Journal of Computer Science 5 (12): 905-914, 2009
ISSN 1549-3636

Towards the Development of Speaker-Dependent and Speaker-Independent Hidden Markov Model-Based Thai Speech Synthesis

Suphattharachai Chomphan
Department of Electrical Engineering, Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand

Abstract: Problem statement: Tone distortion in Thai languages can deteriorate not only the intelligibility of speech but also its naturalness. Therefore, the correctness of tone must be carefully taken into account in continuous speech synthesis. The preliminary work confronted this problem when applying HMM-based speech synthesis to Thai. **Approach:** This study presented a study on speaker-dependent and speaker-independent Hidden Markov Model (HMM)-based Thai speech synthesis. In the speaker-dependent system, we developed a simple tone-separated tree structure in the tree-based context clustering process of the training stage to treat the tone distortion problem. In the speaker-independent system or averaged-voice-model system, a number of tonal features are extracted and applied with the Speaker Adaptive Training (SAT) and Shared Decision Tree (STC) techniques to release the tone distortion problem. **Results:** Our objective evaluation revealed that the proposed features could make the F0 contour closer to the target speaker's real contour. The results from our subjective test also revealed that the proposed tonal features could improve the tone intelligibility of all speech-model scenarios of male and female. **Conclusion:** By applying our approach, the problem of tone distortion can be relieved effectively. The better tone correctness can improve the intelligibility and the naturalness of speech significantly.

Key words: Tone correctness, speaker-dependent, speaker-independent, hidden Markov models, speech synthesis

INTRODUCTION

Historically, Thai speech synthesis has been widely developed in two approaches. Unit-selection-based approach has been conducted at the beginning period with high speech quality in both naturalness and intelligibility, while HMM-based approach has just been studied in 2007[1,2]. The first paper describing the development of a Thai TTS engine was published in 1983[3], where a speech unit concatenation algorithm was applied to Thai. This approach was implemented in the latest version of Vaja[4] at National Electronics and Computers Technology Center (NECTEC). Although the newest Vaja engine produces a much higher sound-quality than the former unit-concatenation based engine, the synthetic speech sometimes sounds unnatural, especially when synthesizing non-Thai words written with Thai characters and still cannot synthesize speech with various voice characteristics such as speaker individualities, speaking styles. To achieve various voice characteristics in speech synthesis systems based on this approach, a large amount of speech data is necessary. Unfortunately, it is burdensome to obtain enough speech data. In order to treat this problem, an HMM-based speech synthesis which has been originally developed to support Japanese has been adapted for Thai by Chomphan in 2007.

The main purpose of speech synthesis is to produce sounds of speech that are intelligible and natural. Modeling spectral and prosodic features suitably would bring about better speech quality. Tone distortion in tonal languages, e.g., Thai, can deteriorate not only the speech intelligibility but also its naturalness, since the lexical tone is a suprasegmental feature formed by the basic prosodic feature. Therefore, the tone correctness should be considered in generating continuous speech. Tone correctness was successfully improved in a speaker-dependent HMM-based synthesis of Thai speech[5]. It has been found that the implemented system could provide speech with better reproduction of prosody over the unit-selection-based Vaja TTS system. Specifically, a decision tree with a tone-separated structure significantly improved the tone correctness of the synthesized speech. After that, a speaker-independent HMM-based Thai speech synthesis system with a speech database containing a large number of speakers with a small amount of data for each speaker has been

developed by Chomphan and Kobayashi to generate speech with various speaker characteristics. However, the tone correctness was inadequate despite using the technique that was adopted in the speaker-dependent system. Moreover, the shared decision-tree context clustering technique (STC)[6] was adopted to reduce the effect of speaker bias and the Speaker Adaptive Training technique (SAT)[7,8] was incorporated into the training procedure of the average-voice model to improve its quality. However, the critical problem of improving tone correctness was still not resolved. Thus, some tonal features were included to treat the problem[9].

This study is structured as follows. In the materials and methods, Thai tone characteristics are addressed. The implementation of the speaker-dependent HMM-based speech synthesis system and the implementation of the speaker-independent HMM-based speech synthesis system are subsequently explained. The results and discussion are then presented. Finally, conclusions and potential research directions are given at the end of this study.

MATERIALS AND METHODS

Thai tone characteristics: In Thai, tone, which is indicated by contrasting variations in contour of fundamental Frequency (F0) at the syllabic level, is an important part of spoken language because the meaning of words with the same sequence of phonemes can be different if they have different tones. There are five tonal variations traditionally named according to the characteristics of their F0 contours within a syllable[10].

In the continuous speech context, the F0 patterns of 5 Thai tones are affected from the adjacent syllable tones[11]. Palmer[12] demonstrated that 5 Thai tones showed some changes in height and slope as a function of the preceding or following tone. Changes in height and slope appeared to be confined primarily to the beginning or end of the syllable[13]. Gandour[14] studied the tonal coarticulation including the carry-over effects and the anticipatory effects. These studies indicate the complicate attributes of Thai tones.

Implementation of the speaker-dependent HMM-based speech synthesis system:

Implementation process and basic configuration: A basic structure of the HMM-based TTS system is shown in Fig. 1. There are two main stages including training stage and synthesis stage.

In the training stage, context dependent phoneme HMMs are trained by using a speech database. Spectral parameter and excitation parameter (F0) are extracted at each analysis frame as the static features from the speech database in the spectral parameter extraction and excitation parameter extraction modules, respectively. Thereafter, they are modeled by multi-stream HMMs in which output distributions for the spectral and F0 parts are modeled by using a continuous probability distribution and the multispace probability distribution (MSD)[15], respectively. In addition, to directly model the phone durations, we utilize a framework of Hidden Semi-Markov Model (HSMM)[16], where the model has explicit state duration distributions instead of the transition probabilities. To model variations in the spectrum and F0, we take into account phonetic, prosodic and linguistic contexts, such as phoneme identity contexts, tone-related contexts and locational contexts. Then, the decision-tree-based context clustering technique[17] is applied separately to the spectral and the F0 parts of the context-dependent phoneme HMMs.

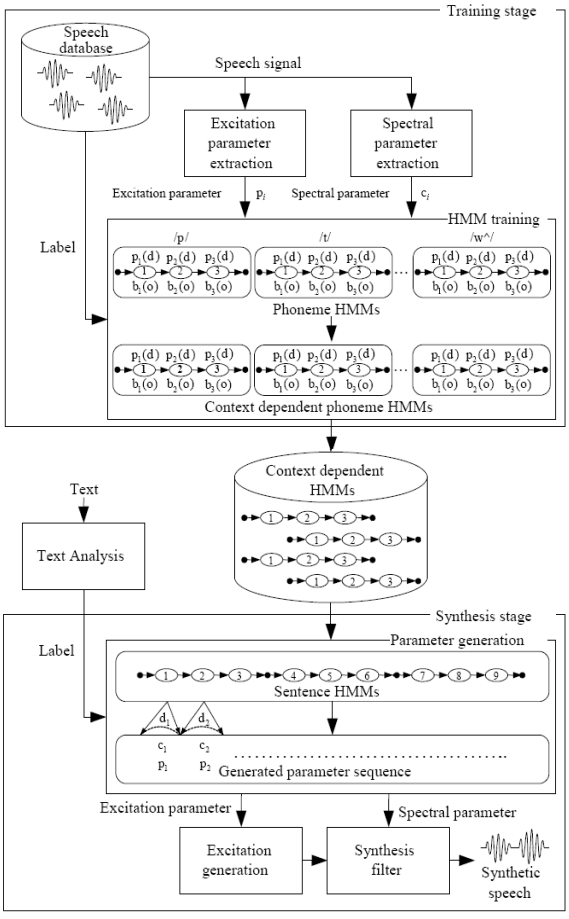

Fig. 1: HMM-based speech synthesis system

In the clustering technique, a decision tree is automatically constructed based on the Minimum Descriptive Length (MDL) criterion. Thereafter, the re-estimation processes of the clustered context-dependent phoneme HMMs are conducted by using the Baum-Welch (EM) algorithm. Finally, state durations are modeled by a multivariate Gaussian distribution[18] and the state clustering technique is also applied to the state duration models.

In the synthesis stage, an arbitrarily given text is transformed into a sequence of context-dependent phoneme labels. A sentence HMM is constructed by concatenating context-dependent phoneme HMMs based on the label sequence. From the sentence HMM, spectral and F0 parameter sequences are obtained based on the Maximum Likelihood (ML) criterion, where the associated phoneme durations are determined by using state duration distributions. Eventually, the output speech is synthesized from the generated mel-cepstral and F0 parameter sequences by using an MLSA (Mel Log Spectral Approximation) filter[19].

Design of decision tree in context clustering: Context dependent models considering several combinations of contextual factors are constructed in the training stage. However, as the number of contextual factors increases, their combinations also increase exponentially. As a result, model parameters with sufficient accuracy cannot be estimated with limited training data. In other words, it is impossible to prepare the speech database which includes all combinations of contextual factors. To release this problem, the decision-tree based context clustering technique is employed to the distributions of the associated speech features.

At the beginning, we used a conventional single binary tree structure in the decision tree-based context clustering process as shown in Fig. 2a. Due to the imbalance of tone frequency, some tones dominate the tree over the others. As a result, the single binary tree context clustering gives high tone error percentage or about 20% when using 2500 training utterances. It causes an unacceptable intelligibility to the synthesized speech. An obvious example of F0 contour distortion between the natural speech and the synthesized speech can be shown in Fig. 3 (at the first full-shape syllable contour). To improve the tone correctness of the synthesized speech, the simple tone-separated decision-tree structure was designed as depicted in Fig. 2b. It is supposed that separating the structure into sub-tree for each tone can reduce the influences across tones.

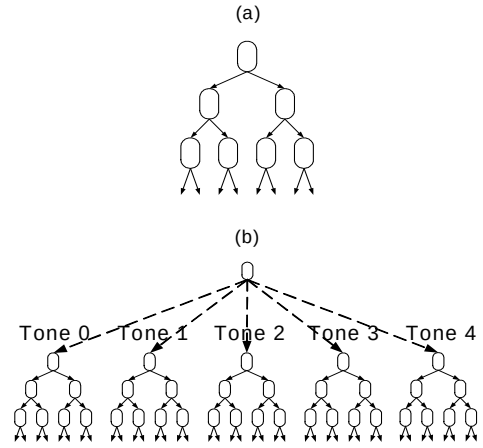

Fig. 2: Tree structures for context clustering: (a) single binary tree structure, (b) simple tone-separated tree structure

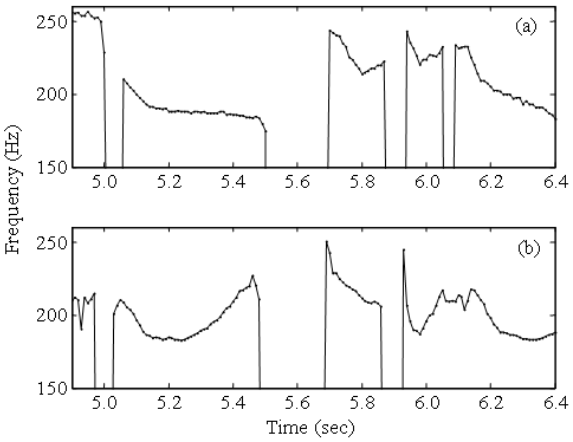

Fig. 3: F0 contours of (a) synthesized speech from the clustering style of single binary tree without tone type questions and (b) natural speech

In the evaluation process, four context clustering styles are designed as follows:

- Single binary tree context clustering without tone type questions
- Simple tone-separated tree context clustering without tone type questions
- Single binary tree context clustering with tone type questions
- Simple tone-separated tree context clustering with tone type questions

Implementation of the speaker-independent HMM-based speech synthesis system:

Implementation process and basic configuration: The conventional HMM-based speech synthesis system

which adopted from the speaker-dependent system[5] embedding the STC technique with a tone-separated tree structure and SAT technique[8] is called baseline training system. A vital problem for generating Thai average-voice speech from baseline training system is tone distortion. In other words, the tone correctness of synthetic speech is considerably degraded. This problem does not exist in the speaker-dependent system because of using single-speaker voice for training. The problem emerges in the speaker-independent system where a multi-speaker speech database is applied. The tying mechanism in decision-tree-based context clustering without an appropriate criterion causes an unexpected phenomenon called tone neutralization. For example, when a short vowel with no final consonant appears as the first element of a compound word, this is often accompanied by a neutralization of tone especially when the original tone is low[20]. Tone neutralization for our context has frequently occurred in synthetic speech and caused several vital tone distortions. Fig. 4 compares the F0 contour of natural speech and the F0 contour generated from the baseline training system for the Thai sentence /κηρ-ο⌉-N^-0 κ-α⌉-ν^-0 ν-ι⌉-?^-3 φ-υ⌉-?^-1 ρ-α-?^-3 ω-α⌉-N^-1 κ-α⌉-ν^-0/ (the hat ^ and number correspond to the final consonant and tone number), which is not included in the training sentences. Tone neutralization can be observed in the lower contour of the conventional approach compared with the upper contour of natural speech.

Integrating a set of tonal features that consist of phrase-intonation and tone-geometrical features into the context-clustering process of the HMM-based speech synthesis system was proposed to treat the problem of tone neutralization[9]. They are expected to be promising factors to reduce the effect of modeling F0 from a variety of speaker characteristics, since HMMs with an associated F0 contour are clustered and share their parameters. Generally, phrase-intonation features represent the F0 contour on the global level, while tone-geometrical features reflect the F0 contour on the local level. Both of them complementarily represent the F0 contour and are derived from an extension of the generative and geometrical models.

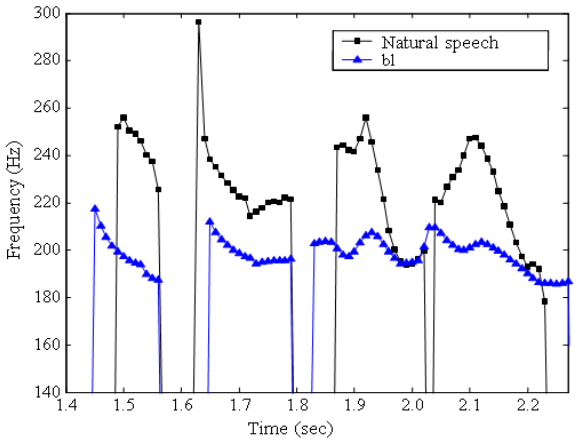

Fig. 4: Comparison of F0 contour of natural speech and F0 contour generated from a baseline training system (bl)

The generative model depicted in Fig. 5 effectively represents the F0 contour of speech in both tonal and non-tonal languages[21-24], therefore we applied it to our research on both the global and local levels. As the geometrical model is a simple method of representing a small portion of the F0 contour, we chose it to model the F0 contour of syllables on the local level.

Integration of tonal features into HMM-based speech synthesis system: we explain how to integrate the tonal features into the baseline training system. The module for extracting tonal features is integrated into the conventional HMM-based speech synthesis system embedding the STC technique with a tone-separated tree structure and SAT technique, as depicted in Fig. 6. The highlighted dark components represent modifications from the conventional system.

Tonal features in the form of codewords are employed in all stages of the systems. In the training stage, tonal feature extraction module adopts the F0 values in a series of speech utterances from the excitation parameter extraction module, the time label that provides the boundary of all syllables in the corresponding speech utterance and the context label that contains contextual information from associated texts. A set of tonal features is subsequently transformed into a set of appropriated codewords. They are finally embedded into the contextual information and an output context label is applied to the tree-based context-clustering process in the HMM training module.

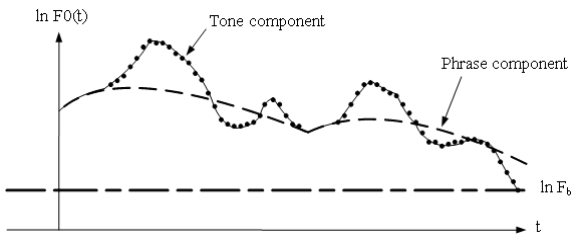

Fig. 5: Representation of F0 contour by generative model

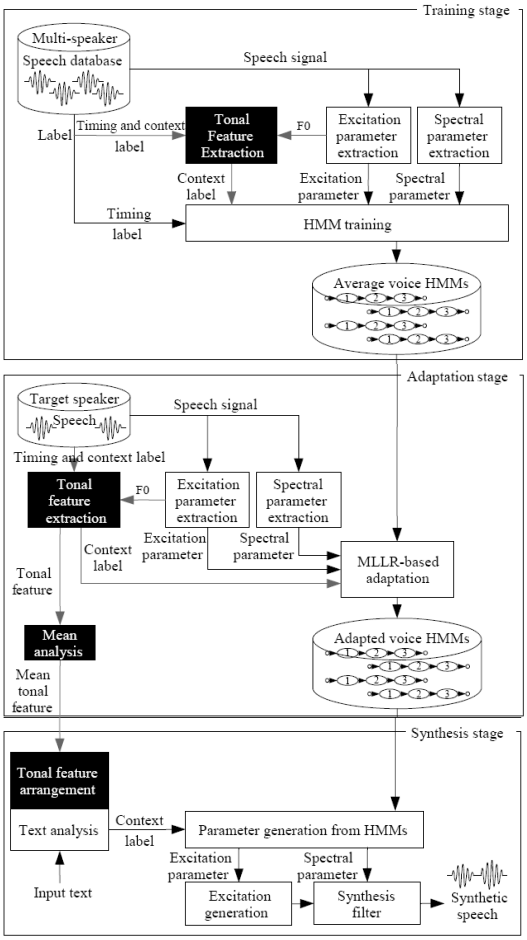

Fig. 6: Block diagram of a tonal-feature-embedded HMM-based speech synthesis system.

The implemented decision tree is binary, where a question splitting contexts into two sub-groups is prepared within each intermediate node and the MDL criterion is used to stop the nodes from splitting[17]. An associated node can be selected for all contexts by traversing the tree, starting from the root node, then selecting the next node depending on the answer to a question about the current context. Therefore, once the decision tree is constructed, unseen contexts can be prepared[18,25].

We arrange the tonal features of target-speaker speech in the adaptation stage by using the same procedure as the training stage and exploit them in the MLLR-based adaptation module. Finally, the representative mean values of the tonal features of the target-speaker speech are calculated by the mean analysis module and then embedded with other contextual information analyzed from the input text to form a context label by tonal feature arrangement and text analysis module in the synthesis stage. It is necessary to include this process in the system, since the embedded representative mean values of the tonal features will reflect the target speaker's characteristics. We update these tonal features for every new target speaker, since they reflect the speaker's individuality.

The speaker-independent system generates the arbitrary target speaker's speech from the average-voice model by conducting the speaker adaptation. The preliminary experiments showed that the tone correctness of average-voice speech as well as adapted-voice speech was degraded. Therefore, the tonal features were incorporated into the speaker-independent system in both the training and adaptation stages. Without incorporating tonal features into the adaptation stage, the tonal features of the target speaker would have been ignored and may have caused unexpected tone distortion.

Arrangement of contextual information: A number of contextual factors that affect the spectrum, F0 pattern and duration, e.g., phoneme identity factors and locational factors, are prepared the same as those used in the speaker-dependent system[5]. They are divided into five levels of speech units, including phoneme, syllable, word, phrase and utterance.

The extraction algorithms for tonal features[9] were used with the F0 series of all training utterances to prepare the tonal features to be employed in the context-clustering process. Each of the tonal-feature ranges determined from analyzing the tonal features is equally divided into several sub-ranges and then the quantization process is applied. The baseline value of F0 and the amplitude of the phrase command for the phrase-intonation features were linearly quantized into eight classes with an assigned codeword of 0-7. These features were then grouped into two sets (S15, S16) at the phrase level of contextual factors as shown in the following list. It is noted that our purpose is to indicate the level of phrase intonation for the current phoneme; therefore, both features have to be used together. As a result, the feature of the baseline value of F0 is not classified into the utterance level, although each utterance has its own unique value.

The initial F0 of the syllable, its duration, its slope and the amplitude of the tone command for the tone-geometrical features were linearly quantized in the same way as that applied to the phrase-intonation features. These features were then grouped into four sets (S6-S9) in the syllable level. Since the current-tone characteristics greatly depend on its adjacent tones; in other words, these are known as tonal coarticulation

effects, which include carry-over and anticipatory effects[12,14]. Therefore, we also provided the contextual factors for these features with preceding, current and succeeding syllable positions.

Phoneme level:
- S1: {preceding, current, succeeding} phonetic type
- S2: {preceding, current, succeeding} part of syllable structure

Syllable level:
- S3: {preceding, current, succeeding} tone type
- S4: Number of phonemes in {preceding, current, succeeding} syllable
- S5: Current phoneme position in current syllable
- S6: {preceding, current, succeeding} codeword of initial F0 of syllable
- S7: {preceding, current, succeeding} codeword of syllable duration
- S8: {preceding, current, succeeding} codeword of syllable slope
- S9: {preceding, current, succeeding} codeword of amplitude of tone command

Word level:
- S10: Current syllable position in current word
- S11: Part of speech of current word
- S12: Number of syllables in {preceding, current, succeeding} word

Phrase level:
- S13: Current word position in current phrase
- S14: Number of syllables in {preceding, current, succeeding} phrase
- S15: Codeword of baseline value of F0
- S16: Codeword of amplitude of phrase command

• Utterance level:
- S17: Current phrase position in current sentence
- S18: Number of syllables in current sentence
- S19: Number of words in current sentence

RESULTS

Experiments for speaker-dependent HMM-based speech synthesis system:

Speech database and training condition: A Thai speech database named TSynC-1 from NECTEC[4] was used for training HMMs. The speech in the database was uttered by a professional female speaker with clear articulation and standard Thai accent. The phoneme labels included in TSynC-1 and the utterance structure from ORCHID text database were used to construct the context dependent labels with 79 different phonemes including silence and pause in the case of tone-independent phonemes and 246 different phonemes including silence and pause.

Speech signal were sampled at a rate of 16 kHz and windowed by a 25 ms Blackman window with a 5 ms shift. Then mel-cepstral coefficients were extracted by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of F0 and their delta and delta-delta coefficients.

We used 5 state left-to-right HSMMs (hidden semi-Markov models) in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution[26]. Using the HSMMs, the explicit state duration probability is incorporated into HMMs and the state duration probability is reestimated by using EM algorithm[27]. It is noted that each context dependent HSMM corresponds to a phoneme-sized speech unit. The number of training utterances was varied as follows: 100, 200, 300, 400, 500, 1000, 1500, 2000 and 2500.

Evaluation of tone correctness: we present how the correctness of the synthesized tone is improved by using the constructed contextual factors and four different tree-based context clustering styles. Fig. 7 shows an example of F0 contours of the natural speech and the synthesized speech with different clustering styles. The first full-shape syllable of Fig. 7 conveys tone 4 or rising tone. Figure 7a is of the single tree context clustering without tone questions, however this syllable contour is misshaped.

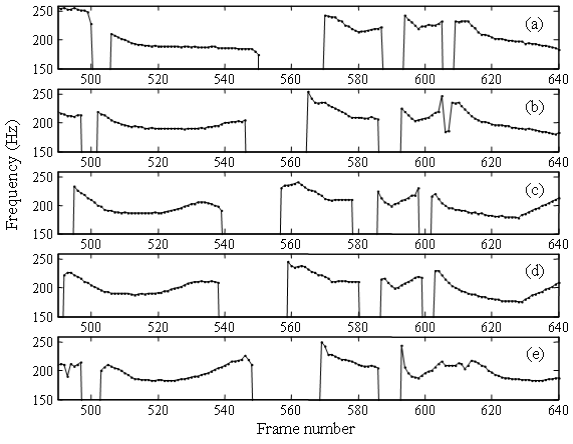

Fig. 7: F0 contours of synthesized speech from 4 different clustering styles; (a) single tree without tone type questions, (b) single tree with tone type questions, (c) tone-separated tree without tone type questions, (d) tone-separated tree with tone type questions and (e) F0 contour of natural speech

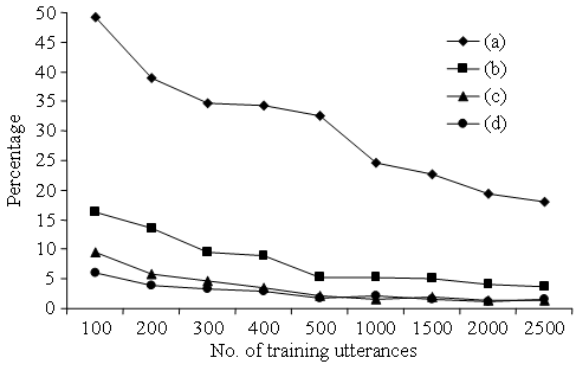

Fig. 8: Tone error percentages of synthesized speech from 4 different clustering styles; (a) single tree without tone type questions, (b) single tree with tone type questions, (c) tone-separated tree without tone type questions and (d) tone-separated tree with tone type questions

As a result, most listeners perceive it with wrong tone. Meanwhile Fig. 7b-d are of the other styles and they show the improvement of the F0 contour shape conforming to that of the natural speech as depicted in Fig. 7e. We investigated 2,289 syllables from 100 synthesized speech utterances to evaluate tone correctness of our system. The tone error percentages for all clustering styles are summarized in Fig. 8.

Experiments and discussions for speaker-independent HMM-based speech synthesis system:

Speech database and training condition: A set of phonetically balanced sentences from the Thai-speech database called LOTUS from NECTEC[28] was used to train the HMMs. TSynC-1 was used for the adaptation of a female target speaker and also for a speaker-dependent system. Another set of phonetically balanced 500 sentences of Thai-male speech was used for the adaptation of a male target speaker. In LOTUS, speech was uttered by 24 female and 24 male speakers with clear articulation and a standard Thai accent.

The average voice model was trained using 35 sentences for each speaker from the 24 female and 24 male speakers from the LOTUS speech data, i.e., the total number of training utterances was 1,680. We constructed two different speech scenarios to evaluate the proposed approach: male and female models. They are defined as follows:

- Male-model scenario: The average-voice model was trained using 840 male speech utterances; i.e., 35 sentences for each speaker from the 24 male speakers of the LOTUS speech data. The adaptation data were from 35 training utterances selected from the local set of the male target speaker
- Female-model scenario: The average-voice model was trained using 840 female speech utterances; i.e., 35 sentences for each speaker from the 24 female speakers of the LOTUS speech data. The adaptation data were from 35 training utterances selected

Comparison of tone intelligibility was conducted for male- and female-model scenarios in the subjective evaluations. The synthetic speech of a speaker-dependent system with 1,500 training utterances is evaluated for comparison. Both the average voice and the adapted voice were used in both evaluations. An objective evaluation by using RMS logarithmic F0 error was done in parallel. MLLR-based speaker adaptation[29] was used to generate the adapted voice.

The entries for “male” and “female” correspond to male- and female-model scenarios in the following results. The entries for “avg.” and “adt.” correspond to average-voice speech and adapted-voice speech. The entry for “sd.” corresponds to the speech generated from the speaker-dependent model. The entries for “bl”, “bl+pi” and “bl+pi+tg” correspond to baseline training, baseline training with phrase-intonation features and baseline training with phrase-intonation features and tone-geometrical features. Since tone-geometrical features are minor components that are complementary to phrase-intonation features, the case of "bl + tg" has been ignored.

Tone intelligibility for male-speech and female-speech models: The results show how the overall tone correctness of the average voice and adapted voice is improved by embedding the tonal features into the system with different scenarios of speech models for males and females. The percentage in tone error[2] is the measured value in this comparison. A subjective test was performed to calculate the tone-error percentage in our implemented systems. The subjects were eight native Thai speakers. For each subject, thirteen sets of 100 tested sentences (i.e., 2,289 syllables), which were not contained in the training data, were presented. Twelve sets were generated from the speaker-independent system conducted by varying the speech-model scenarios (male and female), the kinds of voice (average and adapted) and the training approaches (“bl”, “bl+pi” and “bl+pi+tg”), while another reference set was generated from the speaker-dependent system. The subjects were asked to decide whether the syllables had tones corresponding with

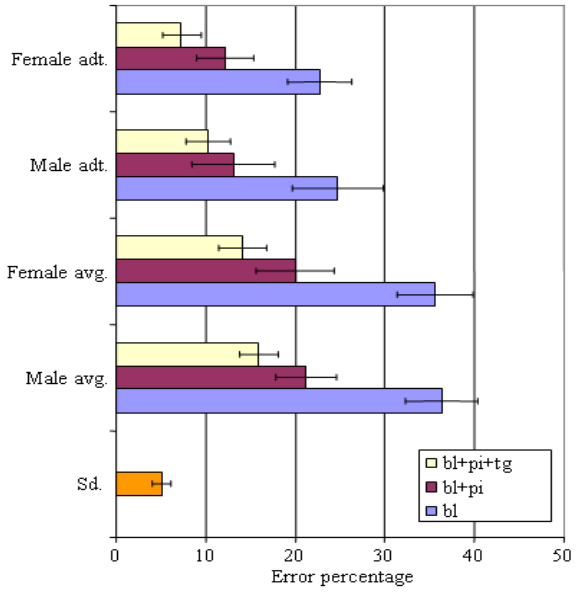

Fig. 9: Tone error percentage of average voice and adapted voice for male and female speech models synthesized from different training approaches

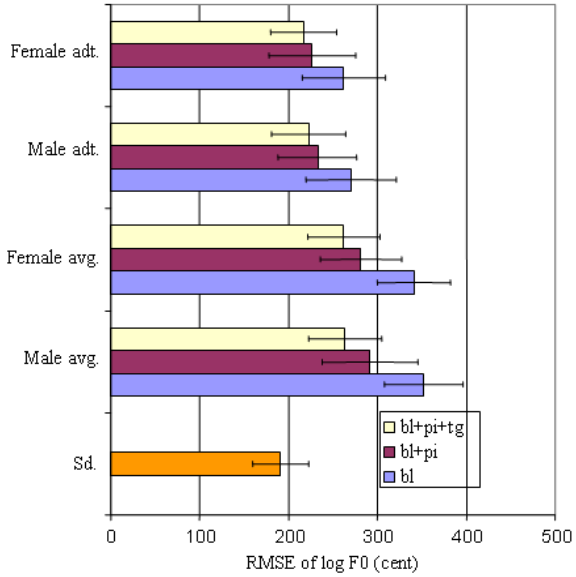

Fig. 10: RMS logarithmic F0 error in (cent) for 100 test sentences of average voice and adapted voice for male and female speech models synthesized from different training approaches

the given texts or not. The average tone-error percentages with a 95% confidence interval for different training styles are summarized in Fig. 9.

F0 error analysis for male-speech and female-speech models: an objective evaluation by using RMS logarithmic F0 error which is performed in parallel with the subjective evaluation is explained. Fig. 10 shows the RMS logarithmic F0 error between generated logarithmic F0 and that extracted from a target speaker's real utterances for both scenarios.

DISCUSSION

Discussions for speaker-dependent HMM-based speech synthesis system: From Fig. 8, the significant reduction of the tone error percentage of the second style comparing with the first style can be seen. It indicates that the tone type questions play a very important role in the generation of F0 contour. The third style can further reduce the error percentage, while the last style gives the error percentage closed to that of the third one. In other words, the separation of tree has more effectiveness than using only simple tone type questions. Considering the number of training utterances, the tone error percentage is decreased as the number of training utterances is increased. It is noted that some distortions of the generated syllable duration are unavoidable when using the tone-separated tree context clustering with small training data due to the limited data in each tone. However, it can be relieved when the number of training utterances is raised over 500.

Discussions for speaker-independent HMM-based speech synthesis system: From Fig. 9, it can be seen that the proposed tonal features can reduce the tone-error percentage using the baseline-training approach more than the phrase-intonation features. As for the F0 error analysis, the state duration of each model was adjusted after Viterbi alignment with the target's speaker real utterances to calculate the error. Since no F0 values were observed in the unvoiced region, the RMS logarithmic F0 error was calculated in the region where both the generated and real contours were voiced. From Fig. 10, comparing the average voice and the adapted voice, it is obvious that the RMSEs of the former are rather larger than those of the latter. From both male- and female-speech-model scenarios, the RMSEs of baseline training with phrase-intonation features (bl+pi) are smaller than those of baseline training (bl). It can also be seen that the proposed tonal features (bl+pi+tg) make the F0 contour closer to the target speaker's real contour than that of baseline training with phrase-intonation features for both scenarios. These results confirm the improvements caused by using the proposed tonal features.

CONCLUSION

An approach of HMM-based Thai speech synthesis is presented in this study. First, the speaker-dependent system was implemented with high tone intelligibility when using a simple tone-separated tree context clustering. Subsequently the speaker-independent system was studied. A group of tonal features including phrase-intonation features and tone-geometrical features were proposed to be embedded in the contextual factors for the context-clustering process of a speaker-independent HMM-based Thai speech synthesis system to treat the problem of tone neutralization. These features were extracted based on optimizing the parameters of the generative model and extracting the geometrical parameters. Our objective evaluation revealed that the proposed features could make the F0 contour closer to the target speaker's real contour. The results from our subjective test also revealed that the proposed tonal features could improve the tone intelligibility of all speech-model scenarios of male and female.

For the future direction, we focus on the study of generating expressive speech. Moreover, applying our approach to other tonal languages is possible thanks to the independence of language in our F0 modeling.

ACKNOWLEDGEMENT

The researcher is grateful to NECTEC. Without LOTUS and TSynC-1 speech databases from NECTEC, we could not achieve this study.

REFERENCES

1. Chomphan, S. and T. Kobayashi, 2007. Design of tree-based context clustering for an HMM-based Thai speech synthesis system. Proceeding of the 6th ISCA Workshop on Speech Synthesis, Aug. 2007, ISCA, Bonn, Germany, pp: 160-165. http://www.isca-speech.org/archive/ssw6/ssw6_160.html
2. Chomphan, S. and T. Kobayashi, 2007. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceeding of the 8th Annual Conference of the International Speech Communication Association, Aug. 2007, Antwerp, Belgium, pp: 2849-2852. http://www.isca-speech.org/archive/interspeech_2007/i07_2849.html
3. Saravari, C. and S. Imai, 1983. A demisyllable approach to speech synthesis of Thai, a tonal language. J. Acoust. Soc. Jap., 4: 97-106. http://ci.nii.ac.jp/naid/110003105666/en
4. Hansakunbuntheung, C., A. Rugchatjaroen and C. Wutiwiwatchai, 2005. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. Proceeding of the International Symposium on Natural Language Processing, Dec. 2005, Bangkok, Thailand, pp: 127-132, http://www.hlt.nectec.or.th/publications.php
5. Chomphan, S. and T. Kobayashi, 2008. Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. Speech Commun., 50: 392-404. DOI:10.1016/j.specom.2007.12.002
6. Yamagishi, J., M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, 2002. A context clustering technique for average voice model in HMM-based speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Sept. 2002, Colorado, USA., pp: 133-136. http://www.isca-speech.org/archive/icslp_2002/i02_0133.html
7. Anastasakos, T., J. McDonough, R. Schwartz and J. Makhoul, 1996. A compact model for speaker adaptive training. Proceeding of the International Conference on Spoken Language Processing, Oct. 1996, Philadelphia, USA., pp: 1137-1140. DOI: 10.1109/ICSLP.1996.607807
8. Yamagishi, J., T. Masuko, K. Tokuda and T. Kobayashi, 2003. A training method for average voice model based on shared decision tree context clustering and speaker adaptive training. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 2003, Hong Kong, pp: 716-719. DOI: 10.1109/ICASSP.2003.1198881
9. Chomphan, S. and T. Kobayashi, 2009. Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis. Speech Commun., 51: 330-343. DOI:10.1016/j.specom.2008.10.003
10. Thathong, U., S. Jitapunkul and V. Ahkuputra, 2000. Classification of Thai consonants naming using Thai tone. Proceeding of the International Conference on Spoken Language Processing, Oct. 2000, Beijing, China, pp: 47-50. http://www.isca-speech.org/archive/icslp_2000/i00_3047.html
11. Wutiwiwatchai, C. and S. Furui, 2007. Thai speech processing technology: A review. Speech Commun., 49: 8-27. DOI: 10.1016/j.specom.2006.10.004
12. Palmer, A., 1969. Thai tone variants and the language teacher. Language Learn., 19: 287-299. DOI: 10.1111/j.1467-1770.1969.tb00469.x

13. Abramson, A.S., 1979. Lexical tone and sentence prosody in Thai. Proceeding of the International Congress of Phonetics Science, Aug. 1979, Copenhagen, Denmark,. pp: 380-387.
14. Gandour, J.T., S. Potisuk and S. Dechongkit, 1994. Tonal coarticulation in Thai. J. Phonet., 22: 477-492.
15. Tokuda, K., T. Masuko, N. Miyazaki and T. Kobayashi, 1999. Hidden Makov models based on multi-space probability distribution for pitch pattern modeling. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 1999, Phoenix, USA., pp: 229-232. DOI: 10.1109/ICASSP.1999.758104
16. Levinson, S.E., 1986. Continuously variable duration hidden Markov models for automatic speech recognition. Comput. Speech Language, 1: 29-45. DOI: 10.1016/S0885-2308(86)80009-2
17. Shinoda, K. and T. Watanabe, 2000. MDL-based context dependent subword modeling for speech recognition. J. Acous. Soc. Jap., 21: 79-86. DOI: 10.1250/ast.21.79
18. Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1998. Duration modeling for HMM-based speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Dec. 1998, Sydney, Australia, pp: 29-32. http://www.shlrc.mq.edu.au/proceedings/icslp98/PDF/AUTHOR/SL980939.PDF
19. Imai, S., K. Sumita and C. Furuichi, 1983. Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. IECE Trans. Fundam., J66-A: 122-129. (In Japanese). DOI :10.1002/ecja.4400660203
20. Iwasaki, S. and I.H. Horie 2005. A Reference Grammar of Thai. Cambridge University Press, Cambridge, ISBN: 0521650852, pp: 392.
21. Fujisaki, H. and H. Sudo, 1971. A model for the generation of fundamental frequency contours of Japanese word accent. J. Acoust. Soc. Jap., 57: 445-452. http://ci.nii.ac.jp/naid/110003107854/en
22. Fujisaki, H. and K. Hirose, 1984. Analysis of voice fundamental frequency contours for decorative sentence of Japanese. J. Acoust. Soc. Jap., 5: 133-142.
23. Fujisaki, H., K. Hirose, P. Halle and H. Lei, 1990. Analysis and modeling of tonal features in polysyllabic words and sentences of the standard Chinese. Proceeding of the International Conference on Spoken Language Processing, Nov. 1990, Kobe, Japan, pp: 841-844. http://www.citeulike.org/user/gpk/article/2946374
24. Fujisaki, H. and S. Ohno, 1998. The use of generative model of F0 contours for multilingual speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Dec. 1998, Sydney, Australia, pp: 714-717.
25. Riley, M.,1989. Statistical tree-based modeling of phonetic segment durations, J. Acoust. Soc. Am. 85, pp: S44-S44, May 1989 DOI: 10.1121/1.2026979
26. Zen, H., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 2004. Hidden semi-Markov model based speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Oct. 2004, Jeju Island, Korea, pp: 1393-1396. http://www.isca-speech.org/archive/interspeech_2004/i04_1393.html
27. Russell, M.J. and R.K. Moore, 1985. Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 1985, Tampa, USA., pp: 5-8. DOI: 10.1109/ICASSP.1985.1168477
28. Kasuriya, S., V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara and N. Thatphithakkul, 2003. Thai speech corpus for Thai speech recognition. Proceeding of the Joint International Conference of SNLP-Oriental June, 2003, Singapore, pp: 54-61. http://www.hlt.nectec.or.th/speech/download/software/view/cocosda2003.pdf
29. Yamagishi, J., T. Masuko and T. Kobayashi, 2004. MLLR adaptation for hidden semi-Markov model based speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Oct. 2004, Jeju Island, Korea, pp: 1213-1216. http://www.isca-speech.org/archive/interspeech_2004/i04_1213.html