American Journal of Applied Sciences 9 (2): 271-277, 2012
ISSN 1546-9239

Thai Speech Phonology for Development of Speech Synthesis: A Review

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: To implementation of the Hidden Markov Model (HMM)-based Thai speech synthesis system, it is necessary to understand the phonology system for a language. Without the phonological information, the contextual factors of tree-based context clustering cannot be completed. **Approach:** The existing speech units in Thai are studied thoroughly so that the synthesis system can provide all of them in an appropriate way. In the study of speech in a specific language, we have to categorize the speech into sounds. Then summarize them in some specific ways including its function, how it appears in speech, the relation with other sounds and also how it contributes to syllable or word. **Results:** The speech units of the phoneme, syllable, word, phrase and sentence level are studied and explained, respectively. **Conclusion:** The important information of Thai phonology system has been summarized. It is expected to apply them efficiently in the HMM-based Thai speech synthesis system.

Key words: Thai phonology system, Thai tone, hidden Markov models, speech synthesis

INTRODUCTION

The study of phonology is the study of the patterned interaction of speech sounds. A fairly obvious observation about human language is that different languages have different sets of possible sounds that can be used to create words. One of the goals of phonology is to describe the rules or conditions on sounds and sound structures that are possible in particular languages. However, in this study we emphasize on the phonological information which will apply with an implementation of the HMM-based speech synthesis (Chomphan and Kobayashi, 2007a; 2007b).

MATERIALS AND METHODS

Vowel Phonemes: Vowels are ones of the most important phonemes in every/most languages. Vowel is the sound that is produced by the wind moving through the vocal chords in the nearly complete closing position. The compressed wind causes the vibration of the vocal chords. The output sound is called voiced sound. This kind of sound is uttered from the mouth without preventing of the wind. However, the mouth organs with different alignments result in different articulatory structures and sounds. There are 21 vowels and 3 composite vowels or diphthongs as shown in Table 1 (Thathong *et al*., 2000; Wutiwiwatchai and Furui, 2007).

Single Vowel Phonemes' Attributes: There are almost 18 single vowel phonemes, i.e., 9 short vowel and 9 long vowel phonemes. Their places of articulation are conforming to the phonetic chart. As for ↦ ↦/ vowels, their places of articulation are much far from advanced vowels, but are little lower than low middle vowels. As a result, the // ↦ ↦// and the /↦ ↦// vowels have very closed places of articulation (Table 1 and 2). For convenience in chart manipulation, these 6 vowels may be allocated as low vowels. The simplified chart is systematized as the Table 3.

Diphthongs' attributes: There are 3 diphthongs in Thai; /◌↦ ½↦ ◌↦/. They are all falling diphthongs, that is, combining between high vowels with a low vowel /↦/. However, in the phonetics study, it may be considered that there are another kind of diphthongs such as the vowels in the following words; ไว //, ลาย /ℓ↦/◌/, เรา /◌↦◌/, หิว /ℏ◌◌/, เร็ว /◌⊙◌/, เลว /ℓ⊙/◌/, แป้ว /≺◌/, แล้ว /ℓ/◌/, คุย / ℏ◌◌/, โชย /⇳ℏ○/◌/, ต่อย /◌◌/, คอย / ℏ/◌/. These vowels are all rising diphthongs resulting from concatenating of low vowels /↦ ↦/ ◌ ⊙ ⊙// ◌ ○// with high vowels /◌/ and /◌/.

Table 1: Thai consonants, vowels and tones

			Places of articulation				
			Labial	Alveolar	Palatal	Velar	Glottal
Manners of articulation	Stops	Voiceless Unaspired	≺	↔	⇳		≫
		Voiceless Aspired	≺ʱ	↔ʱ	⇳ʱ	ʱ	□
		Voiced	↕	○	□	□	□
	Non-stops	Nasal	∓	◆	□	⋰	□
		Fricative	≻	←	□	□	ħ
		Trill	□	→	□	□	□
		Lateral	□	ℓ	□	□	□
		Approximant	⌐	□	◌	□	□
	□	□	Front	Central	Back		
High			◌ ◌/	½ ½/	_ _/		
High-mid			⊙ ⊙/	/	○ ○/		
Low-mid			∵ ∵/		/		
Low				↦ ↦/			
Tone			tone0 `	tone1	tone2 Ý	tone3	tone4[1]

Table 2: Thai vowel system

		Vowel advancement		
		Front	Central	Back
Vowel Height	High	◌ ◌/	½ ½/	_ _/
	High-Mid	⊙ ⊙/	/	○ ○/
	Low-Mid	∵ ∵/ □		/
	Low	□	↦ ↦/	□

Table 3: Simplified Thai vowel system

		Vowel advancement		
		Front	Central	Back
Vowel Height	High	◌ ◌/	½ ½/	_ _/
	Mid	⊙ ⊙/	/	○ ○/
	Low	∵ ∵/	↦ ↦/	/

Even though they are considered phonetically diphthongs, they can be considered a combination of a single vowel and a final consonant of /◌ ⌐/. In other words, the diphthongs /↦◌ ↦/◌ ↦_ ↦/_ ◌_ ⊙_ ⊙/_ _ /_ _◌ ○/◌ ◌ /◌/ should be considered. /↦◌ ↦/◌ ↦⌐ ↦/⌐ ◌_ When comparing among single vowels /◌ ◌/ / /, falling diphthongs /◌↦ ½↦ ⌐↦/ and rising diphthongs /↦◌ ↦/◌ / in the function aspect, the single vowels and the falling diphthongs could have all final consonants, while the rising diphthongs could not appear with any of the final consonants. Moreover, the writing appearance of the words with these rising diphthongs shows explicitly that these words have a final consonant of /◌ ⌐/ in nearly most of existing words. As a result, these rising diphthongs should be considered as a single vowel with a final consonant of /◌ ⌐/. The word ไว /⌐↦◌/ is therefore analyzed that it consists of 4 phonemes including an initial consonant /⌐/, a vowel /↦/, a final consonant /◌/ and a middle tone. It is represented by the phonetic /⌐↦◌0/. The word เรา /⌐↦⌐/ is also analyzed that it consists of 4 phonemes including an initial consonant /⌐/, a vowel /↦/, a final consonant /⌐/ and a middle tone. Therefore, it is represented phonetically as /⌐↦⌐0/. The phonemes /◌/ and /⌐/ can be considered as special phonemes which have at least 2 allophones. The phonemes /◌/ has an allophone set of {/◌/ /◌/}. The allophone /◌/ appears in initial consonant, while the allophone /◌/ appears after the vowels /↦ ↦/ ◌ ⊙ ⊙/ //. These 21 phonemes of vowels (9 short single vowels, 9 long single vowels and 3 diphthongs) contribute as the core of a syllable in Thai. Its function is to form a syllable with an initial consonant and a final consonant. They can be appeared with any of tones, but only some of the initial or final consonants.

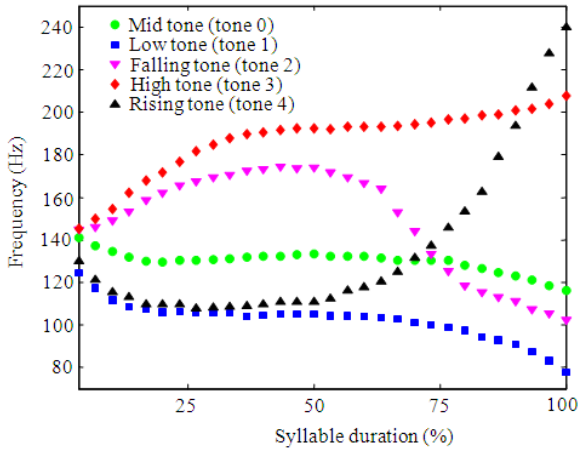


Fig. 1: Standard F0 contours for Thai tones

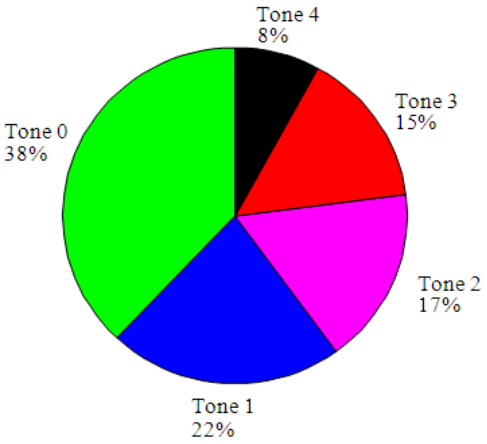


Fig. 2: Proportions of Thai tone occurrence frequencies from TSynC-1 speech database

Consonant phonemes: The consonant is the sound which generated by the output wind from the vocal chords modified by the mouth and nose organs. There are 44 appearances with only 21 sounds (Iwasaki and Horie, 2005).

The initial single consonants are /≺ ≺ħ ↔ ↔ħ ⇳ ⇳ħ ħ ↕ ○ ∓ ◆ ∴ ≻ ← ħ → ℓ ⌙ ○ ≫/. These 21 consonants are of the sounds /ป พ ต ท จ ช ก ค บ ด ม น ง อ ฟ ส ฮ ร ล ว ย อ/. There are 12 composite initial consonants including /≺→ ≺ℓ ↔→ → ℓ ← ≺ħ→ ≺ħℓ ↔ħ→ ħ→ ħℓ ħ⌙/. The first consonant of these composite consonants is of /≺ ≺ħ ↔ ↔ħ ħ/, while the second one is of /→ ℓ ⌙/ only. The 9 final consonants are /ˊ ↔ ⇳ ∓ ◆ ∴ ⌙ ○/. In syllable generation, the combinations between an initial consonant and a vowel are also existed.

Tone: The variation of height of initial consonant and vowel distinguishes in meanings of words in Thai, this is the definition of tone in Thai. Generally, there are 5 tones including /ก, ก่, ก้, ก๊, ก๋/ in Thai or /no-marking, Ý ¹/ in IPA (Palmer, 1969).

For tonal languages such as Thai, tone, which is indicated by contrasting variations in contour of 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. In Thai, there are five tonal variations traditionally named according to the characteristics of their F0 contours within a syllable as shown in Fig. 1. Five IPA tone markers are generally used to indicate Thai tone types; / ˋ/ for middle tone (tone 0), / for low tone (tone 1), / Ý for falling tone (tone 2), / / for high tone (tone 3) and / ¹/ for rising tone (tone 4). The effect of tone on the linguistic meaning is shown in the following examples: the syllable / ħ↦ˋ// (/คา/ in Thai) has tone 0 and means “to get stuck”, the syllable / ħ↦ // (/ข่า/ in Thai) has tone 1 and means “galangal, a kind of spice”, the syllable / ħ↦Ý// (/ฆ่า/ in Thai) has tone 2 and means “to kill”, the syllable / ħ↦ // (/ค้า/ in Thai) has tone 3 and means “to trade” and the syllable / ħ↦¹// (/ขา/ in Thai) has tone 4 and means “leg”. By investigating tone occurrence frequency in TSynC-1 speech database, we found that 77,413 syllables are occupied in descending order by tone 0, tone 1, tone 2, tone 3 and tone 4, respectively. Fig. 2 shows the proportions among all five tone occurrence frequencies.

The most important characteristics of a speech synthesis system are naturalness and intelligibility. Tone distortion can deteriorate not only the speech intelligibility as described above but also the speech naturalness, since the lexical tone is a suprasegmental feature formed by the basic prosodic feature, i.e., F0. Meanwhile the other important basic prosodic features including phrasal pauses, duration and energy can affect mainly the speech naturalness. Therefore the tone correctness must be carefully taken into account in the tonal languages (Abramson, 1979).

T
Ci(Ci) V(V) Cf

Fig. 3: Thai tonal syllable structure

In the continuous speech context, the F0 patterns of 5 Thai tones are affected from the adjacent syllable tones. Palmer demonstrated that the 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. Gandour studied the tonal coarticulation including the carry-over effects and the anticipatory effects. There is a study on tone sandhi in Thai, i.e., Thompson studied on a particular southern Thai dialect. However, it has not been widely applied to the standard Thai. Our approach, in contrast, applies a simple contextual syllable tones in the context clustering process without using any rules or heuristics.

In tone categorization, two criteria are used to categorize Thai tones into tone groups as follows. First, by considering the constancy of the F0 contour, Abramson divided the tones into two groups: the static group (level tone) consists of three tones, high tone, middle tone and low tone; the dynamic group (contour tone) consists of two tones, rising tone and falling tone. Secondly, by considering each contour of Fig. 1, we can see that the F0 patterns of the mid, low, falling, high and rising tones are relatively mid-fall, fall, rise-fall, rise and fall-rise, respectively. As a result, 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.

It should be noted that there is another type of special tone called intensifying tone. It is another kind of tone which is unable to define the writing pattern of Thai. It is usually appeared in the speaking conversation. Its attribute combines both rising and falling tone in one syllable. The F0 level begins at somewhat high level and climbs upward above high level of all other tones and then falls a little bit at the end of syllable. This kind of tone appears only in the repeating word which intensifies the first syllable to show the special meaning of that word. The following words represent the existing of this intensifying tone.

Syllable: As for meaning and boundary of Syllable, syllable is the smallest unit of speech to communicate with others. Generally, the native speaker can define how many syllables exist in a word. This is called mora in other non-tonal languages such as Japanese. For instances, the word /เรียน/ has only one syllable, /เรียน/ has only one syllable, /นิสิต/ has 2 syllables, /จุฬาลงกรณ์/ has 4 syllables and /มหาวิทยาลัย/ has 6 syllables.

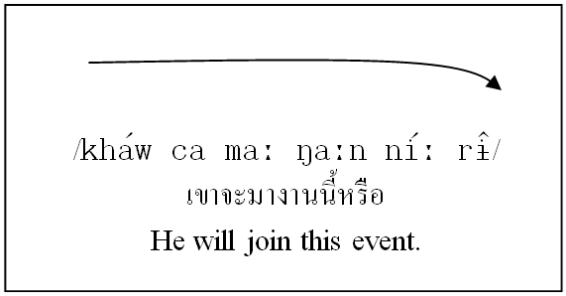
Each syllable existing in a word may have different in dominance. The dominant sound means the sound which is louder than other sounds in the uttered group of sounds.

As for syllable composition (Syllable Structure), a comprehensive description of Thai sound system was published by Lukseneeyanawin (Thathong *et al*., 2000; Wutiwiwatchai and Furui, 2007). Thai sound is often described in a syllable unit as depicted in Fig. 3. 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 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 /≺ ≺ħ ↔ ↔ħ ħ/ with one of the phonemes /→ ℓ ⌟/.

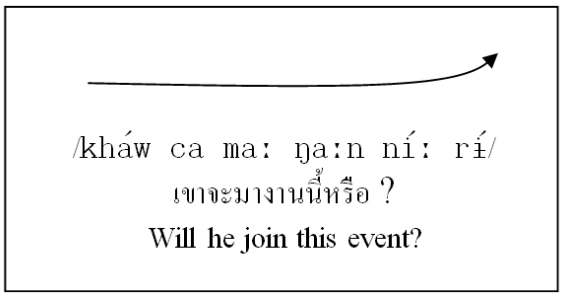
Recently, some loan words which do not conform to the rules of native Thai phonology, such as the initial consonants /↕→ ↕ℓ ≻→ ≻ℓ ○→/ and the final consonants /≻ ← ⇕ħ ℓ/ have begun to appear. These consonants are also included in our speech database. Most of them are used in the training stage of our implemented system, however only some of them are randomly selected into the target texts to be synthesized in the evaluation process.

Word: When considering the pronunciation of syllables, words in Thai can be categorized into monosyllable and multisyllable words. As for the multisyllable words, they may be divided into 2-syllable, 3-syllable, 4-syllable and several syllable words. However, most of words are generally monosyllable and 2-syllable words. As for several syllable words, they are adopted from Bali or Sanskrit, or are composite words. The more syllables are there in word, the less the words are there.

In the multisyllable words, stressing of syllable is rather complicated but is not systematically defined as a rule. That is we do not force the stressing pattern into our system, but let the stressing are formed by training of the observations in our speech database.



(a)



(b)

Fig. 4: Examples of intonation patterns (a) Declarative sentence-falling intonation (b) Question sentence-rising intonation

Part of speech: The part of speech explains the ways that words can be used in various contexts. Every word in the Thai language functions as at least one part of speech; many words can serve, at different times, as two or more parts of speech, depending on the context. The part of speech in Thai are classified for using in constructing the Thai speech text corpus named ORCHID. This classification of the part of speech is used in constructing of the contextual factor in the context clustering process.

Intonation: The intonation is the level of sound that exists along a sentence. It is not a speech unit which varies the meaning of word, but it is an important factor to indicate the meaning of sentence. In other words, the change in intonation causes the derivation in meaning of sentences with the same meaning of words. The intonation is also considered a kind of suprasegmental feature of the natural speech.

There are two dominant patterns of intonation, falling intonation and rising intonation. As for falling intonation, this intonation pattern has general characteristics as follows. The beginning of sentence has high sound level and the end of sentence has low sound level. It appears generally in the declarative sentence. An example of this kind of intonation pattern is shown in Fig. 4a. As for rising intonation, this intonation pattern has general characteristics as follows. The beginning of sentence has low sound level, while the end of sentence has high sound level. It appears normally in the question sentence and some kinds of directive sentences. An example of this kind of intonation pattern is shown in Fig. 4b.

RESULTS

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. 5. 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) (Tokuda *et al*., 1999; Chomphan and Kobayashi, 2008; 2009), respectively. In addition, to directly model the phone durations, we utilize a framework of Hidden Semi-Markov Model (HSMM) (Chomphan and Kobayashi, 2007a; 2007b), 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 is applied separately to the spectral and the F0 parts of the context-dependent phoneme HMMs (Levinson, 1986; Yamagishi *et al*., 2002).

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. They are divided into five levels of speech units, including phoneme, syllable, word, phrase and utterance (Riley, 1989).

The extraction algorithms for tonal features were used with the F0 series of all training utterances to prepare the tonal features to be employed in the context-clustering process.

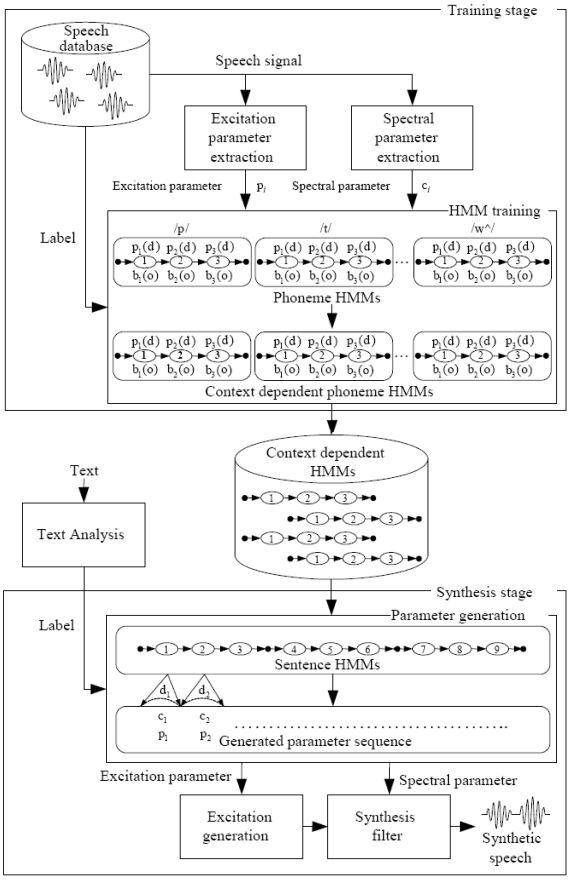


Fig. 5: HMM-based speech synthesis system

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. Therefore, we also provided the contextual factors for these features with preceding, current and succeeding syllable positions (Chomphan and Kobayashi, 2007a; 2007b; Zen *et al*., 2004).

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

DISCUSSION

From the nineteen sets of contextual factors, we can apply it in the context clustering process of the

speaker-dependent HMM-based speech synthesis system. Each set compositely improves the synthesized speech. An approach of HMM-based Thai speech synthesis is shortly presented in this study. The speaker-dependent system was implemented with high tone intelligibility when using the tree-based context clustering.

CONCLUSION

Thai Speech Phonology has been studied in this study. It describes the rules or conditions on sounds and sound structures that are possible in Thai language. The explanations are ranged from phoneme, tone, syllable, word, part of speech, to intonation. The information of these speech units are applied to construct the questions used in tree-based context clustering process of the HMM-based Thai speech synthesis. The implemented speaker-dependent system gives the synthesized speech with high tone intelligibility when using the designed tree-based context clustering.

ACKNOWLEDGEMENT

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

REFERENCES

Abramson, A.S., 1979. Lexical tone and sentence prosody in Thai. Proceedings of the 9th International Congress of Phonetics Science (ICPS' 79), University of Copenhagen, Copenhagen, Denmark, pp: 380-387.

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, 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

Iwasaki, S. and I.P. Horie, 2005. A Reference Grammar of Thai. 1st Edn., Cambridge University Press, Cambridge, ISBN: 0521650852, pp: 392.

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

Palmer, A., 1969. Thai tone variants and the language teachers. Language Learn., 19: 287-300. DOI: 10.1111/j.1467-1770.1969.tb00469.x

Riley, M.D., 1989. Statistical tree-based modeling of phonetic segment durations. J. Acoust. Soc. Am., 85: S44-S44. DOI: 10.1121/1.2026979

Thathong, U., S. Jitapunkul, V. Ahkuputra, E. Maneenoi and B. Thampanitchawong, 2000. Classification of Thai consonant naming using Thai tone. Proceedings of the 6th International Conference on Spoken Language Processing, Oct. 16-20, ISCA Archive, Beijing, China, pp: 47-50.

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

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.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 2004. 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: 1393-1396.