Implementation and Evaluation of an HMM-based Thai Speech Synthesis System

Suphattharachai Chomphan, *Takao Kobayashi*

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
{suphattharachai,takao.kobayashi}@ip.titech.ac.jp

10.21437/Interspeech.2007-533

Abstract

This paper describes a novel approach to the realization of Thai speech synthesis. Spectrum, pitch, and phone duration are modeled simultaneously in a unified framework of HMM, and their parameter distributions are clustered independently by using a decision-tree based context clustering technique with different styles. A group of contextual factors which affect spectrum, pitch, and state duration, i.e., tone type, part of speech, are taken into account especially for a tonal language. The evaluation of the synthesized speech shows that tone correctness is significantly improved in some clustering styles, moreover the implemented system gives the better reproduction of prosody (or naturalness, in some sense) than the unit-selection-based system with the same speech database.

Index Terms: Thai speech, HMM, speech synthesis, tone

1. Introduction

HMM-based TTS system in which each speech synthesis unit is modeled by HMM was proposed in the past decade by Tokuda et al. [1]-[3]. It has also been developed for some other languages as indicated in [4]. A distinctive feature of the system is that the speech parameters used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm.

As for Thai speech synthesis research, the first paper describing the development of a Thai TTS engine was published in 1983 by Saravari and Imai [5], [6], where a speech unit concatenation algorithm was applied to Thai. This approach was implemented in the latest version of Vaja by Hansakunbuntheung et al. [7] in 2005 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, etc. To provide such various voice characteristics in speech synthesis systems based on the speech unit selection approach, a large amount of speech data is necessary. However, it is burdensome to obtain enough speech data. In order to treat this problem, we have developed an HMM-based Thai speech synthesis.

This paper explains how to implement the HMM-based Thai speech synthesis system with a new approach. Thai phonological information and utterance structure of the text corpus are analyzed and then applied in the contextual factor construction. Four different styles of tree-based context clustering are designed to improve tone correctness which is crucial in Thai speech. The implementation and evaluation details are concluded in the following sections.

2. Thai Speech Characteristics

2.1. Thai Phonological System

A comprehensive description of Thai sound system was published in 1992 by Lukseneeyanawin [8]. A brief review is presented in this section. Thai sound is often described in a syllable unit as depicted in Figure 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.

Table 1 illustrates all Thai consonants and vowels in the International Phonetic Alphabet (IPA) and also summarizes

T
Ci(Ci) V(V) Cf

Figure 1: *Thai tonal syllable structure.*

Table 1. *Thai phonemes in IPA and the summarized number of Thai phones(Ph) and characters(Ch).*

Type		List	Ph	Ch
Initial Consonant	Single	p,p^h,t,t^h,c,c^h,k,k^h,b,d,m,n,ŋ,f,s,h,r,l,w,j	32	44
	Cluster	pr,pl,tr,kr,kl,kw,p^hr,p^hl,t^hr,k^hr,k^hl,k^hw		
Vowel	Short	i,ɨ,u,e,ɜ,o,æ,a,ɔ,ia,ɨa,ua	24	16
	Long	iː,ɨː,uː,eː,ɜː,oː,æː,aː,ɔː,iːa,ɨːa,uːa		
Final Consonant		p,t,c,k,m,n,ŋ,w,j	9	37

Table 2. *Thai consonant system.*

			Places of Articulation				
			Labial	Alveolar	Palatal	Velar	Glottal
Manners of Articulation	Stops	Voiceless Unaspired	p	t	c	k	z
		Voiceless Aspired	p^h	t^h	c^h	k^h	
		Voiced	b	d			
	Non-stops	Nasal	m	n		ŋ	
		Fricative	f	s			h
		Trill		r			
		Lateral		l			
		Approximant	w		j		

Table 3. *Thai vowel system.*

		Vowel Advancement		
		Front	Central	Back
Vowel Height	High	i,iː	ɨ,ɨː	u,uː
	Mid	e,eː	ɜ,ɜː	o,oː
	Low	æ,æː	a,aː	ɔ,ɔː
Diphthongs		ia,iːa	ɨa,ɨːa	ua,uːa

the number of the Thai phones and characters according to each part of the syllable structure. The clustered initial consonant can be constructed by combining each of the phonemes /p, p^h, t, t^h, k, k^h/ with one of the phonemes /r, l, w/. In consonantal phonemes classification, the duration of marginal sound is employed to categorize each phoneme with the same manner of articulation into appropriate place of articulation as shown in Table 2. In vowel phonemes classification, the vowels can be grouped by places of articulation using vowel advancement and also manners of articulation using vowel height as shown in Table 3. Moreover vowel duration is computed from a period of fundamental frequency (F0) to classify into short or long vowel as shown in Table 1. Diphthongs are double vowels beginning with one of the phonemes /i, i:, ɨ, ɨ:, u, u:/ followed by the phoneme /a/ [6].

Recently, some loan words which do not conform to the rules of native Thai phonology, such as the initial consonants /br, dr/ and the final consonants /f, c^h/ have begun to appear.

2.2. Thai Syllable Tone

Tone is a suprasegmental feature which uniquely exists in a tonal language. Five IPA tone markers are generally used to indicate Thai tones; / ˉ/ for middle tone (tone 0), / ˋ/ for low tone (tone 1), / ˆ/ for falling tone (tone 2), / ˊ/ for high tone (tone 3), and / ˇ/ for rising tone (tone 4). Each tone type is named according to the characteristic of its F0 contour within a syllable [9]. All five tones can be divided into two groups: 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. In a tonal language the meaning of a syllable changes as the syllable tone changes [10]. For example, the syllable /bā:n/ (/บาน/ in Thai) has a middle tone and means “to widen”, meanwhile syllable /bâ:n/ (/บ้าน/ in Thai) has a falling tone but means “home”. By investigating tone occurrence statistics in TSynC-1 speech database (see section 4.1), we found that 77,413 syllables are occupied firstly by middle tone (38%), low tone (22%), falling tone (17%), high tone (15%), and finally rising tone (8%).

3. Tree-based Context Clustering

3.1. Language-dependent Contextual Factors

Contextual information is language dependent. Besides, a large number of contextual factors do not guarantee the synthesized speech with better quality. There should be efficient factors for a certain language to model context dependent HMMs. The 13 contextual factor sets in 5 levels of speech unit were constructed according to 2 sources of information, including the phonological information as described in section 2.1 and 2.2 (for phoneme and syllable levels), and the utterance structure from Thai text corpus named ORCHID [11] (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

Subsequently, these contextual information sets were transformed into question sets which finally applied at the context clustering process in the training stage.

3.2. Analysis of Contextual Factors

To analyze the contribution of each set of contextual factors constructed in section 3.1, we explored 3 decision trees generated in the clustering process at the training stage of the system including spectrum, pitch, and state duration trees. Two criteria were taken into account. First, the number of the existing questions in each set was counted. The 3 highest proportions among 13 sets are shown in Figure 2. Second, based on the assumption that the question existing near the root node is more important than the further one, a dominance score given to each question was calculated as the reciprocal of the distance from root node to the question node. Subsequently, the dominance scores for each question set were summed up. The 3 highest proportions among 13 sets are shown in Figure 3.

Considering the first criteria from Figure 2, it can be seen that the most tree-occupied question sets are phonetic type (S1), tone type (S3), and part of speech (S7), respectively for all trees. As for the second criteria from Figure 3, the most dominant question sets are little different for each decision trees, i.e., phonetic type (S1), part of speech (S7), and the number of syllables in word (S8) for spectrum tree, phonetic type (S1), tone type (S3), and part of speech (S7) for logF0

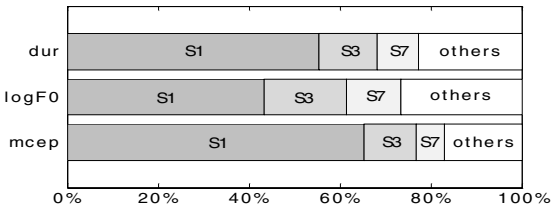

Figure 2: *Three highest proportions of tree occupancy among 13 question sets for spectrum (mcep), pitch (logF0), and state duration (dur) trees. (1st criteria)*

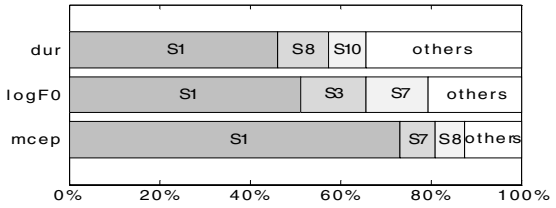

Figure 3: *Three highest proportions of dominance among 13 question sets for spectrum (mcep), pitch (logF0), and state duration (dur) trees. (2nd criteria)*

tree, and phonetic type (S1), the number of syllables in word (S8), and the number of syllables in phrase (S10) for state duration tree. When considering the contribution of tone type question set (S3), the pitch tree is affected at most among all 3 trees.

3.3. Design of Tree-based Context Clustering

First, the full context labels of the speech utterances were constructed according to the necessary context information for all of the questions. These full context labels were employed at the context clustering process in the training stage. Since the syllable tone is very sensitive for Thai perception, the correctness of tone of the synthesized speech must be carefully considered. As a result, four different tree-based context clustering styles were designed; 1) single tree context clustering without tone type questions, 2) single tree context clustering with tone type questions, 3) tone-separated tree context clustering without tone type questions, and 4) tone-separated tree context clustering with tone type questions. The decision tree structure in each state of HMM in the last two styles was modified as shown in Figure 4.

4. Experiments

4.1. Speech Database and Training Condition

A set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC [7] was used for training HMMs. The whole 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-1 and the utterance structure from ORCHID were used to construct the context dependent labels with 79 different phonemes including silence and pause.

Speech signal were sampled at a rate of 16kHz and windowed by a 25ms 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 zeroth coefficient, logarithm of F0, and their delta and delta-delta coefficients [2].

We used 5-state left-to-right HSMMs in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution. Note 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.

4.2. Evaluation of Synthesized Speech

Two experiments were conducted to evaluate the synthesized speech. First, evaluation of tone correctness of the synthesized speech generated from HMM-based system with 4 different tree-based context clustering styles was done. Secondly, evaluation of speech quality in terms of naturalness, clearness, and overall aspects was performed through not only the conventional absolute category rating of mean opinion score (MOS) test but also the comparison category rating (CCR) method.

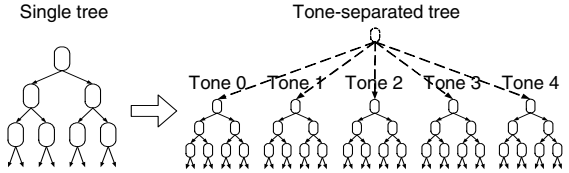

Figure 4: *Tree separation for each tone in Thai.*

4.2.1. Evaluation of Tone Correctness

This section presents how the correctness of the synthesized tone is improved by using contextual factors constructed in section 3.1 and four different tree-based context clustering styles described in section 3.3. Figure 5 shows an example of F0 contours of the natural speech and synthesized speech with different clustering styles. The first full-shape syllable of Figure 5 pronounced as /thă/ conveys tone 4 or rising tone. Figure 5 (a) is of the single tree context clustering without tone questions, however this syllable contour is misshaped. As a result, most listeners perceive it with wrong tone. Meanwhile Figures 5 (b) - (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 Figure 5 (e). To evaluate tone correctness of our system, we investigated 2,289 syllables from 100 synthesized speech utterances. The tone error percentages for all 4 styles are summarized in Figure 6.

We can notice the obvious reduction of the tone error percentage of the second style comparing with the first style. It indicates that the tone type questions play a very important

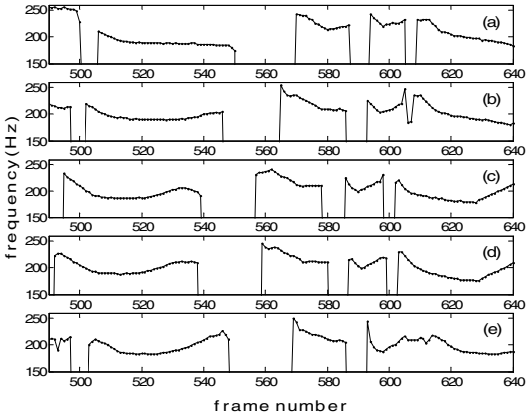

Figure 5: *F0 contours of synthesized speech from 4 different clustering styles; (a) single tree without tone type questions, (b) single tree with tone type questions, (c) tone-separated tree without tone type questions, (d) tone-separated tree with tone type questions, and (e) F0 contour of natural speech.*

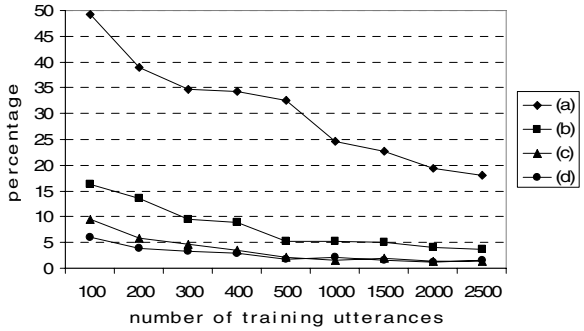

Figure 6: *Tone error percentages of synthesized speech from 4 different clustering styles; (a) single tree without tone type questions, (b) single tree with tone type questions, (c) tone-separated tree without tone type questions, and (d) tone-separated tree with tone type questions.*

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 because the effects from each other tones are reduced significantly. Considering the number of training utterances, the tone error percentage is decreased as the number of training utterances is increased. Note 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 noticeably relieved when the number of training utterances is increased above 500.

4.2.2. Evaluation of Speech Quality

The subjective tests of MOS and CCR were carried out to compare our system with a unit-selection-based TTS system named Vaja from NECTEC. In our HMM-based system, the tone-separated tree context clustering with tone type questions was chosen. In Vaja TTS system, TD-PSOLA and LSP smoothing techniques were applied in the concatenation of demisyllable inventory units extracted from 5200-utterance TSynC-1 database. We used 50 tested utterances for both listening tests. In the MOS test, the synthesized speech material, typically organized in utterances, was presented to 8 Thai subjects for evaluation in overall aspect. The listeners were required to make a single rating from five choices: excellent (5), good (4), fair (3), poor (2), and bad (1). In the CCR test, two different speech samples from Vaja and HMM-based systems were presented to 8 listeners in random order. The listeners then compared the quality of the second one relative to that of the first one in naturalness, clearness, and overall aspects, and gave one out of the seven degrees: much better (3), better (2), slightly better (1), about the same (0), slightly worse (-1), worse (-2), and much worse (-3).

The average scores of all votes in MOS and CCR tests are shown in Figures 7 and 8, respectively. From Figure 7, MOS score of HMM-based system is comparable to Vaja system when the number of training utterances is 1000 or more. From Figure 8, the overall score of HMM-based system is closed to that of Vaja system when the number of training utterances is 1000 or more, that is, the CCR test gives the corresponding results to the MOS test for overall aspect. In clearness aspect,

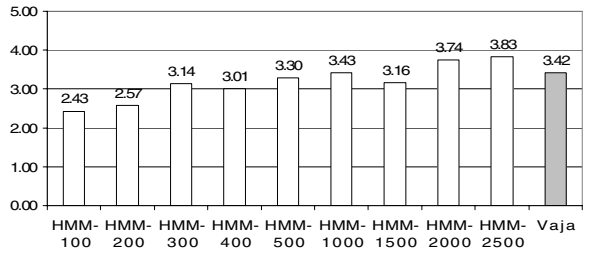

Figure 7: *MOS test for HMM-based & Vaja systems.*

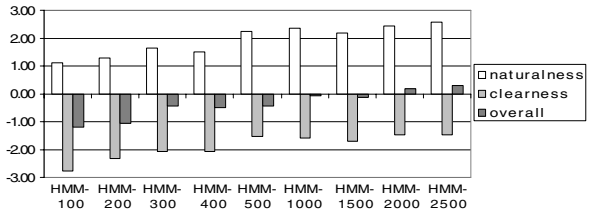

Figure 8: *CCR test for HMM-based system with Vaja system reference.*

HMM-based system is mostly below Vaja system, however in the naturalness aspect, HMM-based system is mostly above Vaja system.

The synthesized speech samples are available on the website: http://www.kbys.ip.titech.ac.jp/demo/thai/index.html

5. Conclusions

A novel approach of HMM-based speech synthesis for Thai language is presented in this paper. A number of specific contextual factors for Thai were constructed to realize the naturalness and the intelligibility of the synthesized speech, while a decision tree structure was designed to improve the tone correctness. The evaluation of the system was conducted by comparing the synthesized speech with that of unit-selection-based system. The results show that our system can generate speech with more naturalness, but less clearness. However, in overall aspect, our system gives a comparable speech quality by using only about one fifth of the speech database size.

6. Acknowledgements

The authors are grateful to NECTEC for providing us the TSynC-1 speech database and the unit-selection-based Vaja TTS system from the website: http://vaja.nectec.or.th/.

7. References

[1] Tokuda, K., Kobayashi, T., and Imai, S., "Speech Parameter Generation from HMM using Dynamic Features", Proc. ICASSP-95, Vol.1, pp.660-663, 1995.

[2] Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S., "Speech Synthesis using HMMs with Dynamics Features", Proc. ICASSP-96, pp.389-392, 1996.

[3] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis", Proc. EUROSPEECH-99, pp.2347-2350, 1999.

[4] Kim, S.J., Kim J.J., and Hahn, M., "Implementation and Evaluation of an HMM-based Korean Speech Synthesis System", IEICE Trans. Inf. & Syst., Vol.E89-D, No.3, pp.1116-1119, 2006.

[5] Saravari, C., and Imai, S., "A Demisyllable Approach to Speech Synthesis of Thai - A Tone Language", J. Acoustic. Soc. Jpn. (E), Vol.4, No.2, pp.97-106, 1983.

[6] Wutiwiwatchai, C., and Furui, S., "Thai Speech Processing Technology: A Review", J. Speech Communication, Vol.49, pp.8-27, 2007.

[7] Hansakunbuntheung, C., Rugchatjaroen, A., and Wutiwiwatchai, C., "Space Reduction of Speech Corpus Based on Quality Perception for Unit Selection Speech Synthesis", Proc. SNLP-2005, pp.127-132, 2005.

[8] Luksaneeyanawin, S., "Linguistics Research and Thai Speech Technology", Proc. of the 5th International Conference on Thai Studies, School of Oriental and African Studies, 1993.

[9] Thathong, U., Jitapunkul, S., and Ahkuputra, V., "Classification of Thai Consonants Naming Using Thai Tone", Proc. ICSLP-2000, Vol.3, pp.47-50, 2000.

[10] Chompun, S., "Fine Granularity Scalability for MP-CELP based Speech Coding with HPDR Technique", Proc. APCCAS-2004, pp.197-200, 2004.

[11] Sornlertlamvanich, V., Takahashi, N., and Isahara, H., "Thai Part-of-speech Tagged Corpus: ORCHID", Proc. Oriental COCOSDA Workshop, pp.131–138, 1998.