IMPROVING NATURALNESS OF THAI TEXT-TO-SPEECH SYNTHESIS BY PROSODIC RULE

Pradit Mittrapiyanuruk, Chatchawarn Hansakunbuntheung,
Virongrong Tesprasit and Virach Sornlertlamvanich

Information R&D Division, National Electronics and Computer Technology Center (NECTEC).
Gypsum Metropolitan Bldg., 22nd Fl., 539/2 Sri Ayudhaya Rd., Rajthevi, Bangkok 10400, Thailand.
(pmittrap, chatchawarnh)@notes.nectec.or.th, (virong, virach)@nectec.or.th

ABSTRACT

This paper presents a method to improve the naturalness of Thai Text-to-speech synthesis, in 4 main parts. In the pausing module, its main function is to determine the break location when synthesizing a Thai text which has no explicit sentence/phrase/word boundary. In the syllable duration and tone generation, a set of rules is provided to generate proper prosodic parameters for synthesizing more natural speech. The syllable duration rule is applied using the Klatt's method to handle the task in syllabic frame. The tonal rule considers the effect of tonal coarticulation and F0 downdrift in generating the F0 contour parameter. In the demisyllable concatenation, the TD-PSOLA technique is applied to modify the waveform for obtaining the required prosody. The LSP-based concatenated boundary smoothing is also included to imitate the cross-syllable coarticulation effect. The result of comparative quality test shows a significant improvement in our proposed method.

1. INTRODUCTION

The Thai language has unique characteristics in both writing and phonological system. For examples, in a common Thai text, there is no explicit use of punctuation marker to determine a sentence, phrase and word in a paragraph. Differing from other languages, Thai is a tonal language in which the tonal perception relates to the meaning. For instance, the word [su:aj0] having a mid tone (0), means "unlucky" while the word [su:aj4] which has the same phoneme but rising tone (4) means "beautiful". The tone in Thai is very meaningful. After having reviewed the approaches which have been successfully applied for English or other languages text-to-speech synthesis (TTS), we found that they can not be applied directly for the Thai language. In this work, we focus our research on handling the unique characteristics of Thai for improving the naturalness of our Thai TTS system.

It is undoubtedly acknowledged that prosody is the most essential impact to improve the naturalness of the synthetic speech. The four topics focussed in this paper are the pausing, the syllable duration rule, the tonal rule and the synthesis method. The rest of this paper is arranged as following. Section 2 discussed the main function of pausing to determine the suitable pause location when synthesizing a large text as the human behavior when speaking a long utterance. The algorithm is to recover the sentence/phrase/word boundary markers that latently reside in the Thai text. In section 3, the syllable duration rule is developed to assign the length of each syllable in utterance. If the ratio of these lengths is mutually proportional then we certainly obtain more natural speech. Furthermore, the result of investigating the spontaneous human speech shows that there are tonal phenomena that made an individual syllabic tonal pattern deviating from the pattern of isolated syllable pronunciation. To simulate these phenomena in our system, the tonal rules are developed as presented in Section 4. Finally, LSP-based concatenated boundary smoothing is introduced to create the output speech, as shown in Section 5. In section 6, the result of preliminary test is reported to evaluate the improvement.

2. PAUSING

Generally, when speakers pronounce any text material they tend to pause at some positions for many purposes such as breathing, separating the phrases, emphasizing the importance of phrases etc. In the view of TTS, these pause locations express the important prosodic characteristic e.g. lengthening duration and the F0 downtrend of syllables. If these locations are accurately determined then the prosody generation module can certainly play an important role in synthesizing a high natural speech. In this work, the pause location is assigned at the end-of-sentence, at the end-of-phrase and after a punctuation mark. Since in the Thai writing system there is no explicit sentence, phrase and word boundary marker, these locations need to be determined.

Thai sentence extraction algorithm [6] is developed to detect the sentence break position from the input text which may be one or more paragraphs. Fortunately, it is conventional to insert the space at the end of a sentence in Thai writing. But not all spaces in the text are the end-of-sentence marker. They are also used in other purposes such as using between phrases or clauses in a sentence, before and after numerals, etc. The algorithm extends the POS tagging in probabilistic n-gram model to discriminate the actual sentence break spaces from the other purpose spaces. The detail of this algorithm can be found in [6]. Knowing the sentence boundary makes it possible to processing a large consecutive text with ease.

The phrase break positions in a sentence is detected by using a rule-based approach. The result from the sentence extraction, a sequence of words together with the corresponding part-of-speech (POS), is taken as the input to this rule set. The algorithm consists of 2 steps. First step, it determines the tentative phrase break positions in a sentence. These positions are at any of (1) the punctuation mark or (2) before/after/between some specific grammatical words. The actual phrase break space is also considered as punctuation. The actual phrase break space is distinguished from other spaces in a sentence by using distinctive pattern rules derived from the formal Thai writing pattern [8]. An example of rule (2) is placing a break between a content word and the following function word. The function word can be a conjunction, a preposition or a relative pronoun. Another example is placing a break after the ending word, which is added to indicate the mood of an utterance. Second step, to avoid the overfrequent pausing, the tentative break forming a phrase which number of syllables less than the threshold is eliminated. This threshold is

the result of the study in [5] reporting that the position of pause occurs at an average of once in every 8 syllables (S.D.=4). Normally, the threshold that is selected relating to [5] is in the range of 8 to 12 syllables.

3. SYLLABLE DURATION RULE

In timing aspect, if the ratio of duration of every segment in an utterance is mutually proportional then we certainly obtain a more natural speech. Because the synthesis technique is based on the syllabic frame and the information specifies the boundary smaller than syllable segment does not exist either, the segment we consider in this work is a syllable. With the claim that the duration effects are language-independent in nature [3], we apply the Klatt's method [4] which is successfully done in other languages to Thai to handle the task in the syllabic frame. This prevalent method is a sequential rule system. First it assigns the default duration from its intrinsic property to each syllable and then the duration of a syllable is modified by the set of scaling factor corresponding to each successive rule. The intrinsic duration of each syllable is different according to the phoneme identities of a syllable (initial consonant, vowel and final consonant). Because this value is measured only for the mid tonal (tone 0) syllable, the duration of other tonal syllable is generated from the duration of the mid tonal syllable. From the investigation of Thai speech sample we found that only the falling (tone 2) and rising (tone 4) tonal syllables have a longer duration comparing with the mid tonal syllable. Then the duration of falling (tone 2) and rising (tone 4) tonal syllables are scaled-up by the factor 4/3 to compensate the interactive effect betweetone and duration interactive effect.

For the rule, the contextual effect in word, phrase and sentence level determines the scaling factor in each rule. The rule in phrase and sentence level considers the final-lengthening effect. It lengthens the sentence-final and phrase-final syllable duration with the same factor of 1.2 and inserts a pause after these syllables. The duration of sentence pause is longer than the phrase pause. In word level, the polysyllabic word effect is considered. The duration of a syllable varies according to the position in the word. The duration of when it is in the final position is longer than when it is in the beginning position. And, the duration of when it is in the beginning position is longer than when it is in the intermediate position. As a result, the duration of syllable in the word-intermediate position is shortened by the factor of 0.85 and in the word-initial position is shortened by the factor of 0.9. In addition, the “postvocalic consonant context of vowel” has an effect on the proceeding vowel duration in shortening the vowel duration if the consonant is voiceless. Based on the assumption that the vowel covers most part of syllable, the duration of syllable that has no final consonant or the final consonant is voiced and followed by the syllable whose initial consonant is voiceless, is shortened with the factor of 0.9.

4. TONAL RULE

In Thai, there are five tones i.e., ‘Mid’ or ‘0’, ‘Low’ or ‘1’, ‘Falling’ or ‘2’, ‘High’ or ‘3’ and ‘Rising’ or ‘4’ as shown in figure 1. Tones play an important role in Thai. The meaning of words with the same phoneme sequence may be different if they comprise of different tone. Furthermore, tones also play an another role in naturalness, especially in the prosody aspect. A tonal rule and some tonal effects are used to generate and smooth the F0 contour. In this paper, we present the paragraph reading model. There are the two classified levels of the rule and the effects. The first level is the suprasegmental level. This level consists of phenomena over segmental level i.e., downdrift which is a phenomena of intonation. Another level is segmental level that i.e., of the tonal coarticulation.

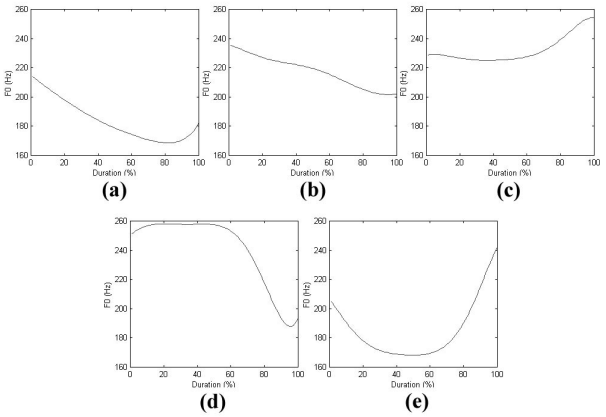

Figure 1 : Five Thai tone-contour prototype.
(a) Low tone, (b) Mid tone, (c) High tone,
(d) Falling tone, (d) Rising tone

4.1 Downdrift

Like many languages, Thai has a gradual downdrift in the value of F_0 across a phrase. This phenomenon makes a F_0 level of the preceding tone contour higher than the following one. For this reason, the same tone-type contours in a same phrase, but at different time, will be placed at the different F_0 level. However, This inclination does not affect a tonal recognition.

The downdrift is integrated and simplified as an asymptote; called baseline, as shown in figure 2. All generated tone contours refer to this baseline. In this case, slopes of asymptote of a Thai male speech and a Thai female speech prototypes are 20 Hz/sec and 30 Hz/sec respectively. A baseline frequency is always reset to a maximum value at the beginning of each phrase.

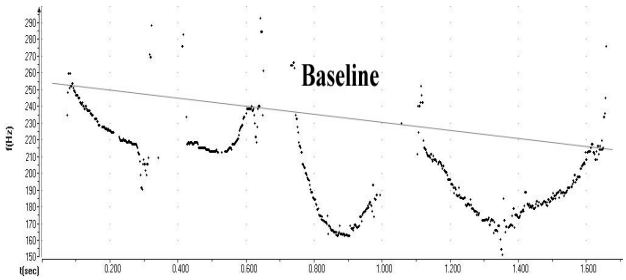

Figure 2 : An asymptote of downdrift in Thai female speech.

According to the downdrift, the tone contours are located relatively with the baseline. In order to locate a tone contour, we define a reference point to all tonal contour patterns. The reference point is calculated by computing an average F_0 of the entire tone-contour prototypes. The reference point comparing with Thai tones is shown in figure 3. In addition, each tone contour has a F_0 offset which is used to refer to the reference point. A F_0 offset is a displacement between a starting F_0 value of a tone contour and its reference points as shown in figure 4.

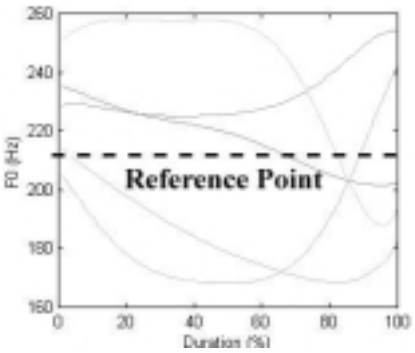

Figure 3 : The reference point comparing Thai tones.

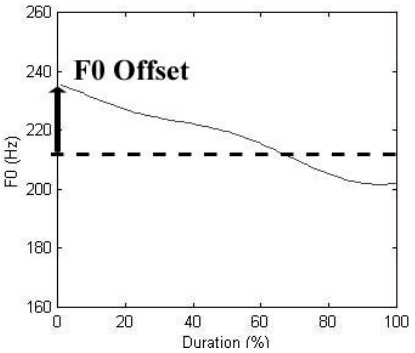

Figure 4 : An example of a F_0 offset.

To locate a tone contour on a baseline, the reference point of the tone contour is located on the baseline. The starting F_0 is calculated from its F_0 offset. An example of locating tonal contours is shown in figure 5.

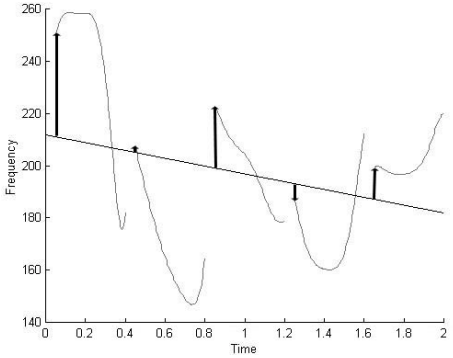

Figure 5 : Locating tone contours.

4.2 Coarticulation

One natural phenomenon of continuous speech, which makes tone patterns alternate from their normal patterns, is coarticulation. In Gandour's work [2], he reported that the Thai tonal coarticulation affects both height and slope of tonal contour bent to adjacent tones, a preceding and a succeeding one. An effect of the preceding tone is called carry-over coarticulation and an effect of the succeeding tone is called anticipatory coarticulation. These effects makes the natural speech differs from a synthetic speech, without prosodic modification, in which the tone patterns of the natural speech are deviated from individual isolated tone pattern.

The change on slope in Thai is insignificantly affected comparing with the change on height [2]. Therefore, only the change on height will be implemented. By computing a pitch difference with adjacent tones, a bending direction and a pitch variation will be known. When the pitch difference is positive, a current tone is higher and the bending direction is downward to the adjacent tone and vice versa. In addition, each tone has a different potential to bend the adjacent tone pattern. The change does not affect through its entire duration but only 75% of its duration for carry-over effect and 50% of its duration for anticipatory effect [2] as expressed in equation (1) and shown in figure 6. For an example, the figure 7 shows a result of coarticulation on the F_0 contour from figure 5.

$$FF(t) = (D_{pt}-t)*c_{pt}*\Delta FF_{pt}*u(D_{pt}-t) + FF_{org}(t) + (t-D_{entire}+D_{st})*c_{st}*\Delta FF_{st}*u(t-D_{entire}+D_{st}) \quad (1)$$

$FF(t)$	New F_0 Value at time t
D_{entire}	Entire duration
D_{pt}	Duration that influenced by a preceding tone
D_{st}	Duration that influenced by a succeeding tone
c_{pt}	Carry-over coefficient
c_{st}	Anticipatory coefficient
ΔFF_{pt}	F_0 difference comparing with a preceding tone
ΔFF_{st}	F_0 difference comparing with a succeeding tone

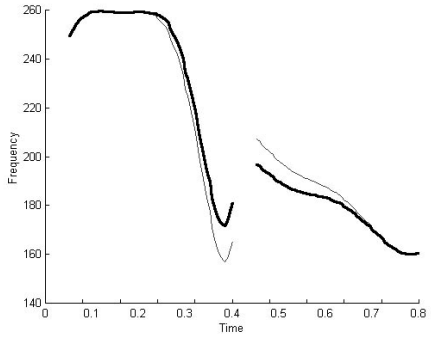

Figure 6 : A coarticulation effect on a change in tonal height.

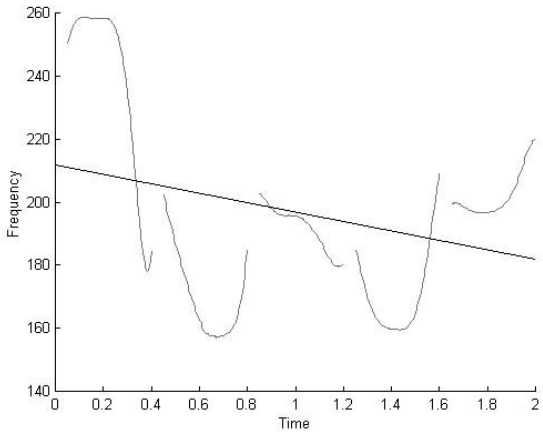

Figure 7 : An example of the coarticulation.

5. SYNTHESIS METHODS

To synthesize any utterance waveform from the linguistic description and prosody parameter derived from the text, the waveform is created from the concatenated sequence of syllabic segment waveform. Each syllable waveform is formed by the concatenation of demisyllable inventory unit. The total number of unit in the inventory is 605. The prosody parameters generated from the previous sections are captured into the waveform to obtain the required prosodic speech by the most widespread prosodic modification technique, Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) [1]. The Inter-demisyllable boundary is smoothed on the spectral, amplitude and pitch domain. For the pitch domain, the TD-PSOLA implicitly handles the pitch smoothing. In spectral and amplitude domain, the Line Spectrum Pairs (LSP) [7] is utilized. Two advantages of LSP are that (1) its parameter sequence and its parameter values correspond to speech formants (2) it has a stability on interpolating parameters. For this reason, the waveform of joining area is recreated by LSP synthesis technique. The waveform of several pitch-synchronous frames in the area of concatenated boundary is transformed into the LSP parameters. These parameters are smoothed by the linear interpolation and then they are transformed back into the smoothly concatenated waveform.

For the syllable concatenation, the coarticulation effect at cross-syllable boundary is modeled. We derive two syllabic joining patterns from investigating the natural speech sample. The first pattern is smoothly continuous linkage where the pitch, amplitude and spectral assimilation occur at the boundary. This pattern occurs when the boundary phonemes of joining syllable are voiced such as /b/, /m/, /ng/ etc. The pitch or tonal contour assimilation is already handled by the tonal rule. To imitate the spectral and amplitude assimilation at cross-syllable boundary, the waveform of joining area is recreated by the same technique used in the inter-demisyllable smoothing. The another joining pattern is the simple touching. This pattern occurs when one or both of the boundary phonemes of joining syllables is unvoiced. If the boundary phonemes are either plosive or glottal stop then the pre-plosive or glottal closure pause with 25 ms in length is inserted between them.

6. EVALUATION

To evaluate the success of our proposed method, we establish the comparative quality test between the synthetic speech with and without prosodic adjustment. With 20 subjects, each subject listens 50 different synthetic sentences both in prosody and no prosody version. Then they ask to value the judgement score 5 point scales (5 – Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 – Bad) by considering the naturalness aspect. The MOS (Mean opinion score) is calculated. For the non-prosodic adjusted speech, the MOS is 2.6 (quality), and the prosodic adjusted speech is 3.1 (quality). From the result, it shows that the proposed improvement in this work made significantly better natural quality.

7. CONCLUSION

We have presented the method to improve our Thai TTS. Most of them are the rule based. The topics we present are the pausing, the syllable duration rule, the tonal rule and the synthesis method. The pausing module consists of 2 major functions: the sentence extraction and the rule-based phrase boundary determination. The syllable duration rule is adapted from the Klatt's method [4] to handle the task in the syllabic frame. The tonal rule considers the F0 downdrift and tonal coarticulation effect [2] in generating the F0 contour. The demisyllable concatenation, TD-PSOLA and LSP smoothing are applied in the synthesis method. The result of preliminary test shows significant improvement.

8. REFERENCES

1. Charpentier, F. and Moulines, E. 1989. *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones.* European Conference on Speech Communication and Technology, pp. 013-019.
2. Gandour, J. T., Potisuk, S., and Dechongkit, S. 1994. *Tonal Coarticulation in Thai*. Journal of Phonetics, vol 22, pp.477-492.
3. Klatt, D.H. 1976. *Linguistic uses of segment duration in English.* Journal of the Acoustic Society of America, 59, pp. 1208-1221.
4. Klatt, D.H. 1987. *Review of text to speech synthesis conversion for English.* Journal of Acoustic Society America, Vol 82, pp.737-793.
5. Luksaneeyanawin, S. et al. 1992. *A Thai text-to-speech system.* Proceedings of 4th NECTEC conference, pp.65-78. (in Thai).
6. Mittrapiyanuruk, P. and Sornlertlamvanich, V. 2000. *The Automatic Thai Sentence Extraction.* Proceedings of 4th symposium on Natural Language Processing (SNLP'2000), pp.23-28.
7. Smith, A.C. and van Schalkwyk, J.J.D. 1988. *Line-spectrum pairs-a review*. Proceedings of Southern African Conference on Communications and Signal Processing (COMSIG 88), pp.7-11.
8. Thavaranon, K. 1978. *Spacing in Thai writing.* Master Thesis, Department of Thai, Chulalongkorn University (in Thai).