American Journal of Applied Sciences 9 (3): 313-320, 2012
ISSN 1546-9239

Design of Tree Structure in Context Clustering Process of 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: In HMM-based Thai speech synthesis, the tone degradation due to the imbalance of training data of all tones. Some distortion of syllable duration is obviously noticeable when the system is trained with a small amount of data. These problems cause the decrement in naturalness and intelligibility of the synthesized speech. **Approach:** This study proposes an approach to improve the correctness of tone of the synthesized speech which is generated by an HMM-based Thai speech synthesis system. In the tree-based context clustering process, tone groups and tone types are used to design four different structures of decision tree including a single binary tree structure, a simple tone-separated tree structure, a constancy-based-tone-separated tree structure and a trend-based-tone-separated tree structure. **Results:** A subjective evaluation of tone correctness is conducted by using tone perception of eight Thai listeners. The simple tone-separated tree structure gives the highest level of tone correctness, while the single binary tree structure gives the lowest level of tone correctness. The additional contextual tone information which is applied to all structures of the decision tree achieves a significant improvement of tone correctness. Finally, the evaluation of syllable duration distortion among the four structures shows that the constancy-based-tone-separated and the trend-based-tone-separated tree structures can alleviate the distortions that appear when using the simple tone-separated tree structure. **Conclusion:** The appropriate structure of tree in context clustering process with the additional contextual tone information can improve the correctness of tones, while the constancy-based-tone-separated and the trend-based-tone-separated tree structures can alleviate the syllable duration distortions.

Key words: Thai speech, speech synthesis, tree-based context clustering, HMM-based speech synthesis, tone correctness, syllable duration distortion, important suprasegmental

INTRODUCTION

Tone is a very important suprasegmental feature of syllable for tonal languages such as Thai, Mandarin and Vietnamese. The words with the same phoneme sequence may have different meanings if they have different tones (Seresangtakul and Takara, 2003). Thus, tone must be carefully taken into account in speech synthesis systems of tonal languages. In the present day, HMM-based speech synthesis system is becoming popular. It has been developed for Japanese by for years (Tokuda *et al*., 1999; Masuko *et al*., 1996; Yoshimura *et al*., 1999) and has also been developed for many other languages such as Korean, English, Portuguese, Chinese and German (Chomphan, 2009). The HMM-based speech synthesis has been applied to many speech synthesis systems of tonal languages successfully.

In Thai, the HMM-based speech synthesis system has been developed for years (Chomphan and Kobayashi, 2007). In the system, a group of contextual factors which affect spectrum, pitch 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 language. It has been found that it can provide speech with the better reproduction of prosody over the unit-selection-based Vaja TTS system (Hansakunbuntheung *et al*., 2005). Specifically, a decision tree with a tone-separated structure shows the significant improvement of tone correctness of the synthesized speech. However, some distortion of syllable duration is obviously noticeable

when the system is trained with a small amount of data. To treat this problem, this study proposes some other structures of the decision tree designed for the purpose of maximal correctness of tone and the purpose of elimination of the syllable duration distortion. Moreover, the contextual tone information (tone types of the preceding and the succeeding syllables) has also been applied to the designed decision-tree structures.

MATERIALS AND METHODS

Characteristics of Thai tones: From the study of Thai sound system by Lukseneeyanawin (Wutiwiwatchai and Furui, 2007; Chomphan and Kobayashi, 2008), Thai sound is described in a syllable unit as shown in Fig. 1. The basic Thai textual syllable structure is composed of consonants, vowels and tone, where Ci, V, Cf and T denotes an initial consonant, a vowel, a final consonant and a tone, respectively.

In tonal languages, tone, which is indicated by contrasting variations in contour of fundamental frequency (F_0) 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. In Thai language, there are five tonal variations traditionally named according to the characteristics of their F_0 contours within a syllable as shown in Fig. 2 (Chomphan 2010a, 2010b). The effect of tone on the linguistic meaning is presented in the following examples: the syllable /คา/ has tone 0 and means “to get stuck”, the syllable /ข่า/ has tone 1 and means “galangal, a kind of spice”, the syllable /ฆ่า/ has tone 2 and means “to kill”, the syllable /ค้า/ has tone 3 and means “to trade” and the syllable /ขา/ has tone 4 and means “leg”. In the investigation of tone occurrence statistics in TSynC speech database (In Thai), It has been found that 77,413 syllables are occupied by tone 0 (38), tone 1 (22%), tone 2 (17%), tone 3 (15%) and finally tone 4 (8%), respectively. It can be seen that the training data of tone 0 dominates those of the others.

Categorizations of Thai tones: Two criteria of categorization of Thai tones into tone groups are as follows.

$$\mathrm{Ci(Ci)\ V\overset{T}{(V)}\ Cf}$$

Fig. 1: Thai tonal syllable structure

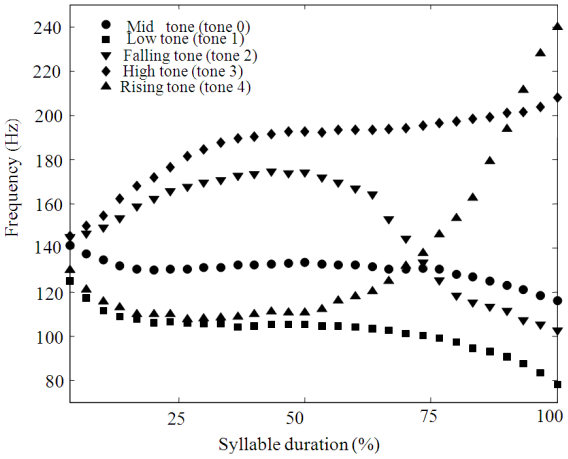

Fig. 2: Standard F_0 contours for thai tones

First, by considering the constancy of the F_0 contour, Abramson divided the tones into two groups (Chomphan 2010c): the static group consists of three tones, high tone, middle tone and low tone; the dynamic group consists of two tones, rising tone and falling tone. Second, by considering each contour of Fig. 2, it can be seen that the pitch patterns of the mid, low, falling, high and rising tones are relatively mid-fall, fall, rise-fall, rise and fall-rise, respectively. Therefore, they can be divided according to the final trend of their contours: the upward trend group consists of two tones, high tone and rising tone; the downward trend group consists of three tones, mid tone, low tone and falling tone.

Construction of contextual factors in Tree-based context clustering: In the HMM-based speech synthesis system, context clustering is an important process to treat the problem of limitation of training data. Information sharing of training data in the same cluster or leaf node in the decision-tree-based context clustering is the important concept, therefore the construction of contextual factors and design of tree structure for the decision-tree-based context clustering should be conducted appropriately.

A number of language-dependent contextual factors has been implemented for Thai (Chomphan and Kobayashi, 2007) to model context dependent HMMs. The following 13 contextual factor sets in 5 levels of speech unit have been constructed according to 2 sources of information, including the phonological information (Chomphan and Kobayashi, 2008) (for phoneme and syllable levels) and the utterance structure from Thai text corpus named ORCHID (Chomphan 2010c; 2010d) (for word, phrase and utterance levels):

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. the number of phones in {preceding, current, succeeding} syllable
- S5. current phone position in current syllable

Word level:

- S6. current syllable position in current word
- S7. part of speech
- S8. the number of syllables in {preceding, current, succeeding} word

Phrase level:

- S9. current word position in current phrase
- S10. the number of syllables in {preceding, current, succeeding} phrase

Utterance level:

- S11. current phrase position in current sentence
- S12. the number of syllables in current sentence
- S13. the number of words in current sentence

These contextual information sets were thereafter transformed into question sets which were applied at the context clustering process in the training stage with the total question number of 1156. An analysis of these question sets was conducted in (Chomphan and Kobayashi, 2007) to evaluate the contribution of each set. Figure 3 shows an example of decision trees for spectrum, pitch and state duration by using all of the constructed question sets for the single binary tree context clustering. It can be obviously seen that the root node question in each tree (C_sil from the spectrum tree, C_Silence from the pitch tree and C_Long from the state duration tree) is of the phonetic type question set. It corresponds to the previous analysis that phonetic type question set is the most important set among all thirteen sets.

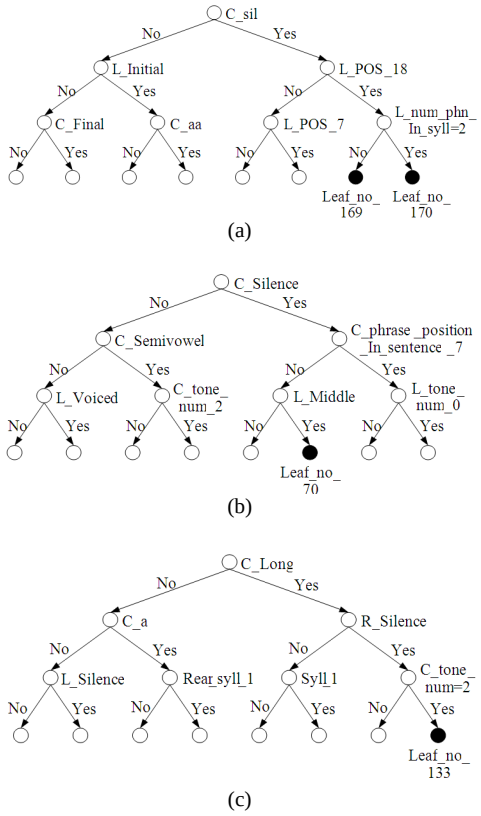

Fig. 3: Example of decision trees for: (a) spectrum (3rd state), (b) pitch (2nd state) and (c) state duration

Design of decision-tree structures: Basically, the single binary tree structure is used in the decision tree-based context clustering process as shown in Fig. 4a. The imbalance of tone frequency causes the prevalence of some tones to the others, as a result, the single binary tree context clustering gives high tone error percentage in the synthesized speech (Chomphan and Kobayashi, 2007). To increase the tone correctness, the simple tone-separated decision-tree structure was designed as depicted in Fig. 4b (Chomphan and Kobayashi, 2007). It has been seen that the significant distortion of the generated syllable duration are unavoidable when using the simple tone-separated tree context clustering with small training data due to the limited data of each tone. To alleviate this problem, the other two structures were designed by considering tone groups and tone types, respectively.

Tone groups categorized in terms of constancy of the F_0 contour were used to design the structure of constancy-based-tone-separated tree as depicted in Fig. 4c. Meanwhile, tone groups categorized by the final trend were used to design the structure of trend-based-tone-separated tree as depicted in Fig. 4d.

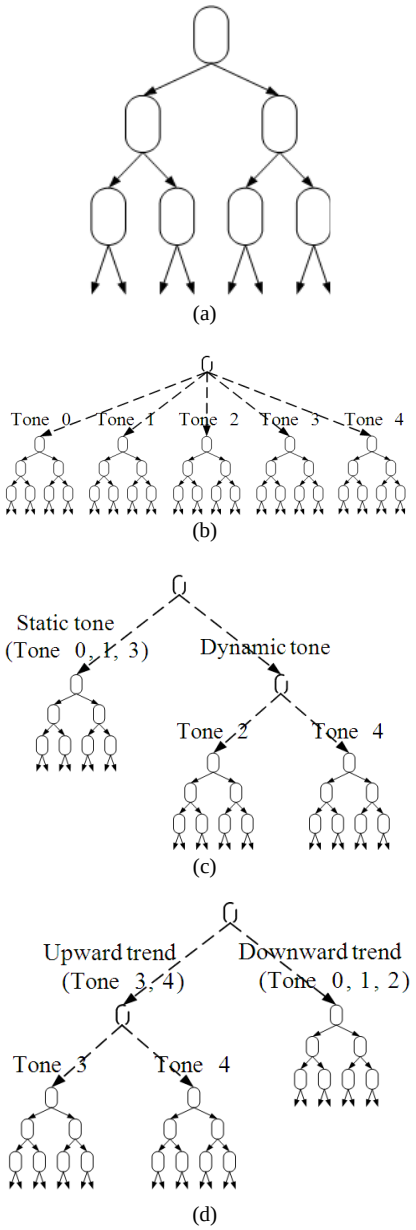

Fig. 4: Tree structures for context clustering: (a) single binary tree structure, (b) simple tone-separated tree structure, (c) constancy-based-tone-separated tree structure and (d) trend-based-tone-separated tree structure

In the static tone group of the constancy-based-tone-separated tree and the downward trend group of the trend-based-tone-separated tree, no tone-separations are applied because the data sharing among the tones within those groups is expected to treat the problem of syllable duration distortion.

Design of context clustering styles: At the beginning, the four structures of decision tree are employed directly in the context clustering process of the training stage without using the tone type question set. The first four styles of context clustering are listed as follows:

- Single binary tree context clustering without tone type questions
- Simple tone-separated tree context clustering without tone type questions
- Constancy-based-tone-separated tree context clustering without tone type questions
- Trend-based-tone-separated tree context clustering without tone type questions

It has been noted that only tone information of the current syllable is concerned in the tone-separated tree structures, while no tone information is concerned in the single binary tree structure. Moreover no other tone information in the neighboring syllables is considered in all structures. To exploit the ignored tone information, the tone type question set is incorporated into all of the designed tree structures to form another four styles of context clustering. These styles of context clustering process are listed as follows:

- Single binary tree context clustering with tone type questions
- Simple tone-separated tree context clustering with tone type questions
- Constancy-based-tone-separated tree context clustering with tone type questions
- Trend-based-tone-separated tree context clustering with tone type questions

RESULTS

Speech database and training conditions: A set of phonetically balanced sentences of Thai speech database named TSynC from National Electronics and Computers Technology Center was used for training the HMMs (Hansakunbuntheung *et al*., 2005). The used sentence text was collected from Thai part-of-speech tagged ORCHID corpus. 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 and the utterance structure from ORCHID were used to construct the context dependent labels with 79 different phonemes including 65 phonemes from original Thai words, 12 phonemes from some loan words and 2 phonemes of silence and pause.

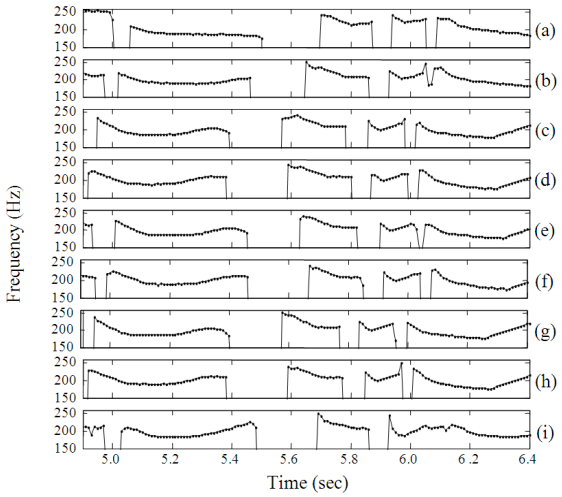

Fig. 5: F_0 contours of synthesized speech from 8 different clustering styles; (a) single binary tree without tone type questions, (b) simple tone-separated tree without tone type questions, (c) constancy-based-tone-separated tree without tone type questions, (d) trend-based-tone-separated tree without tone type questions (e) single binary tree with tone type questions, (f) simple tone-separated tree with tone type questions, (g) constancy-based-tone-separated tree with tone type questions, (h) trend-based-tone-separated tree with tone type questions and (i) F_0 contour of natural speech

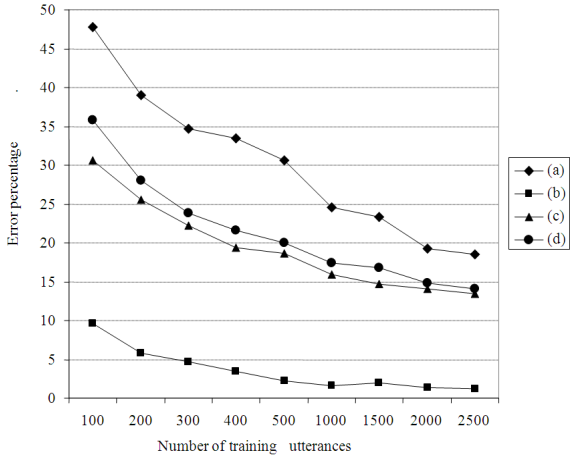

Fig. 6: Tone error percentages of synthesized speech from 4 different clustering styles; (a) single binary tree without tone type questions, (b) simple tone-separated tree without tone type questions, (c) constancy-based-tone-separated tree without tone type questions and (d) trend-based-tone-separated tree without tone type questions

The speech signals were sampled at 16 kHz and windowed by a 25 m sec 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, logarithm of F_0 and their delta and delta-delta coefficients (Tokuda *et al*., 1999).

The 5-state left-to-right phoneme-sized HSMMs is used. The number of training utterances was varied as follows: 100, 200, 300, 400, 500, 1000, 1500, 2000 and 2500, respectively.

Evaluation of overall tone correctness: It has been done to present how the overall tone correctness of the synthesized speech is improved by using eight different tree-based context clustering styles. Figure 5 shows an example of F_0 contours of the natural speech and synthesized speech with different clustering styles. The first full-shape syllable of Fig. 5 conveys tone 4 or rising tone. To evaluate the overall tone correctness of our implemented system, a subjective test was conducted. The 2,289 syllables of 100 synthesized speech utterances were presented to eight native subjects. Then the subjects were requested to decide whether the syllables have the same tones as the given texts or not. The average tone error percentages for the first four styles and another four styles are summarized in Fig. 6 and 7, respectively.

Evaluation of tone correctness for each tone type: The tone correctness in terms of the tone types is presented. The result is shown in Fig. 8. For 100 synthesized speech utterances, the numbers of the syllables with tone 0, tone 1, tone 2, tone 3 and tone 4 are 750, 560, 449, 339 and 191, respectively.

Evaluation of syllable duration distortion: A paired-comparison test among all tree structures (the styles (e)-(h)) of the context clustering with tone type questions was performed. Ten test sentences selected randomly from 100 synthesized speech utterances were used in this evaluation. For each comparison, a pair of utterances from two of the four structures is presented to eight subjects and then the subjects are requested to choose the one which has more natural duration without considering the correctness of tone. The average scores in percentage of each tree structure with the different number of training utterances are shown in Fig. 9.

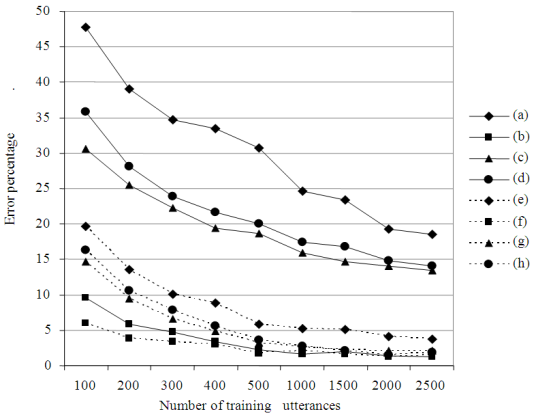

Fig. 7: Tone error percentages of synthesized speech from 8 different clustering styles; (a) single binary tree without tone type questions, (b) simple tone-separated tree without tone type questions, (c) constancy-based-tone-separated tree without tone type questions, (d) trend-based-tone-separated tree without tone type questions (e) single binary tree with tone type questions, (f) simple tone-separated tree with tone type questions, (g) constancy-based-tone-separated tree with tone type questions and (h) trend-based-tone-separated tree with tone type questions

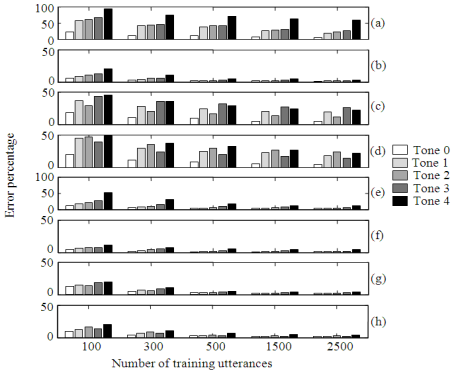

Fig. 8: Tone error percentages of synthesized speech from 8 different clustering styles categorized by tone types; (a) single binary tree without tone type questions, (b) simple tone-separated tree without tone type questions, (c) constancy-based-tone-separated tree without tone type questions, (d) trend-based-tone-separated tree without tone type questions (e) single binary tree with tone type questions, (f) simple tone-separated tree with tone type questions, (g) constancy-based-tone-separated tree with tone type questions and (h) trend-based-tone-separated tree with tone type questions

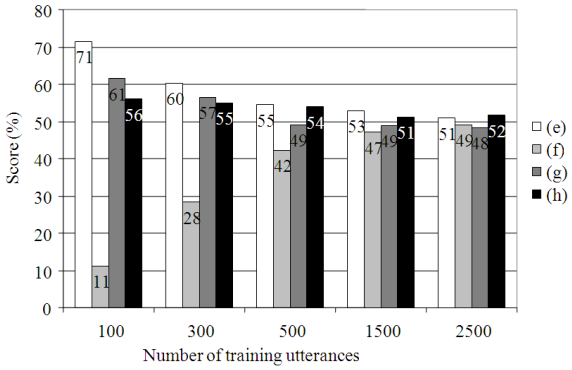

Fig. 9: Scores of a paired-comparison test for natural duration among 4 different clustering styles; (e) single binary tree with tone type questions, (f) simple tone-separated tree with tone type questions, (g) constancy-based-tone-separated tree with tone type questions and (h) trend-based-tone-separated tree with tone type questions

DISCUSSION

In the evaluation of overall tone correctness, Fig. 5a is of the single binary tree context clustering without tone type questions; however this syllable contour is misshaped. As a result, most listeners perceived it with wrong tone. Meanwhile Fig. 5b-h are of the other styles and they show the improvement of the F_0 contour shape conforming to that of the natural speech as depicted in Fig. 5i.

Figure 6 shows the tone error percentages of synthesized speech from the context clustering styles (a)-(d). Comparing the four designed tree structures, we can see that the style (a) (single binary tree without tone type questions) gives the highest level of error percentage, the style (c) (constancy-based-tone-separated tree without tone type questions) and the style (d) (trend-based-tone-separated tree without tone type questions) can reduce the error percentage significantly, while the style (b) (simple tone-separated tree without tone type questions) gives the lowest error percentage. In other words, the context clustering with the tone-separated tree structure has more effectiveness than the context clustering with the single binary tree structure. We can also see that the tone error percentage is decreased as the number of training utterances is increased.

Figure 7 shows the tone error percentages of synthesized speech from the context clustering styles (e)-(h) relative to the styles (a)-(d), respectively. The tone type question set was applied to those four styles. It can be seen that the contextual tone information in

syllable level causes a drastic reduction of the tone error percentage for all tree structures except the simple tone-separated tree context clustering. The reason is that the simple tone-separated tree structure exploits all tone type questions of the current syllable in the separation of tree structure, while the single tree structure does not exploit the tone information at all and the constancy-based-tone-separated, trend-based-tone-separated tree structures exploit partly of the tone information. Therefore the effect of the tone type question set which is employed afterward to the simple tone-separated tree structure is smallest among that of all other tree structures.

From Fig. 6 and 7, it can be also seen that the constancy-based-tone-separated tree structure is more effective in giving a little lower error percentage than the trend-based-tone-separated tree structure.

In the evaluation of tone correctness for each tone type, from Fig. 8a or the single binary tree without tone type questions and (b) or the simple tone-separated tree without tone type questions, the error percentage of tone 4 is mostly highest among all tones, on the other hand, the error percentage of tone 0 is mostly lowest. The reason is that the proportion of training data of tone 4 is smallest while the proportion of training data of tone 0 is largest according to the statistics of tone occurrence in the speech database. From Fig. 8c or the constancy-based-tone-separated tree without tone type questions, the error percentages of tone 2 and 4 are noticeably reduced as compared to (a) or the single binary tree without tone type questions. Meanwhile, from Fig. 8d or the trend-based-tone-separated tree without tone type questions, the error percentages of tone 3 and 4 are reduced as compared to (a) or the single binary tree without tone type questions.

As for Fig. 8e-h in which the tone type question set is employed, it can be seen that the tone error percentages of tone 0-4 are rather close to each other and also much less than those of Fig. 8a-d.

In the evaluation of syllable duration distortion, from Fig. 9, the single binary tree structure gives the least distortion among all tree structures. On the other hand, the simple tone-separated tree structure gives the worst distortion compared to the other structures, because there is no sharing of data between each tone for the simple tone-separated tree structure, meanwhile there is some data sharing for the single binary tree structure, the constancy-based-tone-separated tree structure and the trend-based-tone-separated tree structure as seen in Fig. 4. However it can be seen that the distinction between the scores disappears when the number of training utterances is increased above 1000.

CONCLUSION

Four structures of decision tree were designed according to tone groups and tone types to obtain higher correctness of tone of synthesized speech in HMM-based Thai speech synthesis. The experimental results show that the tone-separated tree structures can reduce the tone error percentage of the synthesized speech compared with the single binary tree structure significantly. By using the contextual tone information in the syllable level, it can improve the tone correctness for all structures of decision tree. There are some distortions of the syllable duration appearing in the case of using the simple tone-separated tree context clustering with a small amount of training data, however it can be treated by using the constancy-based-tone-separated or the trend-based-tone-separated tree context clustering.

ACKNOWLEDGEMENT

The reaches are grateful to Kasetsart University at Si Racha campus for the research scholarship through the board of research.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007. 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., 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

Chomphan, S., 2010a. Analytical study on fundamental frequency contours of Thai expressive speech using Fujisaki's model. J. Comput. Sci., 6: 36-42. DOI: 10.3844/jcssp.2010.36.42

Chomphan, S., 2010b. Multi-pulse based code excited linear predictive speech coder with fine granularity scalability for tonal language. J. Comput. Sci., 6: 1288-1292. DOI: 10.3844/jcssp.2010.1288.1292

Chomphan, S., 2010c. Fujisaki's model of fundamental frequency contours for thai dialects. J. Comput. Sci., 6: 1263-1271. DOI: 10.3844/jcssp.2010.1263.1271

Chomphan, S., 2010d. Performance evaluation of multi-pulse based code excited linear predictive speech coder with bitrate scalable tool over additive white Gaussian noise and Rayleigh fading channels. J. Comput. Sci., 6: 1438-1442. DOI: 10.3844/jcssp.2010.1438.1442

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

Tokuda, K., T. Masuko, N. Miyazaki and T. Kobayashi, 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 15-19, IEEE Xplore Press, Phoenix, USA., pp: 229-232. DOI: 10.1109/ICASSP.1999.758104

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

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of 6th European Conference on Speech Communication and Technology, Sep. 5-9, ISCA Archive, Budapest, Hungary, pp: 2347-2350.