# Prosody-based Naturalness Improvement in Thai Unit-selection Speech Synthesis

A. Rugchatjaroen, A. Thangthai, S. Saychum,
N. Thatphithakkul and C. Wutiwiwatchai
Human Language Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC)
Patumthani, Thailand

*Abstract*-**This paper presents naturalness improvement in Thai unit-selection text-to-speech synthesis (TTS) based on prosody modeling. Although several modeling approaches of prosodic parameters in Thai speech have been proposed, they have not been proven to provide a promising performance when practically assembling in a synthesizer. In this paper, two learning machines for phrase break and phoneme duration prediction are integrated in a Thai unit-selection TTS system. Evaluations focusing on the naturalness improvement are conducted both subjectively and automatically. The listening test gives 0.47 MOS improvement from the baseline system and the objective measures obviously show the better quality of the proposed system.**

## I. Introduction

Text-to-speech synthesis (TTS) has been widely implemented in many kinds of applications. It is particularly important for people with disabilities. At present, there are some developed systems for Thai, which almost present in the way of unit-selection algorithm. There exist some TTS systems developed based on the demisyllable-unit concatenation algorithm with prosody modification. However, automatic controlling of prosodic parameters is still far from natural human sounds.

To improve the naturalness of synthetic speech in Thai, several components need to be considered, including word segmentation [9]-[11], sentence boundary detection [12], and grapheme-to-phoneme conversion [14]-[16]. Prosodic prediction often employs intelligent learning machines or statistical models trained by a large prosody-annotated speech corpus. Given the predicted parameters, modifying speech signals can be performed efficiently by Time Domain Pitch Synchronous OverLab-Add (TD-PSOLA) [4].

At present, Vaja [1], a Thai speech synthesizer developed by NECTEC in Thailand, is utilizing the unit-selection technique on a well-organized speech corpus named TSynC [6]. The current version, treated as baseline system in this paper, incorporated a simple rule-based phrase breaking model and had no prediction of phoneme duration. This paper presents the improvements by objective and subjective tests on a modified system, where in a machine-learning based phrase break detection and a linear-regression based duration model are incorporated. The next section will describe two Thai prosodic-parameter prediction algorithms, including phrase breaking and duration modeling. Section III presents experiments, and results. Conclusion and future works are given in Section IV.

## II. Prosody-modified thai tts system

Recently, modeling of two prosodic parameters, phrasal pauses and phoneme durations for Thai, has been extensively researched [2, 3, 7, 8]. This section reviews proposed algorithms, which will be integrated in the baseline TTS system. Figure 1 shows an overall structure of our TTS system with the two additional modules for phrase breaking and phoneme duration prediction. An example of intermediated outputs is given beside.

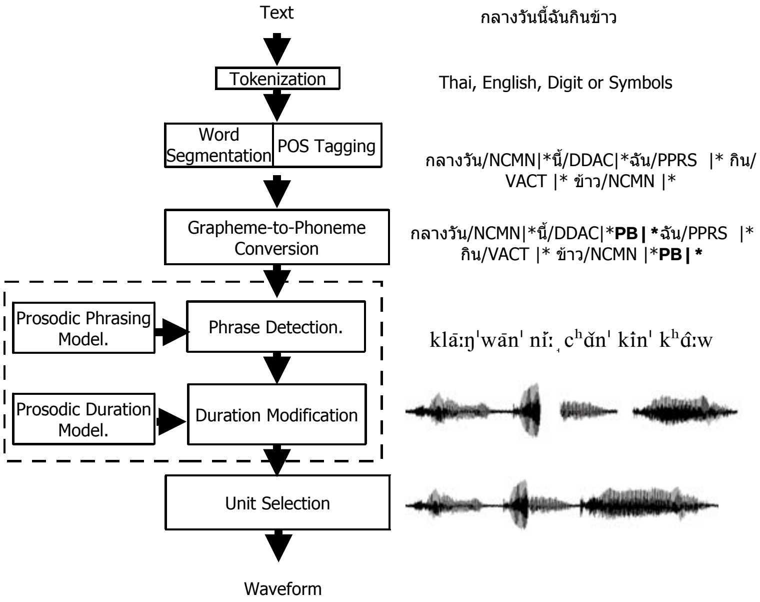
Text
กลางวันนี้ฉันกินข้าว
Tokenization
Thai, English, Digit or Symbols
Word Segmentation
POS Tagging
กลางวัน/NCMN|*นี้/DDAC|*ฉัน/PPRS |* กิน/VACT |* ข้าว/NCMN |*
Grapheme-to-Phoneme Conversion
กลางวัน/NCMN|*นี้/DDAC|*PB|*ฉัน/PPRS |* กิน/VACT |* ข้าว/NCMN |*PB|*
Prosodic Phrasing Model.
Phrase Detection.
klāːŋˈwānˈ níːˌ cʰǎnˈ kīnˈ kʰâːw
Prosodic Duration Model.
Duration Modification
Unit Selection
Waveform

Figure 1. A structure of baseline Thai TTS system.

To predict prosodic parameters, a set of information (features) is extracted from an annotated speech corpus as shown in Table I. P and D denotes the task of phrase break and phoneme duration prediction.

TABLE I
Features used for prosody prediction

| Feature | Description | P | D |
|---|---|---|---|
| *POS* | Part-of-speech (POS) of four contextual | | |

| Feature | Description | P | D |
|---|---|---|---|
| | words | | |
| *NSyl* | The number of syllables between the previous phrase break and the current juncture | | |
| *NWrd* | The number of words between the previous phrase break and the current juncture | | |
| *SylRatio* | The ratio of *NSyl* and the total number of syllables in the previous phrase | | |
| *WrdRatio* | The ratio of *NWrd* and the total number of syllables in the previous phrase | | |
| *Phone* | Five phonemes including the current phoneme and four contextual phonemes | | |
| *PosWrd* | Position in word (begin, mid, end) | | |
| *PosPhr* | Position in phrase (begin, mid, end) | | |
| *Tone* | Tone of the current syllable (five Thai tones) | | |

### A. *Phrase break prediction*

In this section, we describe a practical approach for phrase detection. In literatures [7, 8], five candidates of the learning methods; for Thai POS sequence model, CART, RIPPER, SLIPPER and neural network, were used in prosodic phrase-break prediction. Experiments mostly showed that CART gave almost the highest performance. Since CART can be simply implemented and consumes a small processing time, this paper uses CART to predict phrase breaks.

Some features related to sentence boundaries proposed in [7, 8] are difficult to incorporate in the practical system due to unsolved problems of Thai sentence boundary detection. Not only an unreliable result, but also a high computational time is required by the current sentence breaking module. Therefore this paper ignores such features and proposes a new set of features independent of sentence boundaries as shown in Table I.

### B. *Phoneme duration prediction*

Duration modeling for Thai speech synthesis is dominated by control factors (*Phone*, *PosWrd*, *PosPhr*, *Tone*), which depend on the structural design of Thai syllables, positional difference in phrase and syllable, and syllabic tones as shown in Table I.

Prediction is done by statistical linear regression based on the quantification theory as shown in the following equation.

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc} \delta_{fc}(i) \qquad i = 1,2,3,\ldots,N \qquad (l)$$

where $N$ represents the total number of data,
$\hat{y}_i$ represents the predicted phoneme duration of the *i*-th sample,
$y$ represents the average duration of the phoneme,
$x_{fc}$ represents the regression coefficient, and
$\delta_{fc}($ represents the characteristic function [3] producing 0 or 1 corresponding to affected factors.

The predicted duration is exploited as the first criterion for speech unit selection in TTS. It serves later as a target value for waveform duration modification if the duration difference between the selected unit and the target value exceeds a threshold value. Duration modification is based on Time-Domain Pitch Synchronous OverLap-and-Add (TD-PSOLA) [4].

These trained models and wave-form modification technique are integrated into the baseline system resulting a new naturalness improved Vaja system.

## III. Experiments

Experiments are performed in two approaches, an objective and subjective test. Each test used different data sets and measurements as described in two following subsections. The last subsection shows experimental results.

### A. *Experimental Data*

The TTS system uses a Thai speech corpus named TSynC [6], containing 2,644 paragraph, 15,716 phrase breaks, 5,200 utterances, 5,089 unique words and 30,096 tonal syllable patterns, varied on 89 Thai phonemes. Every paragraph was manually tagged with sentence-end markers, word and phrase breaks, and word POSs, by linguists. The data were divided into five subsets and performed cross-validation. Each of five cross-validation experiments utilized a training set of 2,115 paragraphs and 12,639 phrase breaks in average and a test set of the remaining data. Average results of those five sets are reported.

### B. *Experimental Measurement*

Both objective (automatic) and subjective (listening) tests are performed to reflect the improvement of the system.

- *Objective evaluation of phrase breaking*

Missing or misplaced phrase breaks can distort the meaning of utterances, as well as the naturalness of synthetic speech. To measure the quality of phrase detection automatically, we compute

$$Precisio = \frac{t_B}{t_B + f_B} \qquad (2)$$

$$Reca = \frac{t_B}{t_B + f_{\neg B}} \qquad (3)$$

$$F - measure = 2 \frac{Precision\ Recall}{Precision + Recall} \qquad (4)$$

where $t_B$, $f_B$ and $f_{\neg B}$ denote correct acceptance, false acceptance, and false rejection rates of phrase break prediction.

- *Objective evaluation of duration modeling*

Fluctuation of phoneme duration in a syllable conveys a different meaning of the whole word in an utterance. Differences between the predicted duration and the targeted duration observed from an actual human speech are shown as an average duration error (in ms), which is calculated by the following equation.

$$\textit{Average duration error}\ \underset{)}{=} \frac{\sum_{i=1}^{N}\left|D_{pi}-D_{ti}\right|}{N}$$

where$D_{pi}$ is the predicted duration of the *i*-th phoneme,
$D_{ti}$ is the targeted duration of the *i*-th phoneme,
$N$ is the total number of phonemes in the test set.

- *Subjective evaluation*

A listening test using the MOS scale [5] is conducted. The values 1 to 5 in the MOS scale rank the quality of speech utterance from the worst to the best. Three test cases (S1, S2 and S3) were set up in an office room using 45 test utterances (3 cases * 15 sentences/case). The three cases are as follows.

- S1 contains 15 utterances of natural speech.
- S2 contains 15 utterances generated by the baseline TTS system.
- S3 contains 15 utterances generated by the prosody-modified TTS system.

The test is performed by prompting three versions of each of 15 sentences to a listener, starting by S1 followed by S2 and S3, which can be shuffled. The listener gives scores to each version based on the question "how natural is the utterance?" and steps to the next sentence.

The experiment is done by 20 subjects, who are NECTEC staffs aged between 22 – 38 years old (27.9 years in average). Seven of them are females and thirteen are males.

### C. Experimental Result

According to the objective and subjective tests, results are presented as follows.

- *Objective Test*

This subsection shows results averaged from the five cross-validation data sets. Phrase detection in the several research works revealed that the accuracy was varied by the different features. In real situation, some features such as *POS*, *NSyl*, and *NWrd* are not precise since the previous modules; word segmentation, POS tagging, and grapheme-to-phoneme (G2P) conversion are not perfect.

TABLE II
RESULTS OF FOUR PHRASE BREAKING MODELS.

| Model | Feature | Precision | Recall | F-measure |
|---|---|---|---|---|
| Baseline | *Rule based* | 33.2 | 34.6 | 33.9 |
| 1 | *POS* | 77.8 | 65.9 | 71.4 |
| 2 | *POS + NSyl* | 74.1 | 70.2 | 72.1 |
| 3 | *POS + NSyl + NWrd* | 73.8 | 70.8 | 72.3 |
| 4 | *POS + NSyl + NWrd + SylRatio + WrdRatio* | 74.2 | 69.4 | 71.7 |

Table II compares four models of phrase breaking with different input features. All features are base on perfect word segmentation, POS tagging, and G2P conversion.

The results show that using *POS*, *NSyl* and *NWrd* achieves the best performance. Adding more features, e.g. in the 4th model, is noisy and hence lowers the overall accuracy.

Our word segmentation with POS tagging yields 66.6% accuracy, whereas our G2P module achieves 87.2% accuracy when the input text is perfectly segmented into words. We also compute F-measure of the phrase break prediction model when working with such imperfect word segmentation and G2P conversion. Results are illustrated in Figure 2. In the case of using actual word segmentation and G2P conversion (white pieces), which is the real situation, results show that using *POS* and *NSyl* gives the highest accuracy. Interestingly, no matter what perfect or imperfect G2P is used with the ideal word segmentation, only little changes of accuracy are obtained. The reason is that both still produce the similar number of syllables, which is one of a key feature in the phrase breaking module.

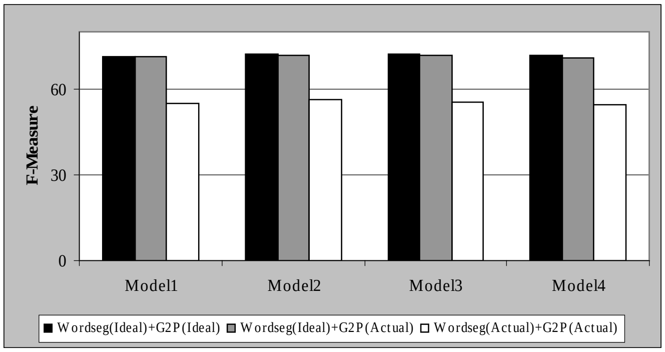

Figure 2. F-measure results of 4 models based on ideal and actual word segmentation and G2P modules.

The graph in Figure 2 obviously shows that the accuracy of phrase breaking can be significantly improved if a better word segmentation is obtained. Improving the word segmentation module can be done by a carefully treat of abbreviation, numbering, and special symbols in the input text.

In the same way, evaluations of duration modeling were also set up using five cross-validation training and test data sets. The average duration error achieves 21.4 ms and 22.9 ms given ideal and actual phrase breaks respectively. Comparing to the baseline TTS system where average duration errors are 24.9 and 25.5 ms for the ideal and actual phrase breaks, 3.5 and 2.6 ms error reductions are obtained. To express more detailed results, we count the number of phonemes, whose predicted durations lie in a ±*t* interval from their targeted durations. Figure 3 plots the relative counts with respect to the value of *t*.

We can observe that durations of over 80% of phoneme samples are well predicted, i.e. less than 20 ms average duration error. In Thai, the phoneme duration plays a critically important role on the vowel length. Short and long vowels in Thai carry different meanings and have been designed as separated units. Therefore, the system must be able to generate at least the distinguishable durations of these two vowel groups. Hansakunbuntheung [3] showed that short and long vowels differ at approximately 50 ms in average. We then set a threshold to modify the phoneme duration by TD-PSOLA if the duration difference between the unit selected from TSynC and its targeted value exceeds 40 ms.

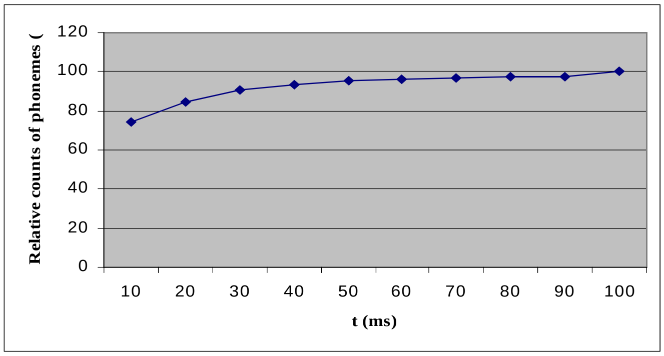

Figure 3. Relative counts of phonemes whose durations lie in ±*t* distance from their targeted values.

- *Subjective Test*

Table III reports MOS results of the listening test described in Section II.

TABLE III
MOS OF THREE TEST CASES

| Case | Source | Naturalness score |
|---|---|---|
| S1 | Natural speech | 4.85 |
| S2 | Synthetic speech from the baseline system | 2.14 |
| S3 | Synthetic speech from the prosody-modified system | 2.56 |

The results show that a 0.42 improvement of MOS is obtained by the prosody-modified TTS system comparing to the baseline system. A major reason of the improvement comes from better phoneme durations predicted by the intelligent duration model. It is noted that some subjects gave a lower score to some prosody-modified utterances, which contain unnatural false-breaks. Such false-breaks happen more in the proposed phrase-break detector, where breaks are determined at every word juncture; while the baseline system considers only at white-spaces.

## IV. Conclusion and Future works

In this paper, phrase-break and phoneme-duration prediction modeling were optimized the practical use in a Thai unit-selection TTS system. In phrase-break improvement, a new set of features was proposed as an input to CART. To predict appropriate phoneme durations, a linear regression model was developed and TD-PSOLA was applied to modify speech signals. These proposed models were evaluated by an objective test and gave 73.8% precision, 70.8% recall and 72.2% F-measure. Moreover, duration prediction gave 3.5 and 2.6 ms reduction of average duration errors for ideal and actual phrase-break detection. A subjective test reported that the prosody-modified TTS system achieved 0.42 MOS improvement of naturalness over the baseline system.

Although the implemented system reaches an acceptable quality, there are several issues of naturalness, which can improve for example, G2P conversion of hardly literate words such as named entities, word segmentation, and other prosody information includes tonal information etc.

## Acknowledgment

We would like to thank the members of HLT and IMG laboratory of NECTEC who spent their valuable time to join our subjective test.

## References

[1] P. Mittrapiyanurak, C. Hansakunbuntheung, V. Tesprasit and V. Sormlertlumvanich, “Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach”, *NECTEC Technical Journal Thailand*, vol. 2, No. 7, pp. 36-47, June 2000.

[2] C. Hansakunbuntheung, V. Tesprasit, R. Siricharoenchai, Y. Sagisaka, “Analysis and Modeling of Syllable Duration for Thai Speech Synthesis”, *Proc. European Conference on Speech Communication and Technology*, Geneva-Switzerland, pp.93-96 (2003-9).

[3] V. Tesprasit, P. Charoenpornsawat and V. Sornlertlamvanich, “Analysis and Modeling of Syllable Duration for Thai Speech Synthesis”, *Proc. European Conference on Speech Communication and Technology*, Geneva-Switzerland, pp.325-328 (2003-9).

[4] F. Charpentier and E. Moulines, “Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *European Conference on Speech Communication and Technology*, vol. I, pp. 013-019, 1989.

[5] M. Viswanathan, “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale”, *Computer, Speech and Language19*, pp. 55-83, 2005.

[6] C. Hansakunbuntheung, V. Tesprasit and V. Sornlertlamvanich, “Thai Tagged Speech Corpus for Speech Synthesis”, *The Oriental COCOSDA 2003*, pp. 97-104, 2003.

[7] V. Tesprasit, P. Charoenpornsawat and V. Sornlertlamvanich, “Learning Phrase Break Detection in Thai Text-to-Speech”, *Interspeech 2003*, pp. 113-116, 2003.

[8] C. Hansakunbutheung, A. Thangthai, C. Wutiwiwatchai and R. Siricharoenchai. “Learning Methods and Features for Corpus-Based Phrase Break Prediction on Thai”, *Interspeech 2005*, pp. 325-328, 2005.

[9] W. Aroonmanakun, N. Thubthong, P. Wattuya, B. Kasisopa, and S. Luksaneeyanawin, “Automatic Thai transcriptions of English words”, *Southeast Asian Linguistics Society Conference 14 (SEALS 14)*, 2004.

[10] J. Inrut, P. Yuanghirun, S. Paludkong, S. Nitsuwat, and P. Limmaneepraserth, “Thai word segmentation using combination of forward and backward longest matching techniques”, *International Symposium on Communications and Information Technology (ISCIT)*, pp. 37–40, 2001.

[11] T. Theeramunkong and S. Usanavasin, “Non-dictionary-based Thai word segmentation using decision trees”, *The First International Conference on Human Language Technology Research (HLT)*, pp. 1–5, 2000.

[12] P. Mittrapiyanurak and V. Sornlertlamvanich, “The automatic Thai sentence extraction”, *International Symposium on Natural Language Processing (SNLP)*, pp. 23–28, 2000.

[13] K. Sileverman, “The Structure and Processing of Fundamental Frequency Contours”, *Ph.D. Thesis, University of Cambridge,*1987.

[14] S. Meknavin and B. Kijsirikul, “Thai grapheme-to-phoneme conversion, Burnham, D. (Ed.)”, *Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing of Language and Speech*, NECTEC, 1997.

[15] P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, “Thai Grapheme-to-Phoneme using Probabilistic GLR Parser” In Proc. Eurospeech, volume 2, pp. 1057 – 1060, 2001.

[16] P. Charoenpornsawat and T. Schultz, “Example-Based Grapheme-to-Phoneme Conversion for Thai”, *Interspeech 2006*, pp. 1268-1271, 2006.