American Journal of Applied Sciences 9 (3): 358-364, 2012
ISSN 1546-9239

Improvement of Tone Intelligibility for Average-Voice-Based Thai Speech Synthesis

[1]Suphattharachai Chomphan and [2]Chutarat Chompunth
[1]Department of Electrical Engineering,
Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6,
Tungsukhla, Si Racha, Chonburi, 20230, Thailand
[2]School of Social and Environmental Development,
National Institute of Development
Administration, 118 M.3, Serithai Road,
Klong-Chan, Bangkapi, Bangkok, 10240, Thailand

Abstract: Problem statement: Tone intelligibility in speech synthesis is an important attribute that should be taken into account. The tone correctness of the synthetic speech is degraded considerably in the average-voice-based HMM-based Thai speech synthesis. The tying mechanism in the decision tree based context clustering without appropriate criterion causes unexpected tone neutralization. Incorporation of the phrase intonation to the context clustering process in the training stage was proposed early. However, the tone correctness is not satisfied. **Approach:** This study proposes a number of tonal features including tone-geometrical features and phrase intonation features to be exploited in the context clustering process of HMM training stage. **Results:** In the experiments, subjective evaluations of both average voice and adapted voice in terms of the intelligibility of tone are conducted. Effects on decision trees of the extracted features are also evaluated. By considering gender in training speech, two core experiments were conducted. The first experiment shows that the proposed tonal features can improve the tone intelligibility for female speech model above that of male speech model, while the second experiment shows that the proposed tonal features improve the tone intelligibility for gender dependent model than for gender independent model. **Conclusion:** All of the experimental results confirm that the tone correctness of the synthesized speech from the average-voice-based HMM-based Thai speech synthesis is significantly improved when using most of the extracted features.

Key words: Thai speech, speech synthesis, tone intelligibility, tone correctness, generative model, context clustering, average voice, hidden Markov models

INTRODUCTION

In HMM-based speech synthesis, the prominent attribute is the ability to generate speech with arbitrary speaker's voice characteristics and various speaking styles. There have been proposed a number of TTS techniques and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech. However, to provide various voice characteristics in speech synthesis systems based on the speech unit selection approach, a large amount of speech data is needed and it is very tough to obtain enough speech data (Yamagishi *et al*., 2002).

For tonal languages such as Thai, tone is a very important suprasegmental feature of syllables. The words with the same phoneme sequence may have different meanings if they have different tones (Seresangtakul and Takara, 2003). Therefore, tone must be carefully considered in tonal speech synthesis.

The most important characteristics of a speech synthesis system are naturalness and intelligibility. Tone distortion can deteriorate not only the speech intelligibility but also the speech naturalness as well, since the tone is a suprasegmental feature formed by the basic prosodic feature, i.e., F0 (Wutiwiwatchai and Furui, 2007). Meanwhile the other important basic prosodic features including phrasal pauses, duration and energy can affect the speech naturalness mostly (Chomphan, 2009). As a result, the tone correctness must be carefully considered in the tonal speech synthesis.

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

As for speaker dependent HMM-based Thai speech synthesis research, a speech synthesizer has been implemented (Chomphan and Kobayashi, 2007a). In the system, a group of contextual factors which affect spectrum, fundamental frequency (F0) and state duration, such as tone type and part of speech are taken into account especially for the purpose of producing natural sounding prosody of the tonal speech. It has been found that it can provide speech with the better reproduction of prosody over the unit-selection-based Vaja TTS system from NECTEC (National Electronics and Computers Technology Center) (Chomphan and Kobayashi, 2007b). A decision tree with a tone-separated structure (Chomphan and Kobayashi, 2007a) presents the considerable improvement of tone correctness of the synthesized speech. However, some distortion of syllable duration is noticeable when the system is trained with a small amount of data. The other structures of the decision tree are designed for not only the purpose of maximal correctness of tone but also the purpose of elimination of the syllable duration distortion (Chomphan and Kobayashi, 2007a).

In the area of speaker independent HMM-based Thai speech synthesis, a system with a speech database containing quite a large number of speakers with a small amount of data for each speaker has been developed. Although it is desirable that sentence sets of speakers are different from each other to make database rich in phonetic and linguistic contexts, the synthetic speech generated from the average voice model (Yamagishi *et al*., 2002) trained using different sentence set for each speaker sounds unnatural compared to the model trained using the same sentence set for all speakers, especially when the amount of training data of each speaker is limited. To overcome the problem, the Shared decision Tree Context clustering (STC) (Yamagishi *et al*., 2003) is adopted, where every node of the decision tree always has training data from all speakers so that each distribution of the average voice model reflects the statistics of all speakers. Moreover, Speaker Adaptive Training (SAT) (Chomphan and Kobayashi, 2008) is incorporated into the training procedure of the average voice model to improve the quality of the average voice model.

The naturalness of the synthetic speech generated from the mentioned system is comparable to that of the speaker dependent system, however, the tone correctness of the synthetic speech is degraded considerably. The tying mechanism in the decision tree based context clustering without appropriate criterion causes unexpected tone neutralization. This problem does not always occur in the speaker dependent system, but it is obviously seen in the speaker independent system when multi-speaker speech database is used for training. Subsequently, an incorporation of the phrase intonation to the context clustering process in the training stage was proposed (Chomphan and Kobayashi, 2009). The phrase intonation features are extracted by using a Fujisaki's model including a baseline value of F0 and a magnitude of phrase command. However, the tone correctness is not

satisfied. This study, therefore, proposes a number of tonal features including tone-geometrical features and phrase intonation features to be used in the context clustering process of HMM training stage. The experimental results of subjective evaluations of the proposed technique are also discussed in the study.

MATERIALS AND METHODS

Phrase intonation features: The F0 contours are varied considerably in continuous speech due to the influences of such factors as tones of the adjacent syllables, syntactic and pragmatic information of the whole utterance and the overall speaking rate. The main factors causing these variations are tone coarticulation, tone enhancement/suppression and phrase intonation (Fujisaki and Ohno, 1998). In the conventional HMM-based speech synthesis approach, the speech features including F0 values are modeled statistically. As for speaker dependent system, the intelligibility of tone is degraded significantly when using simple binary tree-based context clustering. It can be treated by applying the tree structure of tone-separated structure. Moreover, in the case of speaker independent system, the variety of the speaker characteristics and the small amount of training data from one speaker cause the tone neutralization. As a result, the intelligibility of tone is considerably degraded in spite of modifying the tree structure. To relieve the effect of the variety of such F0 contours, the phrase intonation is thought to be a promising factor. The relevant features are the baseline value of F0 and the magnitude of phrase command of the Fujisaki's model as shown in Fig. 1. These features are subsequently incorporated into the contextual factors to reduce the variations caused by the phrase intonation factor.

Phrase intonation feature extraction: Fujisaki's model treats the F0 contour as a smooth rise-fall pattern in the vicinity of the accented Japanese mora (Fujisaki *et al*., 1990). The F0 contour is mathematically treated as a linear superposition of a global phrase and local accent components on a logarithmic scale. The phrase command produces a baseline component, while the accent command produces the accent component of an F0 contour. We use the two parameters of Fujisaki's model as our phrase intonation features including the baseline value of F0 and the magnitude of phrase command. An F0 contour of an utterance generated from an extension of Fujisaki's model for tonal languages has the following expressions Eq. 1-3:

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

$$G_{pi}(t) = \begin{cases} (\alpha_i^2 t)\exp(-\alpha_i t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (2)$$

$$G_{t,jk}(t) = \begin{cases} [1 - (1 + \beta_{jk} t)\exp(-\beta_{jk} t)] & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (3)$$

where, G_{pi} (t) represents the impulse response function of the phrase control mechanism and $G_{t,jk}(t)$ represents the step response function of the tone control mechanism, respectively. The symbols are denoted as follows: F_b is the smallest F0 value in the F0 contour of interest, A_{pi} and $A_{t,jk}$ are the amplitudes of the i-th phrases and of the j-th tone command. T_{0i} is the timing of the i-th phrase command; T_{1jk} and T_{2jk} are the onset and offset of the k-th component of the j-th tone command. α_i and β_{jk} are time constant parameters. I, J, K(j) are the number of phrases, tones and components of the j-th tone contained in the utterance, respectively. To extract the best representative parameters from the model, the optimization is conducted by minimizing the mean squared error in the ln F0(t) domain through the hill-climbing search in the space of model parameters.

Tone-geometrical features: The tone-geometrical features which represent the F0 contour in minor level, i.e., syllable level are applied in this context. From the F0 contour of each training utterance, a portion of the contour is extracted syllable by syllable. Figure 2 shows an example of the syllable portion where the dominant points are marked and the tone-geometrical features are extracted by measuring the distance from these points.

Tone-geometrical feature extraction: The tone-geometrical features are extracted as follows. The F0 contour of a whole utterance is marked for the syllable boundary by using the phoneme labeling information.

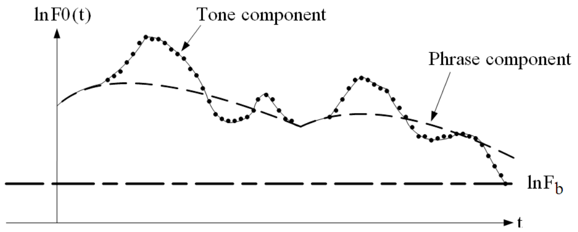

Fig. 1: Representation of F0 contour by generative model

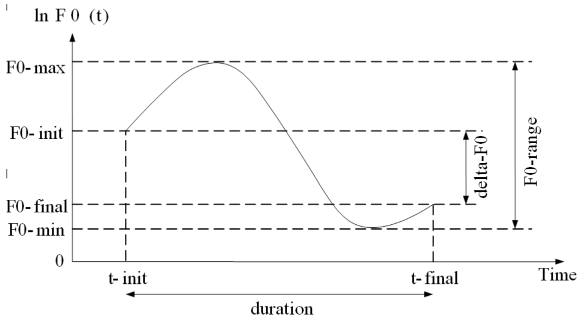

Fig. 2: Tone-geometrical feature representation for F0 contour for a syllable

Subsequently, all points in Fig. 2 are fitted and then the tone-geometrical features are extracted by applying the following definitions. These features are chosen and defined as follows:

- Initial F0 of syllable denoted by F0_init in Fig. 2
- Syllable duration denoted by duration in Fig. 2
- Syllable slope of contour calculated by delta_F0/duration
- Amplitude of accent command ($A_{t,jk}$ in Eq. 1)

Feature arrangement for context clustering: As for the phrase intonation features, the first feature of baseline value of F0 (F_b), it ranged from 67.7-178.8 Hz, while the second feature of amplitude of phrase command (A_{pi}) ranged from 0.00-1.20. Both of them were linearly quantized into 8 classes with an assigned codeword of 0-7. These features were then grouped into two sets (S15, S16) in the phrase level of the following contextual factors. It has been noted that the idea is to indicate the level of phrase intonation for the current phone, therefore both features have to be performed together. As a result, the feature of baseline value of F0 is not classified into the utterance level, although each utterance has its unique value. As for the tone-geometrical features, the feature of initial F0 of syllable, it ranged from 52.4 Hz-289.3 Hz, the second feature of syllable duration ranged from 5 ms to 1235 ms, the third feature of syllable slope, it ranged from -15.49-7.56, while the last feature of amplitude of accent command ranged from 0.03-2.99. All features were linearly quantized in the same manner as the phrase intonation features. These features were then grouped into four sets (S6, S7, S8, S9) in the syllable level of contextual factors.

Phoneme level:

- {preceding, current, succeeding} phonetic type
- {preceding, current, succeeding} part of syllable structure

Syllable level:

- S3. {preceding, current, succeeding} tone type
- S4. the number of phones in {preceding, current, succeeding} syllable
- S5. current phone position in current syllable
- S6. codeword of initial F0 of syllable
- S7. codeword of syllable duration
- S8. codeword of syllable slope
- S9. codeword of amplitude of accent command

Word level:

- S10. current syllable position in current word
- S11.part of speech
- S12. the number of syllables in {preceding, current, succeeding} word

Phrase level:

- S13. current word position in current phrase
- S14. the 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. the number of syllables in current sentence
- S19. the number of words in current sentence

In the synthesis stage, the parameter generation algorithm is mostly the same as the conventional system with adding the appropriate codewords for these proposed features in the context labels. The mean values of the proposed features for a specific target, which are the a priori knowledge, are expected to be the best representatives. For examples, to generate an average voice, the mean values from all speakers in the training data are selected for the phrase intonation features; 127.4 Hz and 0.40 for F_b and A_{p1}, corresponding to the quantization codewords of 4 and 2, respectively. To generate an adapted voice of a female target speaker, the mean values of the female; 154.6 Hz and 0.38 for F_b and A_{p1}, are chosen, corresponding to the quantization codewords of 5 and 2, respectively. In the sentence with more than one phrase, the i-th successive phrase also needs an associated A_{pi}. The representatives for these amplitudes of phrase command can be obtained by the same statistical method. In addition to the phrase intonation features, the tone-geometrical features are treated differently. Since these features are quite different for each tone type, the mean values for each tone type are chosen instead of the means values of all tone types. For example of an target female, the mean values of the initial F0 of syllable, the syllable duration, the syllable slope and the amplitude of accent command, are chosen respectively as; 245.0 Hz, 188 ms, -1.05, 0.35 for Tone0; 243.5 Hz, 129 ms, -2.25, 0.38 for Tone1; 252.0 Hz, 170 ms, 0.70, 0.43 for Tone2; 232.6 Hz, 145 ms, 0.35, 0.36 for Tone3 and 238.4 Hz, 197 ms, -1.10, 0.36 for Tone4. Finally, these selected means are transformed into the corresponding quantization codewords as same as those for the phrase intonation features. It has been noted that the representatives of these tonal features should be adapted to the target speaker to comply with individual statistics.

RESULTS

Experimental conditions: The speech material consists of a set of phonetically balanced sentences of Thai speech database named LOTUS and a set of phonetically balanced sentences of Thai speech database named TSynC-1 (Hansakunbuntheung *et al*., 2005). They are both from NECTEC. The whole sentence text of both databases was collected from Thai part-of-speech tagged ORCHID corpus. In LOTUS speech database, the speech was collected from 24 female and 24 male speakers with clear articulation and standard Thai accent, while the speech in the TSynC-1 was collected from a professional female speaker. The speech signal were sampled at a rate of 16kHz and windowed by a 25 ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were extracted by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zero-th coefficient, logarithm of F0 and their delta and delta-delta coefficients (Masuko *et al*., 1996).

The 5-state left-to-right HSMMs in which the spectral part was modeled by a single diagonal Gaussian output distribution are conducted. Each context dependent HSMM corresponds to a phoneme-sized speech unit. The average voice model was trained using 35 sentences for each speaker from 24 female and 24 male speaker's speech data.

To evaluate the proposed approach, 4 different models were constructed; male, female, gender-independent and gender-dependent models. Each model was trained by the baseline and proposed approaches. The baseline training embedded the shared decision tree context clustering (STC) technique with a tone-separated tree structure and the Speaker Adaptive Training (SAT). The applied approach incorporated the tonal features into the context clustering of the baseline training.

A couple of comparisons of tone intelligibility include that for male and female speech models and that for gender-dependent and gender-independent speech models. Moreover the reference system of a speaker dependent system using 1, 500 training utterances in both comparisons is presented. In each comparison, an average voice is conducted first and then an adapted voice is applied. In the speech adaptation, the MLLR-based speaker adaptation (Yamagishi *et al*., 2004) with 35 utterances of a target speaker was performed.

The notations used in the experimental results are defined as follows. The entries for "sd.", "male avg.", "female avg.", "male adt.", "female adt.", "gi. avg.", "gd. avg.", "gi. adt." and "gd. adt." correspond to speaker dependent model, male average voice model, female average voice model, male adapted voice model, female adapted voice model, gender independent average voice model, gender dependent average voice model, gender independent adapted voice model, gender dependent adapted voice model, respectively, meanwhile the entries for "bl", "bl+pi" and "bl+pi+tg" correspond to baseline training, baseline training with phrase intonation features, baseline training with phrase intonation features and tone-geometrical features, respectively.

Influence of tonal features on clustering trees: We first investigated how the phrase intonation features and tone-geometrical features affect the clustering trees including mel-cepstrum (mcep), logF0 and duration (dur) trees. The increasing of percentage of the number of existing questions and the dominance score for the questions in phrase level and syllable level is summarized in Fig. 3 and 4, respectively. Figure 5 shows the effects of phrase intonation features and tone-geometrical features which cause the changes in the F0 contour levels of some syllables.

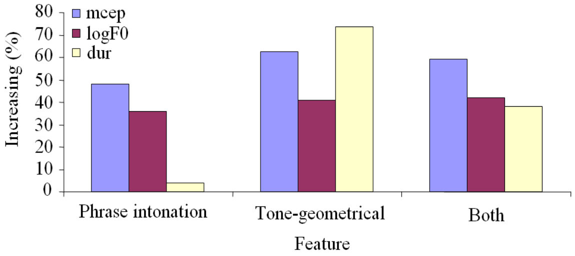

Fig. 3: Increasing percentages of the number of existing questions for the questions in phrase level (for the phrase intonation features) and syllable level (for the tone-geometrical features) from different clustering trees

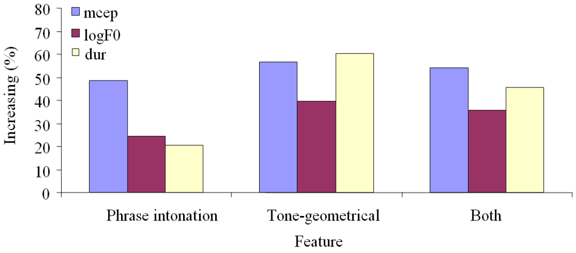

Fig. 4: Increasing percentages of the dominance score for the questions in phrase level (for the phrase intonation features) and syllable level (for the tone-geometrical features) from different clustering trees

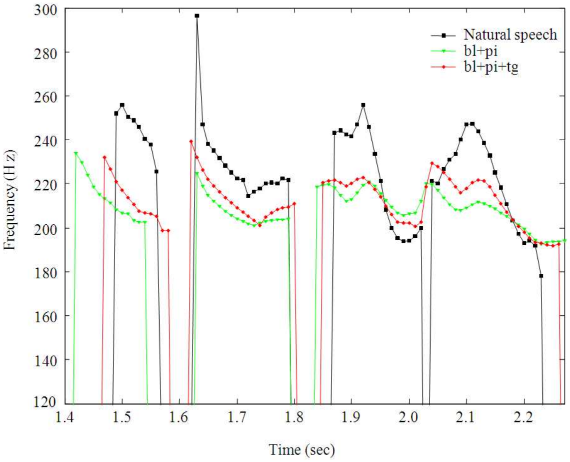

Fig. 5: Examples of F0 contour of natural speech, generated F0 contours from baseline system with phrase intonation features and baseline system with phrase intonation features and tone-geometrical features

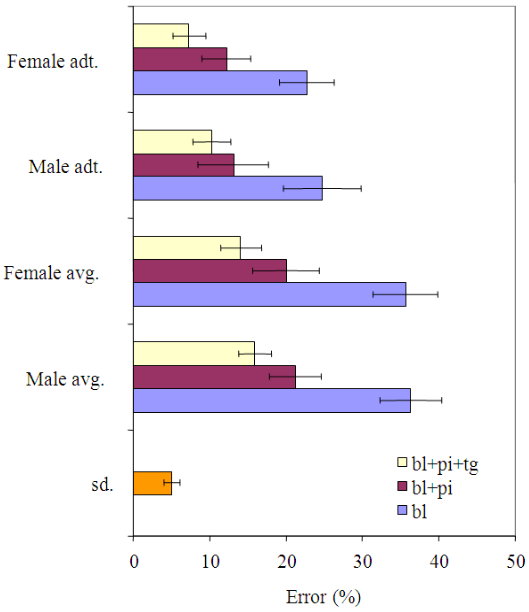

Fig. 6: Tone error percentage of average voice and adapted voice for male and female speech models synthesized from different training approaches.

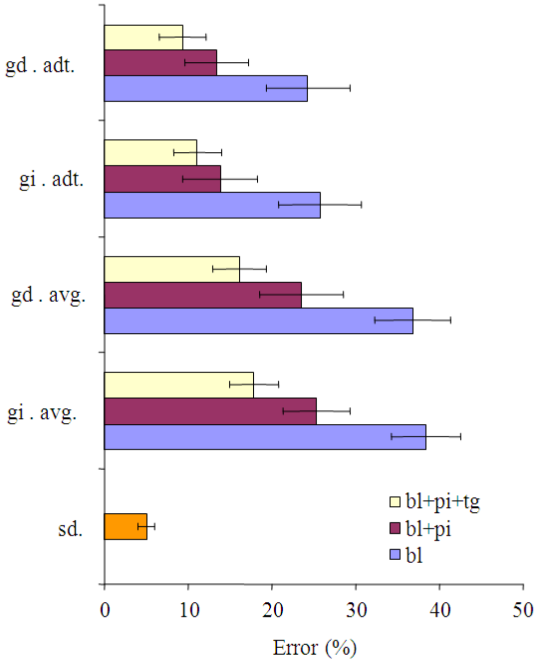

Fig. 7: Tone error percentage of average voice and adapted voice for gender-dependent and gender-independent speech models synthesized from different training approaches

Tone intelligibility for male and female speech models: This result shows how the overall tone correctness of the average voice and adapted voice is improved by embedding the proposed features for the different speech models of male and female. The tone error percentage is the measured value in this comparison. To calculate tone error percentage of the implemented systems, a subjective test was performed. The 2,289 syllables of 100 synthesized speech utterances were presented to eight subjects. Then they were requested to decide whether the syllables have the same tones as the given texts. The average tone error percentages with 95% confidence interval for different training styles are summarized in Fig. 6.

Tone intelligibility for gender-dependent and gender-independent speech models: The result presents the comparison of tone intelligibility for gender-dependent and gender-independent speech models in the overall tone correctness. The average tone error percentages with 95% confidence interval for different training styles are summarized in Fig. 7.

DISCUSSION

From the experimental result of influence of tonal features on clustering trees, it can be noticed from Fig. 3 and 4 that the increasing of percentage by the tone-geometrical features is noticeably larger than that of the phrase intonation features. From Fig. 5, it can be obviously seen that proposed tonal features can reduce the difference between those of the natural speech and the baseline training system with only phrase intonation features.

From the experimental result of tone intelligibility for male and female speech models, it can be noticed from Fig. 6 that the proposed tonal features can reduce the tone error percentage from the baseline training approach more than the phrase intonation features. Moreover, the reduction in tone error percentage of the female speech model is little larger than that of the male speech model.

From the experimental result of tone intelligibility for gender-dependent and gender-independent speech models, both average voice and adapted voice give the result corresponding to the earlier results as seen in Fig. 7. Moreover, the reduction in tone error percentage of the gender-dependent speech model is little greater than that of the gender-independent speech model.

CONCLUSION

A group of tonal features including phrase intonation features and tone-geometrical features are used to be embedded in the contextual factors for the context clustering process of a speaker independent HMM-based Thai speech synthesis system. They are extracted based on the parameter optimization of generative model and the geometrical parameter extraction. It is supposed to reduce the variation of tone caused by phrase intonation both from intra-and inter-speaker. From the experimental results, the proposed tonal features can improve the tone intelligibility for female speech model above that of male speech model. In addition, the proposed tonal features give the better improvement of the tone intelligibility for gender dependent model than for gender independent model. These results confirm that the tone correctness of the synthesized speech is significantly improved when using most of the extracted speech features.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007a. Design of tree-based context clustering for an HMM-based thai speech synthesis system. Proceedings of the 6th ISCA Workshop on Speech Synthesis, Aug. 22-24, ISCA Archive, Bonn, Germany, pp: 160-165.

Chomphan, S. and T. Kobayashi, 2007b. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceedings of the 8th Annual Conference of the International Speech Communication Association, Aug. 27-31, ISCA Archive, Antwerp, Belgium, pp: 2849-2852.

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

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

Chomphan, S., 2009. Towards the development of speaker-dependent and speaker-independent hidden markov model-based Thai speech synthesis. J. Comput. Sci., 5: 905-914. DOI: 10.3844/jcssp.2009.905.914

Fujisaki, H. and S. Ohno, 1998. The use of a generative model of F_0 contours for multilingual speech synthesis. Proceedings of the 4th International Conference on Spoken Language Processing, Oct. 12-16, IEEE Xplore Press, Beijing, China, pp: 714-717. DOI: 10.1109/ICOSP.1998.770311

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. Proceedings of the 1st International Conference on Spoken Language Processing, Nov. 18-22, ISCA Archive, Kobe, Japan, pp: 841-844.

Hansakunbuntheung, C., A. Rugchatjaroen and C. Wutiwiwatchai, 2005. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. National Electronics and Computer Technology Center. http://hlt.nectec.or.th

Masuko, T., K. Tokuda, T. Kobayashi and S. Imai, 1996. Speech synthesis using HMMs with dynamic features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, IEEE Xplore Press, Atlanta, USA., pp: 389-392. DOI: 10.1109/ICASSP.1996.541114

Seresangtakul, P. and T. Takara, 2003. A generative model of fundamental frequency contours for polysyllabic words of Thai tones. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, IEEE Xplore Press, Hong Kong, pp: 452-455. DOI: 10.1109/ICASSP.2003.1198815

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

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. Proceedings of the 7th International Conference on Spoken Language Processing, Sep. 16-20, ISCA Archive, Denver, Colorado, USA., pp: 133-136.

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. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, IEEE Xplore Press, Hong Kong, pp: 716-719. DOI: 10.1109/ICASSP.2003.1198881

Yamagishi, J., T. Masuko and T. Kobayashi, 2004. MLLR adaptation for hidden semi-Markov model based speech synthesis. Proceedings of the 8th International Conference on Spoken Language Processing, Oct. 4-8, ISCA Archive, Jeju Island, Korea, pp: 1213-1216.