged each particular syllable to have the tone error. Most of these syllables are also in the utterances which were evaluated with the under average MOS scores.

TABLE 21: THE 5 MOST SYLLABLE ERRORS EVALUATED IN THE PERCEPTUAL TEST

ID	Syllable	Tone error (%)
1	d-uua-j^2	90.0
2	d-uua-j^2	86.7
3	n-ii-z^3	80.0
4	m-ee-t^3	73.3
5	s-a-n^2	66.7

To analyze which parameter causes the tone error in these syllables, one of 5 parameters generating the F0 contour was manually adjusted to the target value individually and the F0 contours of the utterances were re-generated and then a perceptual test was conducted. The RMSEs of the utterances containing these syllables are shown in Table 22. Most of the modifications reduced the distortions because of the decreases of RMSEs. However, from the perceptual test results, the improvement is clearly noticeable.

Each shading cell with bold fold letters in Table 22 shows that the tone erroneous disappears when the corresponding parameter is adjusted. For syllable with ID 1, 2 and 3, the tone errors apparently came from the TN template identity parameter prediction errors because only this parameter can amend the errors. Likewise, the F0min parameter contributes the erroneous to the syllable ID 4. Differently, for the syllable ID5 the listeners could not judge which single parameter causes the tone erroneous.

TABLE 22: RMSES OF THE F0 CONTOURS OF THE UTTERANCES CONTAINING THE SYLLABLE ID WHOSE SPECIFIC PARAMETER WAS ADJUSTED TO THE TARGET VALUE.

Modified parameter	**RMSE (log Hz)**				
	Syl ID 1	**Syl ID 2**	**Syl ID 3**	**Syl ID 4**	**Syl ID 5**
None (Original)	0.7036	0.6054	0.4242	0.6195	0.5754
TNonset	0.7036	0.6055	0.3548	0.6463	0.5476
TNdur	0.7176	0.5854	0.3541	0.6192	0.6013
F0min	0.7182	0.5853	0.3585	**0.6181**	0.5738
F0band	0.7181	0.6055	0.4223	0.6456	0.6018
TN template identity	**0.7188**	**0.5848**	**0.3848**	0.6195	0.5754

From above analysis, an assumption was raise. We assumed that the effect of each parameter to the generated tone contour is not symmetry. The TN template identity parameter influences the tone erroneous the most whereas the temporal parameters e.g., TNonset and TNdur do not contribute erroneous to the generated F0 contour much. Besides that, very inaccurate TN template identity parameter prediction (about 45.49% in average) causes tone errors. Therefore, we wanted to increase the TN template identity parameter prediction accuracy in order to generate the tone contours correctly.

On our assumption that tone characteristics of the neighboring tones (both preceding and following ones) relate to the variation of the current tone shape especially in the short syllable. This triggered an idea to enhance the TN template prediction accuracy by adding this kind of context when predicting the parameter. In practice, we can take an advantage from the adjacently preceding syllable. Hence, in this case, we focused on how the preceding syllable's appearance contributes to the current syllable. This kind of information can be utilized to predict the TN template identity of the considering syllable. In this study, the preceding syllable's shape can be obtained by the TN template identity parameter; thus, we added the adjacently preceding syllable's TN template identity parameter to the question set when constructing the prediction tree of this parameter. Comparatively, according to the impurity measurement of the prediction trees shown in Table 23, entropies of the trees built with the modified question set are lower than those of the original trees. Lower entropy means the more similar the data are in each leaf of the prediction tree. This can be inferred that the preceding syllable's tone nucleus template identity leads to the better prediction result.

TABLE 23: ENTROPY AND PERPLEXITY OF THE TONE TEMPLATE PREDICTION TREE WITH AND WITHOUT THE TONE TEMPLATE OF THE PRECEDING SYLLABLE IN THE QUESTION SET

Tone	**Entropy (Perplexity)**	
	without preceding tone template	**with preceding tone template**
T0	0.436679 (1.35349)	0.435076 (1.35198)
T1	0.559127 (1.47338)	0.561465 (1.47577)
T2	0.433217 (1.35024)	0.436414 (1.35324)
T3	0.537038 (1.45099)	0.530517 (1.44445)
T4	0.595902 (1.51142)	0.576831 (1.49157)

According to previous researches, F0 value correlates to the syllable duration. We adjusted the procedure in prediction step and add the TNdur into the question set when constructing the decision trees of the F0 related parameters. As a result, the temporal parameters were estimated, and then added to predict the F0 related parameters subsequently. We confirmed this with regarding to the parameters selected by the greedy algorithm run in *stepwise* option in Wagon, a CART building tool. With the stepwise option provided in Wagon, the tree is built with only important attributes which are able to give high accuracy on the given validation set significantly instead of the full set of attributes. The stepwise option, first, builds a tree by searching for the best individual attribute which gives the best accuracy on the validation set. In each step, the next best attribute which gives the best accuracy among the remaining attributes is added incrementally until the accuracy is not improved or some stop criteria is met.

Considering this option, some of selected features by the tool were shown in Table 24, we can see in the last row of the table that the tone nucleus duration is selected to build F0range and F0mean prediction trees of most of the tones. This reveals that TNdur feature is a good candidate when building the F0 related trees.

TABLE 24: MATRIX OF THE SELECTED FEATURES BY STEPWISE OPTION IN WAGON WHEN BUILDING THE PREDICTION TREES OF F0MEAN, F0RANGE AND TN TEMPLATE IDENTITY

Features	F0mean					F0range					TN Template ID				
	T0	T1	T2	T3	T4	T0	T1	T2	T3	T4	T0	T1	T2	T3	T4
Initial consonant of current syllable	X	X		X		X	X			X	X	X	X		
Initial consonant of following syllable			X	X	X	X		X	X		X		X		X
Vowel of current syllable		X	X				X			X	X	X		X	
Vowel of preceding syllable	X					X		X							
Final consonant of current syllable		X	X				X	X			X	X	X	X	X
Final consonant of preceding syllable	X						X	X							
Tone of following syllable	X	X	X	X		X		X		X		X	X	X	
Tone of preceding syllable	X	X		X		X	X	X			X	X	X	X	
Part of Speech of current syllable	X				X		X								X
Position type of current syllable		X	X	X						X			X		X
Position of current syllable in word	X	X		X	X										
Position of current syllable in sentence	X	X		X					X	X			X		
TN duration of current syllables	X	X	X	X		X	X	X	X	X					

Moreover, with regard to the prediction trees, we found that the average RMSE of the predicted F0mean parameters (0.05692 log Hz) was more accurate than the average RMSE of the F0min parameters (0.06466 log Hz). This motivated us to improve how the F0 contour is interpolated from the predicted parameters. It is another possibility to make use of the parameters which were predicted more precisely (F0mean and F0range) to calculate the F0 contour by linear interpolation. Therefore, instead of directly using F0min parameter, the F0mean and F0range parameters were used to calculate the absolute value to construct the F0 contour from the normalized tone nucleus template. We performed the validation with 2 synthetic speech sets from the same parameters but different tone nucleus contour constitution method: one with the original method and the other with the modified method. The modified method decreases the distortion about 6.4% in avarage.

To be concrete, our proposals to reduce the F0 distortion and tone erroneous are as follow:

(a) Add preceding tone nucleus template identity into the training question set to help predict the current tone nucleus template identity.
(b) Add tone nucleus duration into the training question set to help predict the F0 related parameters.
(c) Revise using average F0 as a parameter to reconstitute the tone nucleus contour from the normalized tone nucleus template.

In summary, the prediction procedure has been changed as shown in Fig 26. The time related parameters, TNonset and TNdur, were predicted first. The predicted TNdur were, then, added into the question sets to predict the F0mean, F0range and TN template identity. Furthermore, when predicting the current TN template identity, the predicted TN template identity of the adjacently preceding syllable was also included in the question set. This modified method corresponding to our proposal will be henceforth referred as **"MTN approach".**

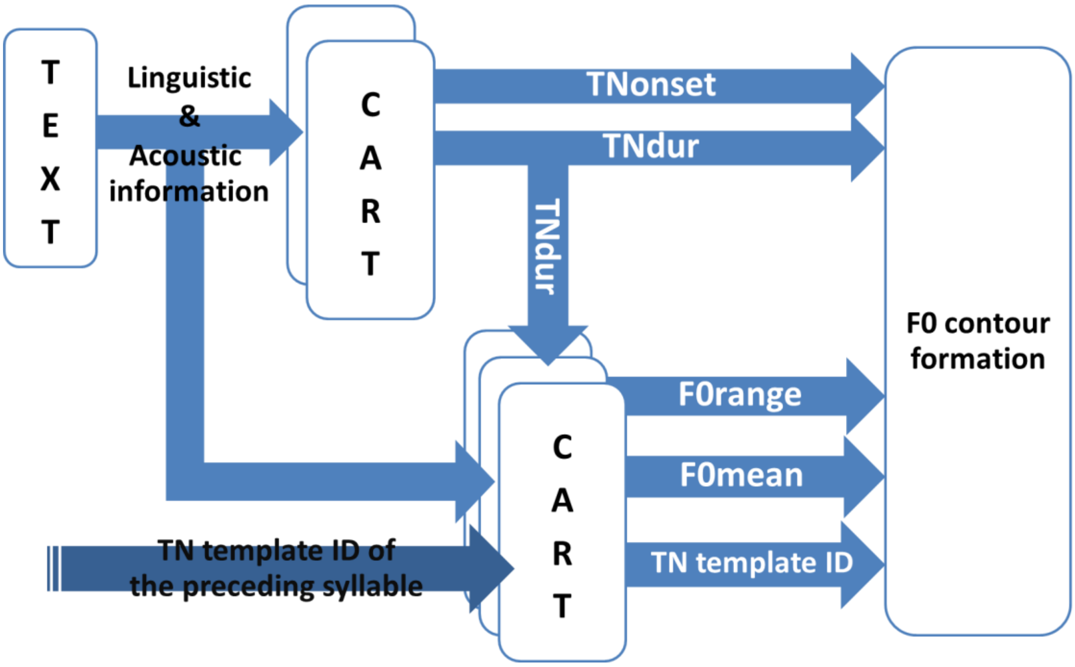


FIG 26: DIAGRAM OF THE MODIFIED PARAMETER PREDICTION (MTN APPROACH)

To evaluate the model, 10-fold cross validation on the training data set was conducted. The results in Table 25 reveal that the modified model (MTN approach) gives less distortion mostly. It is clear to see that all RMSEs decreased except in the case of T3's F0range. However, comparing to the other improvements, the higher error is very low. In average, the MTN approach reduced the F0mean parameter distortion from 0.089 to 0.074 (16.85%) and F0range parameter distortion from 0.080 to 0.071 (11.25%).

TABLE 25: AVERAGE RMSES OF 10-FOLD CROSS VALIDATION OF THE PREDICTED F0MEAN AND F0RANGE IN TN AND MTN APPROACHES

Tone	RMSE (log Hz)			
	F0mean		F0range	
	TN	MTN	TN	MTN
Mid (T0)	0.092	0.067	0.076	0.067
Low (T1)	0.097	0.082	0.086	0.067
Falling (T2)	0.079	0.069	0.083	0.077
High (T3)	0.099	0.086	0.077	0.079
Rising (T4)	0.080	0.067	0.079	0.065
Average	**0.089**	**0.074**	**0.080**	**0.071**

5.2 EXPERIMENTAL DATA PREPARATION

Both the training and target data sets are as same as in section 4.5. Also the tone nucleus detection algorithm was not changed, so we obtained the same contours from the detected tone nuclei. F0mean parameters was extracted from the detected tone nuclei instead of F0min. Indifferently, in the training process, CART trees were trained by Wagon with stepwise option enabled. The difference from the previous chapter is that the TNdur parameter and the TN template identity parameter of the previous syllable were added to the question set as addressed earlier in section 5.1. By the proposed methodology, the tone nucleus contours were reconstituted from the predicted parameters, and then, those generated tone nucleus contours were connected to each other with the same interpolation technique. All 30 synthesized speeches according to above methodology were prepared. For further reference, this prepared speech utterances dataset is named as "**MTN approach**".

5.3 EXPERIMENTS AND RESULTS

5.3.1 OBJECTIVE EVALUATIONS

In this experiment, we also used RMSE and correlation as the indicators. The results are shown in Table 26 and Fig 27. 21 utterance contours generated by MTN approach have less distortion than those generated by TN approach. Also, in average, MTN approach generated less error than WH approach. We can clearly see from Fig 28 that the generated contours of the tone nuclei in a test contour utterance by the MTN approach are much closer to the target contours than the ones from the TN approach. A paired-samples t-test showed that RMSEs of the MTN approach were significantly less than those of the TN approach, *t (29) = 2.67, p < 0.01*.

TABLE 26: RMSE AND CORRELATION RESULTS OF TN AND MTN APPROACHES

Dataset	Evaluation	
	RMSE (log Hz)	Correlation
TN approach	0.6583±0.1554	0.9615±0.0225
MTN approach	0.6124±0.1706	0.9655±0.0223

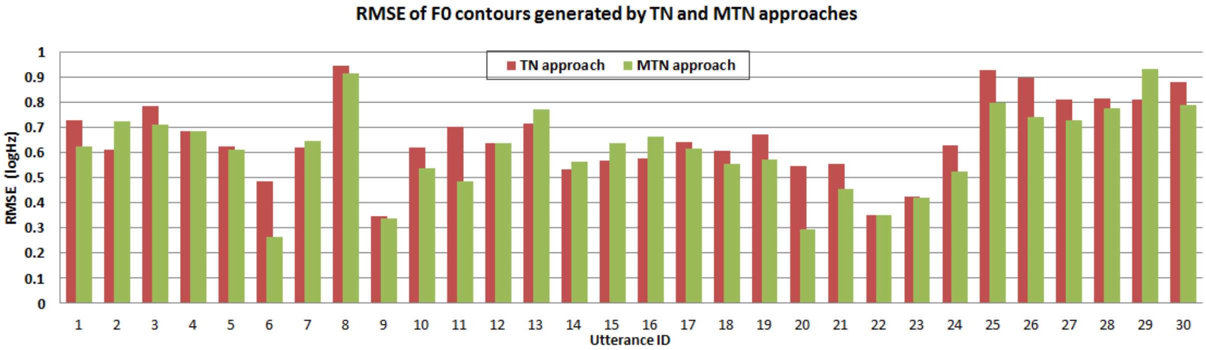


FIG 27: RMSE OF EACH F0 CONTOUR UTTERANCE GENERATED BY TN AND MTN APPROACHES

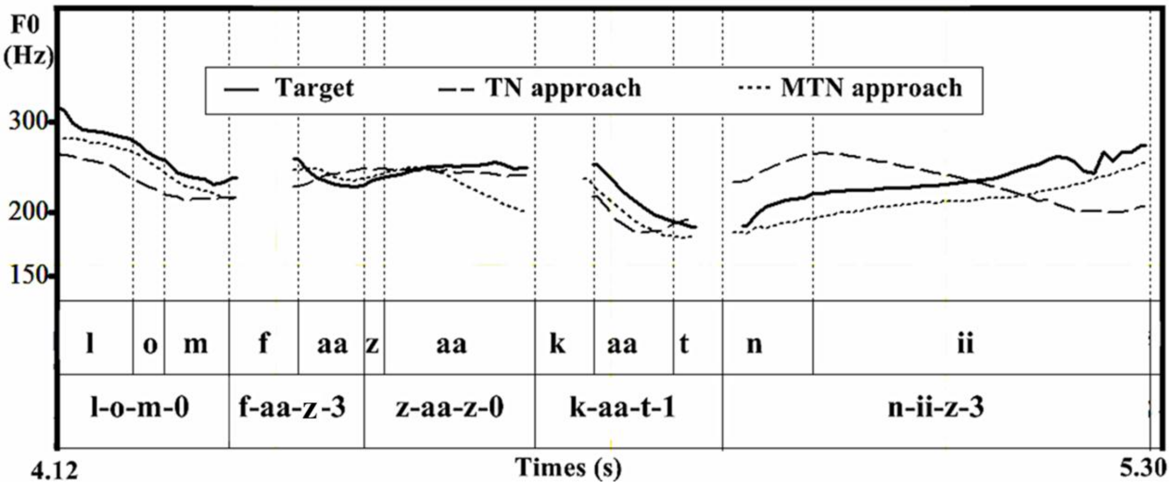


FIG 28: GENERATED F0 CONTOURS FROM TN (DASH LINE) AND MTN (DOT LINE) APPROACHES COMPARING TO THE TARGET CONTOURS (SOLID LINE) IN THE TOP PANEL. THE VERTICAL LINES REPRESENT THE BOUNDARIES ACCORDING TO THE PHONEMES AND SYLLABLES IN LOWER PANELS. TONES ARE MARKED AT THE END OF EACH SYLLABLE ANNOTATION AT THE BOTTOM PANEL.

5.3.2 *TONE INTELLIGIBILITY EVALUATION*

The objective of this task is as same as in section 4.6.2. A total of 19 native Thais aged varying from 22 to 32 years old, who can speak central Thai fluently, took part in the experiment. They were requested to complete a given task on a web-based system[9]. The task is the same as in section 4.6.2. The total 30 synthesized sounds of the test set generated by the MTN approach were randomly presented to the listeners. The result of average tone error percentages is shown comparatively to the TN approach in Table 27.

TABLE 27: TONE ERROR PERCENTAGES OF TN AND MTN APPROACHES

Data set	Tone Error (%)
TN approach	4.87
MTN approach	3.16

5.3.3 NATURALNESS EVALUATIONS

The objective of this section is as same as in section 4.6.3. There were 2 evaluation tasks: the naturalness evaluation by absolute MOS scale and the preference evaluation in the aspect of naturalness.

For the first task, the MOS was also used as the indicator. The same group of listeners in section 5.3.2 participated in this evaluation. They were requested to complete a given task on a web-based system[10]. The task is exactly as same as in section 4.6.3. The total 30 synthesized sounds of the test set generated by the MTN approach were randomly presented to the listeners. The MOS medians of each utterance were calculated and the contingency table shows the number of each category in Table 28. The descriptive statistical data of the evaluated scores are elaborated in Table 29. To see the contribution of the data, the number of the elements in each category were obtained and shown in Fig 29.

TABLE 28: CONTINGENCY TABLE OF MEDIAN MOS OF ALL UTTERANCES IN EACH DATASETS

Dataset	Excellent	Good	Fair	Poor	Bad
TN approach	1	16	11	2	0
MTN approach	11	10	6	3	0

TABLE 29: DESCRIPTIVE STATISTICS OF MOS

Dataset	Mean	Median	Mode	S.D	Number of samples
TN approach	3.5	4.0	4	0.67	30
MTN approach	4.0	4.0	5	0.98	30

[9] The evaluation system can be visited at http://www.gavo.t.u-tokyo.ac.jp/~oraphan/121126/121126te_eval.html

[10] The evaluation system can be visited at http://www.gavo.t.u-tokyo.ac.jp/~oraphan/121126/121126mos_eval.html

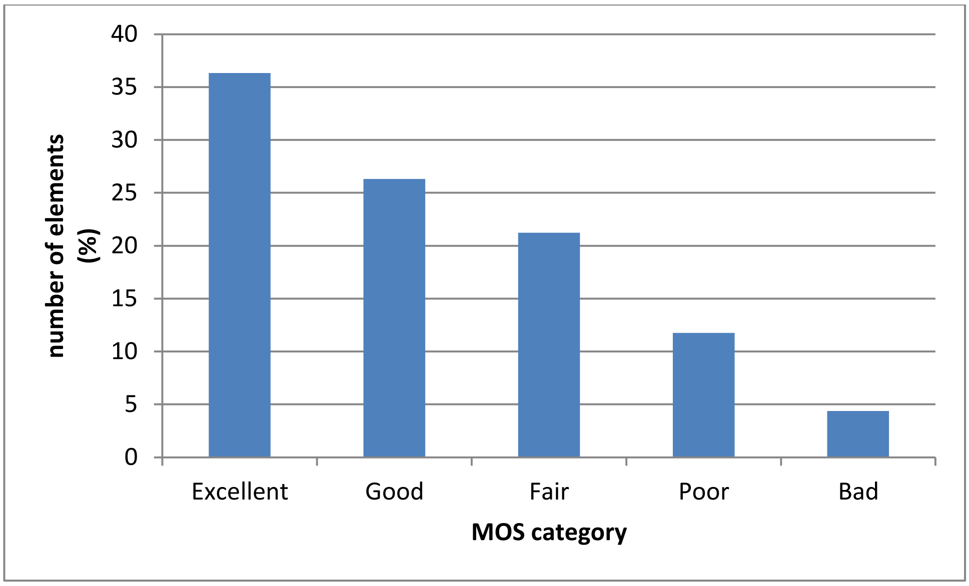


FIG 29: NUMBER OF ELEMENTS IN EACH MOS CATEGORY FOUND IN MTN APPROACH EVALUATION

Both approaches have equal median but the MTN approach has higher mean and higher mode than the TN approach and from Table 28 number of utterances that have higher evaluated score increased. To guarantee that the average MOS of MTN and TN approaches are comparative, we use a statistical test for this task. Typically, the statistical test were used to draw inferences about the mean of two populations if they are statistically significant different or not. In our experiments, they correspond to the rating score for the synthetic speech of 2 models: one came from the TN approach and the other came from the MTN approach. Normally, T-test is a popular parametric statistical test; nevertheless, our data do not meet the assumptions of the T-test. Our testing data do not have equally numerical interval scale. The MOS score is a psychological measure of which the distance between the categories may be not considered to be equal. Moreover, as shown in Fig 29, the data is not normal distribution. The skewedness test can confirm that the data is not normal distribution with skewedness of -0.59594. Therefore, T-test is the inappropriate statistical tool to be used in our case.

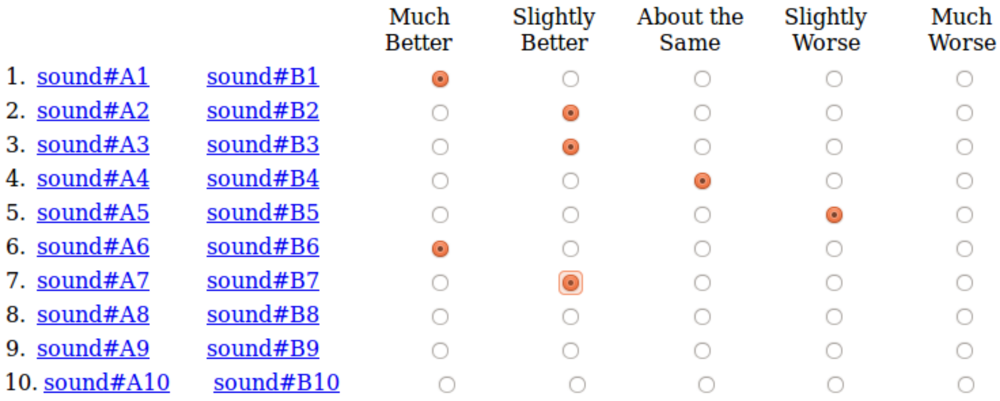
Alternatively, with regarding to a suggestion in speech synthetic evaluations in Blizzard Challenge 2007 [40], the non-parametric Wilcoxon signed rank test was considered being more proper. Wilcoxon signed-rank test failed to reject the null hypothesis (*Z = 1.89, p = .056, r = 0.35*). MTN approach did not elicit the improvement from the TN approach with statistical significance.

For the other task, to assess the impact of the modified methodology, the preference test comparing the mean opinion score directly was conducted in which a pairwise comparison was made between sounds synthesized with the F0 contours generated by TN approach and the one generated by MTN approach. We used the comparison category rating (CCR) with 5-point scale shown in Table 30 (adapted from [41]) as a relative measure unit. Positive and negative numbers are used to describe the directions of preference. This indicator is used to compare between a pair of synthetic sounds on how the second one sounds better than the first one. In our experiment, it is interpreted that the more score, the more natural the second sample sound is.

TABLE 30: COMPARISON CATEGORY RATING SCORE IN PREFERENCE TEST

Category	Score
Much better	2
Slightly better	1
About the same	0
Slight worse	-1
Much worse	-2

In this additional task, the same group of the participants in the previous task was requested to complete this given task on the same web-based system[11] with a different interface as illustrated in Fig 30. In each pair, the listeners were asked to listen to two synthesized sounds: *sound#A* and *sound#B*, respectively. And then, they were asked to judge how much naturalness of the second sound has relatively to that of the first one. They had to do these steps until finishing all of the 30 synthetic sound pairs.



			Much Better	Slightly Better	About the Same	Slightly Worse	Much Worse
1.	sound#A1	sound#B1	●	○	○	○	○
2.	sound#A2	sound#B2	○	●	○	○	○
3.	sound#A3	sound#B3	○	●	○	○	○
4.	sound#A4	sound#B4	○	○	●	○	○
5.	sound#A5	sound#B5	○	○	○	●	○
6.	sound#A6	sound#B6	●	○	○	○	○
7.	sound#A7	sound#B7	○	●	○	○	○
8.	sound#A8	sound#B8	○	○	○	○	○
9.	sound#A9	sound#B9	○	○	○	○	○
10.	sound#A10	sound#B10	○	○	○	○	○

FIG 30: SCREEN FOR THE NATURALNESS PREFERENCE EVALUATION TASK

[11] The evaluation system can be visited at http://www.gavo.t.u-tokyo.ac.jp/~oraphan/121126/121126ccr_eval.html

Also, the CMOS is ordinal measurement. Medians of each utterance were calculated, and then the descriptive statistical data of all utterances were obtained as shown in Table 31. In overall, the scores evaluated by the participants are illustrated in Fig 31.

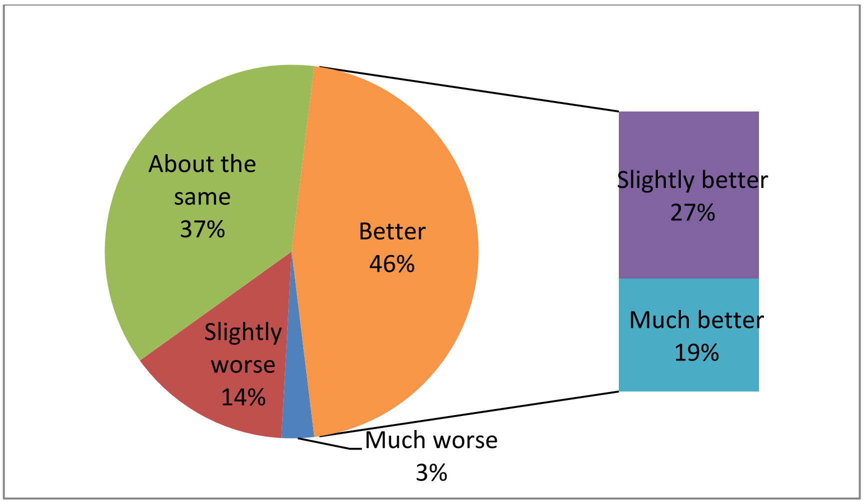


FIG 31: PERCENTAGES OF ALL CMOS EVALUATED BY PARTICIPANTS

We can see that the median of the median evaluated score of all utterances is 0, however, the number of evaluated scores for "slightly better" and "much better" categories were more than that in "about the same" category. In this case, we also use Wilcoxon signed-rank test to check that the median is not equal zero and the test indicated that the data did not statistically come from the distribution of zero median (*Z = -2.36, p < .05, r = 0.43*)

TABLE 31 : DESCRIPTIVE STATISTICS OF CCR

Statistical indicator	Value
Mean	0.4
Median	0.0
Mode	0
S.D.	0.84
Number of samples	30
Much better	3
Slightly better	10
About the same	13
Slightly worse	4
Much worse	0

5.4 DISCUSSION AND CONCLUSION

We proposed the modification in parameter prediction and F0 contour generation. F0mean parameter was suggested using than F0min for generate the tone nucleus contour from the existing normalized tone nucleus template. The contextual parameters were added to the question sets when constructing the decision trees. They are the tone nucleus duration of considering syllable and the tone nucleus template identity of the preceding syllable. Both subjective and objective tests were conducted. The objective test showed that the proposed methodology reduced the distortion of the generated F0 contour. Moreover, a paired-samples t-test indicated that RMSEs were significantly lower for the tone nucleus model with proposed method (*Mean = 0.6583, S.D. = 0.15537*) than the WH approach (*Mean = 0.6658, S.D. = 0.15470*), *t (29) = 2.43, p < 0.05*.

Also, the subjective tests corresponding to the tone intelligibility revealed that the modified methodology raises the tonal intelligibility by decreasing the number of tone errors. For the naturalness evaluation, although the result is not statistically significance on the difference between the proposed methodology and the conventional one, by the overall scores, it performs much better than the conventional one as shown in Fig 32. It is apparent that the numbers of elements in “Excellent” and “Good” categories have been increased drastically comparing to the WH approach. Likewise, it decreased the numbers of elements in “Bad” and “Poor”.

In addition, the preference test also indicates that the proposed approach can yield more preference score than the conventional one; however, there are still some samples which were evaluated to be degraded. Considering the difference between the naturalness of the synthetic speeches by the tone nucleus model with the proposed methodology and the other without the tone nucleus model, the statistical test (Wilcoxon signed rank test) shows that the tone nucleus model with the modified methodology statistically outperformed the one without applying the tone nucleus model on the naturalness of the synthetic speeches (*Z = -2.43, p < 0.05, r = 0.44*).

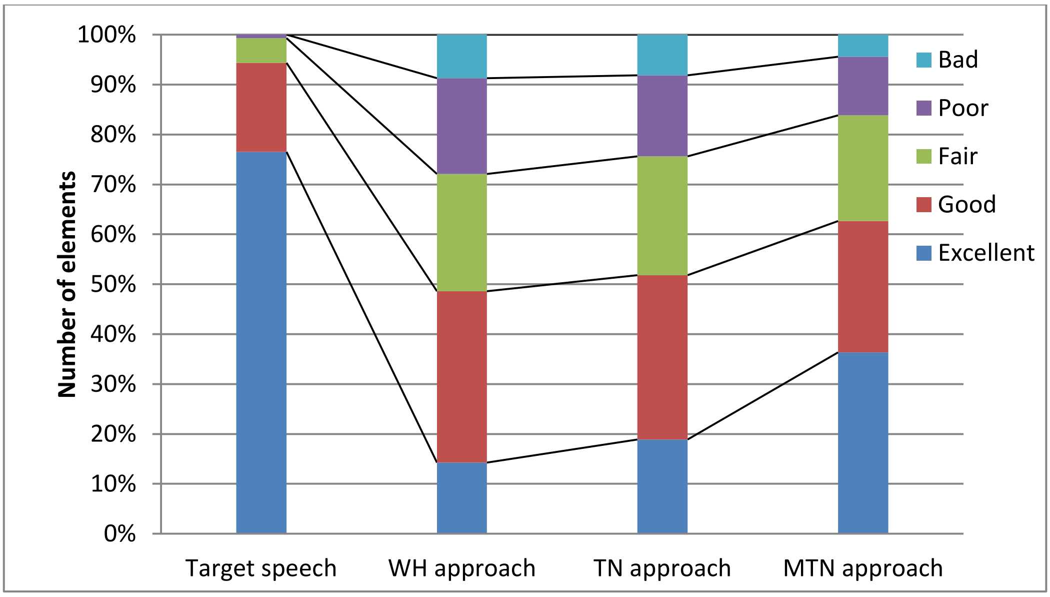


FIG 32 : DISTRIBUTIONS OF THE NATURALNESS CATEGORY IN EACH TESTING DATASET

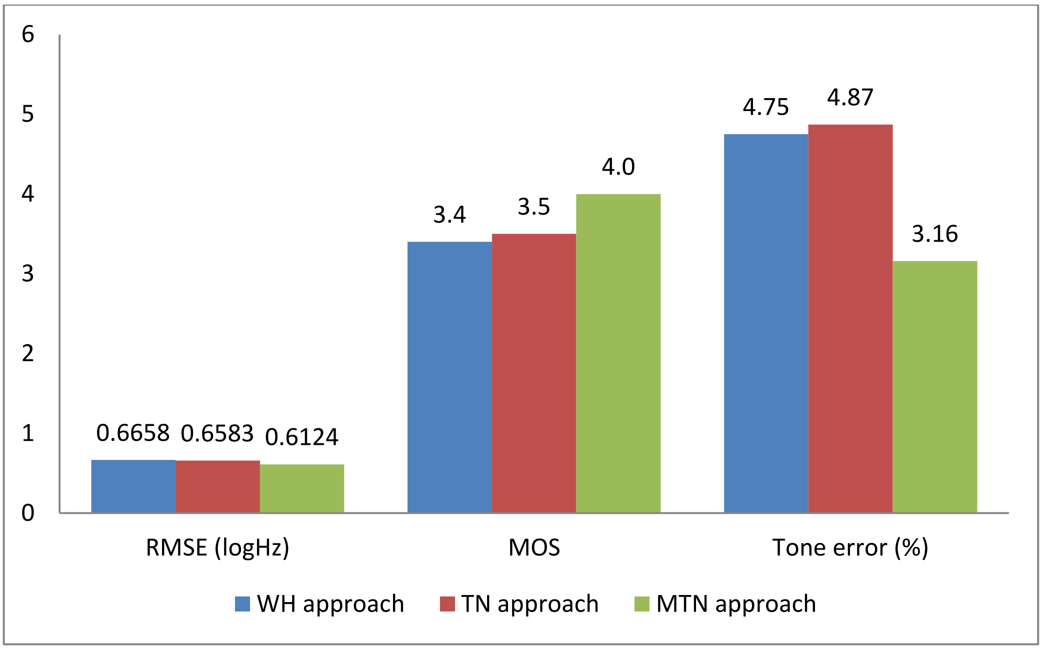


FIG 33 : OBJECTIVE AND SUBJECTIVE RESULTS OF THE F0 CONTOURS GENERATED BY WH, TN AND MTN APPROACHES

To sum up, the results are shown in Fig 33 with improved tone contour generation accuracy, less F0 distortion and more naturalness. This shows that the proposed method improves the generated F0 contour results to yield more naturalness of the synthesized speech and reduce tone errors.

6 Conclusions and future work

6.1 CONCLUSIONS

We have shown research works focusing on the F0 contour generation for Thai language to meet the essential but still challenging requirements of speech synthesis. To the best of our knowledge and surveys, this work is among the first that has applied the tone nucleus model as the F0 modeling in Thai language. As its advantages, the tone nucleus model was successfully adapted with very compact parameter sets on the relatively small size of training data. The tone nuclei were defined for all five Thai tones corresponding to their underlying targets (Chapter 3). From the analysis and observation, the rules detecting the tone nuclei automatically were derived (Chapter 4). All tone nucleus parameters were predicted by Classification and Regression trees with linguistic and acoustic information. The predicted tone nuclei were concatenated by cubic interpolation to form the F0 contours of the utterances. All synthetic speeches, here, were re-synthesized by TD-PSOLA with fixed spectrum from original speech signals. The objective and subjective results has proved that the tone nucleus model is applicable to the other tonal language than Mandarin, which is the first language the model has been originally applied on. Furthermore, the promising process has been proposed to cope with the high distortion of the generated F0 contours (Chapter 5). The objective test indicated the improved performance of the proposed methodology with reduced average RMSE whereas the subjective tests significantly elicited the improvement in tone intelligibility with reduced tone error percentages and increased naturalness with better MOS.

6.2 FUTURE WORK

The methodologies and results presented in this thesis are potentially useful to other applications in higher level e.g., word emphasis. Nevertheless, many issues can be improved and extended. The future work can be focused on the following issues.

- Issue about to meet the tonal intelligibility in speech synthesis, we cannot focus partially only on the F0 contour generation or the short-term spectrum generation. These two factors should be considered together. As interesting evidence, [42] argues that F0 contour is not sufficient enough to identify tone correctly from the perceptual test and the envelope of the spectrum also contributes tonal information to the generated speech.
- In this work, we studied and evaluated the model speaker-dependently. However, I believe that the tone nucleus model can apply to other speakers because the tone nuclei for all five Thai tones were defined mainly by the underlying tone targets which do not depend on the particular speaker.

Bibliography

[1] P. Seresangtakul and T. Takara, "Analysis of pitch contour of Thai tone using Fujisaki's model," in *IEEE*, 2002.

[2] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki and P. Charnvivit, "Perception of tone and vowel quantity in Thai," 2002.

[3] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan,* vol. 5, no. 4, pp. 233-241, 1984.

[4] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol and C. Wutiwiwatchai, "Optimization of t-tilt F0 modeling," 2009.

[5] 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," 2008.

[6] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the acoustical society of America,* vol. 107, p. 1697, 2000.

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

[8] X. Wang, K. Hirose, J. Zhang and N. Minematsu, "Tone recognition of continuous Mandarin speech based on tone nucleus model and neural network," *IEICE-Transactions on Information and Systems,* vol. 91, no. 6, pp. 1748-1755, 2008.

[9] 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.

[10] M. Wen, M. Wang, K. Hirose and N. Minematsu, "Prosody Conversion for Emotional Mandarin Speech Synthesis Using the Tone Nucleus Model," *IPSJ SIG Notes,* vol. 2011, no. 2, pp. 1-6, jul 2011.

[11] P. Seresangtakul and T. Takara, "A generative model of fundamental frequency contours for polysyllabic words of Thai tones," in *IEEE*, 2003.

[12] N. Thubthong, B. Kijsirikul and S. Luksaneeyanawin, "An empirical study for constructing Thai tone models," 2002.

[13] 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.

[14] P. Teeranon, "The change of Standard Thai high tone: An acoustic study and a perceptual experiment.," *SKASE Journal of Theoretical Linguistics [online],* 2007.

[15] A. Abramson, "Lexical Tone and Sentence Prosody in Thai41," 1979.

[16] K. Thepboriruk, "Bangkok Thai Tones Revisited," *JSEALS,* vol. 3, pp. 86-105.

[17] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," 2007.

[18] M. Tingsabadh and A. Abramson, "Thai," *Journal of the International Phonetic Association,* vol. 23, no. 01, pp. 24-28, 1993.

[19] B. Moren and E. Zsiga, "The Lexical and Post-Lexical Phonology of Thai Tones*," *Natural Language & Linguistic Theory,* vol. 24, no. 1, pp. 113-178, 2006.

[20] H.-W. H. Xuedong, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, 1 edition ed., Prentice Hall, 2001, p. 1008.

[21] J. Gandour, "Tonal coarticulation in Thai," *Journal of Phonetics,* vol. 22, no. 4, pp. 477-492, 1994.

[22] N. Satravaha, "Tone Classification of Syllable-Segmented Thai Speech Based on Multilayer Perceptron," 2002.

[23] S. Potisuk, M. Harper and J. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," *Speech and Audio Processing, IEEE Transactions on,* vol. 7, no. 1, pp. 95-102, 1999.

[24] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics,* vol. 27, no. 1, pp. 55-105, 1999.

[25] P. Pittayaporn, "Directionality of tone change," 2007.

[26] V. Boonpiam, A. Rugchatjaroen and C. Wutiwiwatchai, "Cross-language F0 modeling for under-resourced tonal languages: a case study on Thai-Mandarin," 2009.

[27] H. Fujisaki and W. Gu, "Phonological representation of tone systems of some tone languages based on the command-response model for F0 contour generation," 2006.

[28] S. Kallayanamit, "Intonation in Standard Thai: Contours, registers and boundary tones," 2004.

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

[30] Y. Xu and Q. Wang, "What can tone studies tell us about intonation?," 1997.

[31] E. Zsiga and R. Nitisaroj, "Tone features, tone perception, and peak alignment in Thai," *Language and Speech,* vol. 50, no. 3, pp. 343-383, 2007.

[32] J. Ni, S. Sakai, T. Shimizu and S. Nakamura, "Prosody modeling from tone to intonation in Chinese using a functional F0 model," in *IEEE*, 2008.

[33] D. B. Paul, *Praat: doing phonetics by computer [Computer program]. Version 5.2.40, retrieved 2011 from http://www.praat.org/.*

[34] D. Weenink, *Speech Signal Processing with Praat,* 2013.

[35] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, Monterey, CA: Wadsworth and Brooks, 1984.

[36] V. Sornlertlamvanich, T. Charoenporn and H. Isahara, "ORCHID: thai part-of-speech tagged corpus," *Orchid, TR-NECTEC-1997-001,* pp. 5-19, 1997.

[37] Y. Hu, M. Chu, C. Huang and Y. Zhang, "Exploring tonal variations via context-dependent tone models," 2007.

[38] P. Taylor, R. Caley and A. Black, *The Edinburgh Speech Tools Library. 1.0.1 edition http://www.cstr.ed.ac.uk/projects/speechtools.html,* 1998.

[39] G. T. Eduardo, "Statistical Analysis of Ordinal User Opinion Scores," 2012.

[40] R. A. J., M. Podsiad?o, M. Fraser, C. Mayo and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," 2007.

[41] N. S. Assessment, *Speech quality assessment methods - subjective assessment method for speech quality http://www.ntt.co.jp/qos/qoe/eng/technology/sound/03_3.html.*

[42] N. Kertkeidkachorn, S. Vorapatratorn, S. Tangruamsub, P. Punyabukkana and A. Suchato, "Contribution of Spectral Shapes to Tone Perception".

[43] Y. Xu, "Speech prosody: a methodological review," *Journal of Speech Sciences,* vol. 1, no. 1, pp. 85-115, 2012.

[44] N. Thubthong, "A Study of Various Linguistic Effects on Tone Recognition in Thai Continuous Speech," 2001.

[45] K. Hirose and H. Fujisaki, "Analysis and synthesis of voice fundamental frequency contours of spoken sentences," in *IEEE*, 1982.

[46] W. J. Gedney, "A comparative sketch of White, Black, and Red Tai," *Social Science Review,* vol. 1, pp. 1-47, 1964.

[47] P. J. Bee, "Restricted phonology in certain Thai linker-syllables," in *Studies in Tai linguistics in honor of William J. Gedney*, J. G. Harris and J. R. Chamberlain, Eds., Central Institute of

English Language, 1975, pp. 17-32.

[48] P. Taylor, Text-to-Speech Synthesis, Cambridge Univ Press, 2009.

Publications

- Krityakien, O., Hirose, K., and, Minematsu, N., “Tone nucleus model for Thai language speech synthesis,” in Proc. Acoustic Society of Japan (ASJ) meeting, Autumn 2012, pp. 391-392, 2012.

- Krityakien, O., Hirose, K., and, Minematsu, N., “F0 Contour Generation of Thai Speech Using the Tone Nucleus Model,” in Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, 2013. (To be published)