Generalization of Thai Tone Contour in HMM-Based Speech Synthesis

Anocha Rugchatjaroen*, Sittipong Saychum *, Keiichiro Oura† and Keiichi Tokuda†
*NECTEC, National Science and Technology Development Agency (NSTDA), Thailand
E-mail: anocha.rug, sittipong.say @nectec.or.th
†Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan
E-mail: uratec, tokuda @nitech.ac.jp

***Abstract*— This paper presents an improvement of vowel tone modeling from using an idea of generalization of tone contour by sharing F0 probability stream in the conventional HMM-Based Speech Synthesis system. As of the characteristic of parameter-based speech synthesis, the excitation source (F0) is highly significant to the correctness of synthetic speech especially for a tonal language like Thai. This paper is proposing a method of sharing F0 probability distributions within the same syllabic tones by tying them together before adjusting their probabilities. The implementation and evaluation was done on HTS platform which is based-on HTK. The results show that the proposed method can increase Thai tone correctness at about 14.1% objectively and 90% subjectively.**

Keywords— HMM-based Speech synthesis; F0 modeling; Tonal language; Thai

I. INTRODUCTION

Fundamental frequency (F0) contour is one of the most essential features in the acoustic of speech especially for a tonal language. The very first presentation of Thai tone analysis has been found in Bradley's 1911 [1], which depicts the frequency curves analyzed from Thai long-vowel syllables. Afterwards, there have been several studies in Thai tones which some of them are well known and widely referred to until these days such as Abramson (1962, 1979) [2, 3], Gandour, et al. (1991, 1999) [4, 5], Luksaneeyanawin (1989) [6] and Fujisaki, et al. (2003) [7], etc.

From a speech synthesis perspective, F0 modelling is one of the most important ingredients that strongly affects the correctness and prosody of synthetic speech. In both voice and voice-less phones, having F0 or not, their excitation contour together with their neighbouring's are significant [8, 9]. Algorithms proposed to model such the tasks are varies depending on many factors such as unit size (e.g. syllables or phrases), characteristics of languages (e.g. intonational or tonal languages), and the type of F0 models (e.g. parametric and non-parametric). The well-known parametric modelling techniques such as Fujisaki [7] and Tilt models [10, 11] has successfully been proved that they are great for representing a F0 contour. Conventional approaches often require a proper size of speech corpus for analyzing or for training the model. For example, a rich resource language such Mandarin Chinese has been successful in F0 modelling using various methods [12, 13, 14].

Parameters of the equations in modelling are analysed from speech corpora using machine-learning approaches such as Classification and regression tree (CART), etc. On the other hand, an F0 contour can also be a non-parametric one by selecting it directly from a speech corpus [15] but it cannot work with under-resourced languages. Without a speech corpus, F0 contours might be simply produced by rules but, again, the result is known to be unnatural. From all aspects and at the current stage of technology, F0 modelling from a sufficiently large corpus is essential for speech synthesis.

This work proposes a method to improve the accuracy of F0 modelling in Hidden Markov Model (HMM) based synthesis by sharing log-F0 contour probabilities within the same toneme, not phoneme, to establish the generalization of the same tone contours found in a training speech corpus. Section II describes tone contour in Thai. The HMM-based speech synthesis implementation is in Section III, then section IV explains the proposed method and its accuracies in section V, then the conclusion comes in section VI.

II. THAI SYLLABIC TONE CONTOUR

In general, a Thai tone means a Thai syllabic tone. As same as other tonal languages, syllabic tones affect the meaning of syllable/word. Traditionally, Thai people were thought to tune each syllable changing the meaning of them as a melody of the common, the primary, the secondary, the tertiary and the tetrad. Their F0 trajectories have been studies for several decades [2, 16]. They actually move in middlingly, loweringly, fallingly, higheringly and risingly and their toneme notations are /0/, /1/, /2/, /3/ and /4/, respectively. Their F0 contours approximately are as shown in Figure 1.

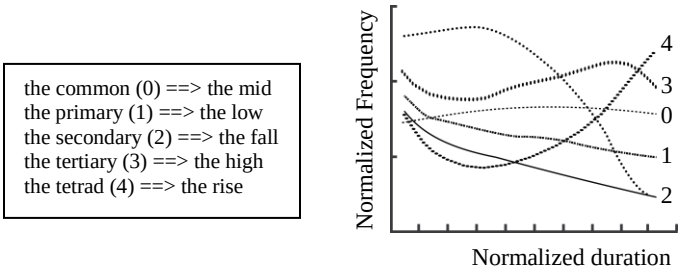


Fig. 1. Thai syllable tones (left) and Thai syllable tone contours (right) from [11].

Internationally speaking, their IPA are /◌̄/ for 0, /◌̀/ for 1, /◌̂/ for 2, /◌́/ for 3 and /◌̌/ for 4. Traditionally, Thai people were thought to tune each syllable changing the meaning of them as a melody. Examples of tuning different tones causing different

meaning are "คา" /kh-a-0/ (means to get stuck), /kh-a-1/ (means galangal), /kh-a-2/ (means to kill), /kh-a-3/ (means to trade), and /kh-a-4/ (means leg).

The five tones in Thai can be trajectorially divided into 2 group: the static group which consists of 3 tones, the high, the mid and the low; and the dynamic or kinetic group which consists of 2 tones, the rising and the falling.

The fall and the high are the two tones that have a dramatic change diachronically [2, 17]. It can be concluded that in isolative citation style the two main features that can be used to distinguish the five tones from each other are the pitch height, and the pitch direction. These two distinctive features need to be represented in terms of the movement of the F0 through the time dimension. Automatic procedure for extracting F0 from isolated monosyllabic source words either for speech synthesis or speech recognition still needs to be improved.

III. Another Implementation of Thai HMM-based Speech Synthesis

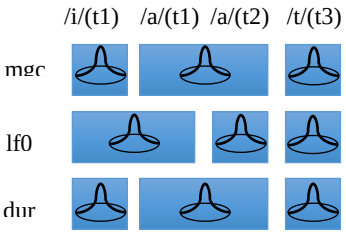
This research is based on HTS version 2.3 [17], which has been released in Dec 2015. HTS is a HMM-based speech synthesizing platform based-on modified modules of HTK [18]. It basically uses HMM to learn probabilities of speech acoustic and prosodic parameters and re-synthesizes those parameters back again from the generated statistics. With its well-designed structure, the platform provides a configuration file, which contains a useful switchboard for controlling stepwise modules.

In the implementation of the proposed system, some modules need to be modified. Before the modification, another proper Thai linguistic architecture analysis has to be re-processed in an attempt to fitting the language into the platform. With a successful implementation of English in a published example [19], Thai adopts some similar structures and adjusts some. The phoneme set is almost the same, just some phoneme codes changed, but the linguistic context-dependent question file are mainly adopted and used. As Thai is the tonal language and there is no stress mark in the training corpus, Thai Speech Synthesis Corpus-1 (TSynC-1) [20], the stress questions are converted into tone questions.

IV. Proposed Method

An important cue for predicting a F0 contour for exciting a syllable is the general trajectory of each syllabic tone type. In a tone type, the F0 contour of syllables are similar. This brings the hypothesis of this work to be "the usage of generalized F0 contour of each tone can improve the tone contour prediction in HMM-based synthetic speech."

The implementation of generalization was done under the meaningful of probabilistic distribution for lf0 in HTS-based architecture. Normally, HTS extracts and stores cepstral coefficients (mgc), duration (dur) and logarithmic F0 (lf0) contour of each phone separately. This work generalizes lf0 contours by tying lf0 stream of phones that have the same tone together and re-estimates the probability distribution when training models.



Pseudo-code:
1. Tying cepstral stream of monophone units
2. Tying log-f0 stream of pause, silence, consonants and tone
3. Tying duration stream of monophone units
4. Re-estimate (EM) all state & stream probabilities
5. Untying cepstral stream
6. Untying log-f0 stream
7. Untying duration stream

Fig. 2. Graphical idea of the proposed system (left) and its pseudo-code (right).

The graphical idea in Fig 2 shows an example of sharing lf0 probability when the training data have the same tone (t1) embedded in two different units of vowels (/a/ and /i/), which the proposed system wishes to share their probability distributions of their lf0 in the proposal to making them more generalized.

V. Evaluation and Results

This research used a set of Thai Speech Corpus named TSynC-1 [20]. It is a reading speech corpus distributed in 16k of sampling rate. It contains 5,200 sentences, uses 79 phonemes in labelling, which X of them is consonants and Y of them is vowels. Tones are marked in syllable level. In total, there are 246 distinct tone-dependent phonemes. The training set was the whole 5,200 sentences and the testing set was a randomly selected hundred sentences of them.

The evaluation was conducted for two main purposes, objective and subjective comparisons. The objective was set to depict differences between F0 trajectories of recorded and synthesized sound. The Euclidean distances from F0 contours of vowels in recorded speech to that of vowels in synthetic sounds were computed phone by phone using dynamic time warping [21, 22]. Figure 3 shows that the proposed system (TONE-TYING) produces a better F0 contour, i.e. a shorter distance to the reference speech, than that produced by the BASELINE system.

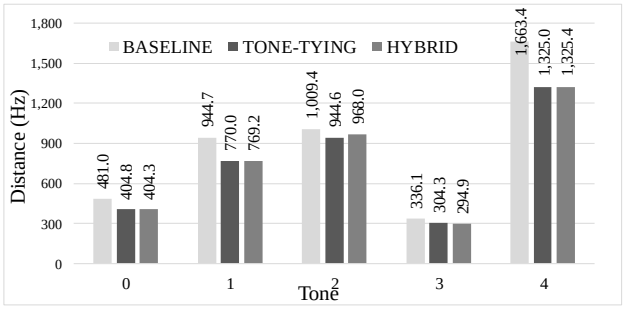


Fig. 3. Euclidean distance (Hz) between F0 contours of vowels in the record speech and that of vowels in synthetic ones using BASELINE, TONE-TYING and HYBRID methods.

The BASELINE results are from using one of Chompan's methods from his work in 2007, it uses a specific type of root in tone question tree to firstly classify types of tone trajectory before getting down to other questions [9]. The test set contains

100 Thai sentences selected randomly from TSynC-1. It consists of 966 vowels in total. There are 345, 212, 194, 145 and 70 units of syllabic tones 0, 1, 2, 3 and 4, respectively. The Figure 3 shows that the average distance frequencies of all vowels are shorten when using the proposed method (TONE-TYING). Moreover, HYBRID results, which are from applying BASELINE into TONE-TYING method, have longer distances than the results from using TONE-TYING method alone.

The subjective test was set in an AB comparison test. Ten sentences of synthetic speech using BASELINE and TONE-TYING methods were randomly presented to listeners in pairs. Eleven native Thai listeners aging between 31-39 years were asked to choose the sentence that has more tone correctness, syllable-tone counting, then specify the distorted-tone syllables. The results show that 9 out of 10 TONE-TYING sentences were preferred to be a better tone choice. Just one of them got almost the same preference as the BASELINE.

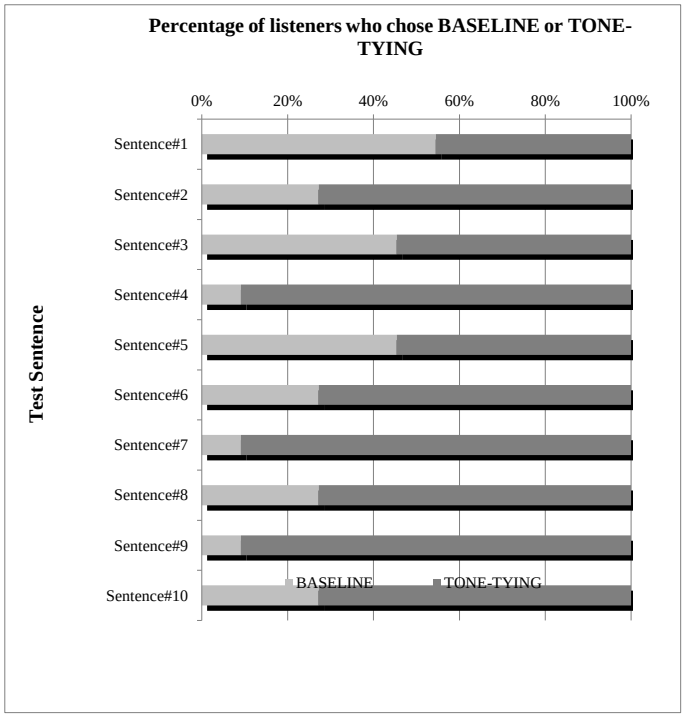


Fig. 4. Percentages of listener preferences counting from their selected choice.

The distorted-tone perception results are summarized separately syllable by syllable. When look at them in sentence level, they shows that there is at least one syllable-tone in all BASELINE sentences found distorted, but only two TONE-TYING sentences found distorted syllables (counting only when there are at least two listeners perceived the same distorted syllable). This means that listeners found more distorted syllables in BASELINE than in TONE-TYINE sentences.

VI. Conclusion and Future Works

In Thai, tying the log-F0 stream of vowels that having the same tone before re-estimate the probability in training step and then untying them before using them helps increasing the generalization of log-F0 stream of the vowel. This can improve the correctness of tone contour generation in the synthesized speech, which is shown in the experimental results.

In the future, another study of adding more linguistic conditions when tying tone stream can be explored such as considering the Thai death/live syllables which highly affects the syllabic tone, etc. Another interesting study is whether voiced/unvoiced phones are influential to the variation of syllabic tones, and how could we include such influent in the model.

Acknowledgment

This research was supported by CREST, JST. The authors would like to thank Dr. Chai Wutiwiwatchai for his helps and supports, all colleagues in Tokuda-Nankaku Laboratory (Japan), all in Speech and Audio Technology Laboratory (Thailand) and the anonymous Thai native listeners for their kind supports.

References

[1] C. Bradley, "Graphic analysis of the tone-accents of Siamese language," *Journal of American Oriental Society*, 1911.

[2] A. Abramson, "The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments," *Indiana U. Research Center in Anthropology, Folklore and Linguistics, Bloomington*, vol. 20, 1962.

[3] A. Abramson, "The coarticulation of tones: an acoustic study of Thai," in *Studies of Tai and Mon-Khmer phonetics*, 1979.

[4] J. Gandour, S. Potisuk, P. S. and S. Dechongkit, "Inter- and intraspeaker variability in fundamental frequency of Thai tones," *Speech Communication*, vol. 10, no. 4, pp. 355-372, 1991.

[5] J. Gandour, A. Tumtavitikul and N. Satthamnuwong, "Effects of speaking rate on Thai tones," *Phonetica*, vol. 56, pp. 123-134, 1999.

[6] S. Luksaneeyanawin, "A Thai Text to Speech System," in *the Region Workshops on Computer Processing of Asian Language*, Thailand, 1989.

[7] H. Fujisaki, S. Ohno and S. Luksaneeyanawin, "Analysis and synthesis of F0 contours of Thai utterances based on the command-response model.," in *International Congress of Phonetic Sciences*, 2003.

[8] C. Wutiwiwatchai and S. Furui, "Thai Speech processing technology: a review," *Speech Communication*, vol. 49, pp. 8-27, 2007.

[9] S. Chomphan and T. Kobayashi, "Tone correctness improvement i nspeaker dependent HMM-based Thai speech synthesis," *Speech Communication*, vol. 50, pp. 392-404, 2008.

[10] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697-1714, 2000.

[11] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, R. A. and S. S., "T-Tilt: A Modified Tilt Model for F0 Analysis and Synthesis in Tonal Languages," in *INTERSPEECH 2008*, Brisbane, 2008.

[12] Y. Qian, F. Soong, Y. Chen and M. Chu, "An HMM-based Mandarin Chinese Text-To-Speech System," in *Lecture Notes in Computer Science on Chinese Spoken Language Processing*, Berlin, Springer, 2006.

[13] Q. Sun, K. Hirose and N. Minematsu, "Two-Step Generation of Mandarin F0 Contours Based on Tone Nucleus and Superposition Models," in *6th ISCA Workshop on Speech Synthesis*, 2007.

[14] K. Hirose and J. Tao, Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis, Berlin: Springer, 2015.

[15] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech97*, Rhodes, Greece, 1997.

[16] K. Thepboriruk, "Bangkok thai tones revisited," Department of linguistics, University of Hawai'i at Manoa, Honolulu, 2009.

[17] HTS Working Group, Nagoya Institute of Technology, "HMM-based Speech Synthesis System (HTS)," 25 12 2015. [Online]. Available: http://hts.sp.nitech.ac.jp/?Download. [Accessed 16 05 2016].

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK Book (for HTK Version 3.4)," Cambridge University Engineering Department, 2006.

[19] P. Woodland, G. Evermann and M. Gales, "HTK History," Cambridge University Engineering Department (CUED), 2006. [Online]. Available: http://htk.eng.cam.ac.uk/docs/history.shtml. [Accessed 2014 12 15].

[20] C. Hansakunbuntheung, T. V. and V. Sornlertlamvanich, "Thai Tagged Speech Corpus for Speech Synthesis," in *Oriental COCOSDA*, 2003.

[21] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561-580, 2007.

[22] A. Rilliard, A. Allauzen and P. Mareuil, "Using Dynamic Time Warping to compute prosodic similarity measures," in *INTERSPEECH*, 2011.