Automatic Syllable-Pattern Induction in Statistical Thai Text-to-Phone Transcription

Ausdang Thangthai, Chatchawarn Hansakunbuntheung, Rungkarn Siricharoenchai, and Chai Wutiwiwatchai

Speech Technology Section, Information Research and Development Division
National Electronics and Computer Technology Center (NECTEC), Thailand
{ausdang.thangthai, chatchawarnh, rungkarn, chai}@nectec.or.th

Abstract

This paper proposes a technique of automatic syllable-pattern induction in statistical Thai text-to-phone transcription. A general process of building a statistical text-to-phone transcription is to first define a set of rules describing syllable patterns, which is used for syllabification. Given an input text, the syllabification process generates all possible syllable sequences, which are then scored and selected using a statistical model. Updating the handcrafted rule set of syllable patterns is time-consuming and requires expert linguists. Instead of the manual process, automatic induction of new syllable patterns from a large raw text if proposed. The process that can deal with raw text is particularly needed for Thai as segmenting Thai text is a very tedious task. Experiments show that the proposed Thai text-to-phone transcription system after applying a large raw text for syllable-pattern induction achieves approximately 2% improvement. A comparison with other Thai text-to-phone transcription models and error analyses are also given in the paper.

Index Terms: speech synthesis, text-to-phone transcription, thai

1. Introduction

Text-to-speech synthesis is constructed of two main modules, text-to-phone transcription where an input text is converted to a phonetic string, and speech synthesis where speech is generated given the phonetic string. Speech synthesis generally utilizes language-independent techniques such as unit concatenation. However, text-to-phone transcription strongly depends on cultural-specific and linguistic-specific usage of written script. To obtain a precise phonetic string, text-to-phone transcription therefore needs to be carefully built for a given language.

For Thai, several text-to-phone transcription systems have been proposed. A dictionary-based approach [1] requires a large dictionary and is unable to solve out-of-vocabulary (OOV) words, i.e. words not appearing in the dictionary. Rule-based approaches aim to solve the problem by writing general rules to cover unseen words [2], [3]. However, the rule-based approach often fails for ambiguities of syllable segmentation. To overcome the problem, machine learning [4] and statistical models [5], [6] have been proposed. A problem of proposed statistical models is that pre-defined patterns of textual syllable are required. Writing rules of these patterns is non-trivial and time-consuming even by expert linguists. Furthermore, it is difficult to update the set of patterns.

Given an original set of rules describing syllable patterns, our proposed method provides an easy way to update the rule set by inducing new syllable-patterns automatically from a large non-segmented text. Since segmenting Thai text is not trivial, the process that uses non-segmented text is essentially needed. Possible syllable sequences with their phone transcriptions are produced in the first pass of our model and the best sequence is chosen statistically in the second pass. Syllable tones, which are specific for tonal languages including Thai, are determined in the final step.

The rest of this paper is organized as follows. We first describe Thai syllable structures and problems of determining word pronunciation. Section 3 describes a general model of statistical text-to-phone transcription as well as our proposed model. Section 4 shows experimental setup and results. Finally, Section 5 concludes our work.

2. Thai Transcription

2.1. Syllable pattern representation

Basic Thai textual syllables can be represented in the form of $\{C_i,V,C_f,T\}$, where C_i, V, C_f and T denotes an initial consonant, a vowel, a final consonant, and a tone respectively. Table 1 summarizes the number of Thai characters and phones according to each part of syllables. Thai is a tonal language where meaning of a syllable changes as the syllable tone changes. Four tone markers are used to indicate 5 Thai tones; middle, low, falling, high and rising. For example, the syllable "ช้าง" (elephant) has the {ช, า, ง,้} pattern, which is pronounced as /cʰáːŋ'/.

Table 1. *The number of Thai characters and phones.*

Type	Character	Phone
Initial consonant (C_i)	44	38
Vowel (V)	16	24
Final consonant (C_f)	37	9
Tone (T)	4	5

2.2. Problems of text-to-phone transcription

Problems of Thai text-to-phone transcription have been described in previous research [4], [5], and [6], which are briefly summarized as follows.

- **Ambiguity in character-sound mapping:** Some Thai characters can be pronounced differently depending on its function and context. For instance, the character "ร" as an

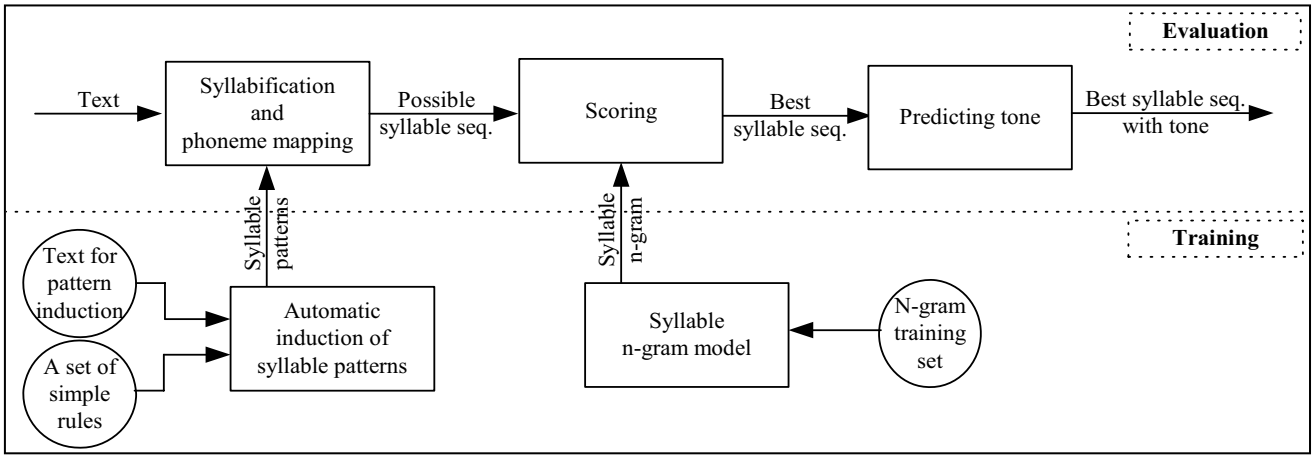

Figure 1 *Diagram of statistical approach for text-to-phone transcription.*

initial consonant is mapped to /r/ sound but as a final consonant is mapped to /n'/ sound. Similarly, the character "ท", which can be pronounced as /d/ or /tʰ/ in different words.

- **Implicit vowel:** In some cases, a vowel may not be represented in the textual form. For example, the syllable "ผม" (hair) /pʰŏm'/ consists of only two characters, an initial and a final consonant.
- **Vowel length:** The problem occurs when a long vowel is written but it is usually pronounced as a short sound. For example, the syllable "ท่าน" (you), which should be read as /tʰâ:n'/ (long vowel), but it is often pronounced as / tʰân'/ (short vowel).
- **Linking syllable:** Pronunciations of some words derived from Pali and Sanskrit often insert linking syllables in between two syllables. For example, "รัฐบาล" (government) is pronounced as /rát'/tʰà/bā:n'/ rather than /rát'/bā:n'/. The syllable /tʰà/ in the middle is a linking syllable not presented in the written form.
- **Function consonants:** A final consonant can be propagated to be the initial consonant of following syllable. For example, in the word "จัตุรัส" (square) /càt'/tù/ràt'/, the character "ต" functions as both a final consonant /t'/ of the first syllable and an initial consonant /t/ of the next syllable.
- **Character ordering:** In some cases, the order of characters is not corresponding to the pronunciation. For example, the word "เฉลียง" (veranda) /cʰà/lǐ:aŋ'/ has two syllables /cʰà/ and /lǐ:aŋ'/, but a part of vowel of the second syllable "เ" is placed in front of the first syllable "ฉ".

3. Statistical Text-to-Phone Transcription

3.1. General statistical model

As described in the introduction, the statistical approach is extensively conducted as it can solve ambiguities. A general diagram of statistical approach for text-to-phone transcription is shown in the top part of Figure 1. Mainly there are two processes; finding possible syllable sequences with their phone transcriptions and selecting the best sequence using a statistically trained model. The last block of tone prediction is specific for our Thai text-to-phone transcription system as syllable tones in Thai can be discovered easily by separated rules. The process of determining possible syllable sequences often utilizes a set of syllable patterns mainly written by hands. In previous research, handcrafted rules are written using context free grammar [5] or patterns of syllable [6].

3.2. Automatic induction of syllable patterns

One of the most problematic parts of the general statistical model described in the previous subsection is the difficulty of updating rules of syllable pattern. Syllable patterns might be updated easily if there exists a large text with textual syllables segmented. Segmenting the large text to sequences of syllables is a very tedious task that requires manual or semi-automatic effort. Therefore, this paper proposes a new method of syllable-pattern induction given non-segmented text. The proposed method is illustrated in Figure 2. The detail of procedure is as follows.

a) **Finding syllable patterns:** Given an input text, the first step is to find all possible syllable patterns. Syllable patterns are discovered using a set of simple rules, which can be written by non-experts. For example, the word "สามารถ" contains many possible syllables such as {ส, สา, สาม, สามา, สามาร, ม, มา, มาร, มารถ, …}. To deal with linking syllables, function consonants, and character ordering, we define that one textual syllable can be pronounced with one or two syllable sounds.

b) **Filtering for pronounceable syllables:** Only some syllable patterns produced in the first step are pronounceable. Those syllable patterns are included in the final syllable-pattern set. Pronounceable syllables are determined by passing each syllable pattern to a syllable-to-phone table [5]. The syllable pattern that does not meet any rule in the table is eliminated from the set.

c) **Storing syllable patterns:** Results from the previous step are all pronounceable syllable patterns in training text. Finally, each syllable sound is given a pattern index such as

/sǎ:m'/	ST0_0_0
/sǎ:/	ST0_0_1
/mā:/	ST0_0_2

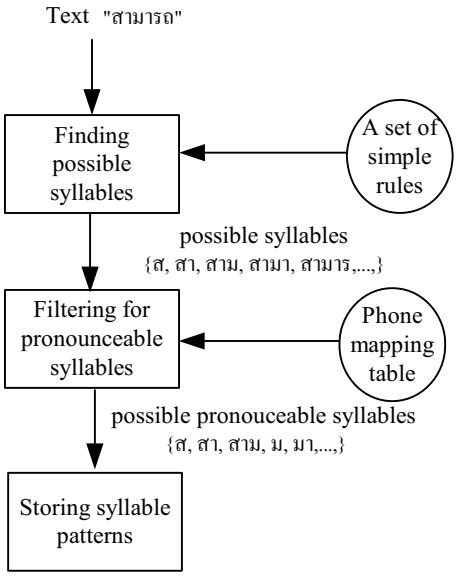

Figure 2 *Automatic induction of syllable patterns.*

The most benefit of this method is that, given a set of simple rules and a syllable-to-phone mapping table, it induces all possible syllable patterns from non-segmented training text. As shown in Figure 2, although the word "สามารถ" is actually pronounced as two syllables, {สา /sǎ:/, มารถ /mâ:t'/}, but the other possible syllable patterns such as {สาม /sǎ:m'/, มา /mā:/, มาร /mā:n'/, …} are also discovered and stored.

The final set of syllable patterns is used in the syllabification and phone-mapping process as shown in the top-left block of Figure 1. This process is similar to the procedure described above. The first step is to find all possible syllable sequences given an input text. Only sequences whose all syllables are pronounceable according to the phone-mapping table are stored. Finally the best sequence is determined using a statistical model described in the next subsection.

3.3. Syllable n-gram model

Given possible syllable sequences with their phone labels, a statistical model is conducted to score each sequence and select the best sequence. We use a well-known probabilistic n-gram model. The probabilistic n-gram model has been shown to be effective for Thai word segmentation [7] and other natural language processing problems. Our problem can be explained by the following equation.

$$\hat{S} = \arg\max_{S} P(S \mid G) = \arg\max_{S} \frac{P(G \mid S)P(S)}{P(G)} \qquad (1)$$

where G is a character string of the input text and $S = s_1 s_2 ... s_n$ is a sequence of syllable patterns. The syllable pattern s_i is actually the pattern index mentioned in the previous subsection. In the current implementation, $P(G|S)$ is assumed to be uniform for every syllable sequence and $P(G)$ is a fixed constant for every candidate. Hence, Equation 1 can be simplified as

$$\hat{S} = \arg\max_{S} P(S \mid G) = \arg\max_{S} P(S) \qquad (2)$$

$P(S)$ can be calculated by using the n-gram model of syllable patterns. Due to sparseness of training data, a bi-gram model is used in our work as presented in Equation 3. Good-Turing smoothing algorithm is conducted in the bi-gram model.

$$P(S) = \prod_{i} P(s_i \mid s_{i-1}) \qquad (3)$$

3.4. Predicting tone

In the final step, handcrafted rules are used to determined tones of syllables. The detail of rule can be found in [8]. There are some special cases where syllable tones do not correspond to the common rule set. Also most of loan words are pronounced with unusual tones. This problem will be investigated in the future work.

4. Experiment

4.1. Experimental setup

Two data sets are used in our experiments. The first set is a word list from Lexitron, a Thai open-source electronic dictionary (http://lexitron.nectec.or.th/), which contains 24,147 words. Each word is given its pronunciation in the form of phone sequences. We will denote the first set as "Lexitron". The second set is a large collection of Thai text from various sources including 1-year newspaper and a large number of Thai websites. The size of the second set is approximately 114.5 Mega-bytes containing around 24-million words and 137-tousand unique words. The second set is called "Webtext" hereafter.

Table 2. *Data sets used in experiments.*

Exp.	Syllable-pattern induction	Bi-gram training	Evaluation
A	4/5 of Lexitron	4/5 of Lexitron	1/5 of Lexitron
B	Webtext	4/5 of Lexitron	1/5 of Lexitron

In the training phase, a training set is used for inducing syllable patterns, whereas another training set is used to train syllable bi-gram. The former set needs only pure text without segmentation but the latter set requires phone sequences tagged to each text chunk. We perform two experiments with data sets shown in Table 2. The purpose of the experiment "B" is to show an improvement of our text-to-phone transcription when a large raw text is given, comparing to the baseline system in the experiment "A". In both experiments, the Lexitron set is divided to five subsets. Five cross-validation experiments, where a subset is used for evaluation and the rest are used for training, are carried out. An average result of five cross-validation experiments is reported.

4.2. Experimental results and discussion

Since the problem of vowel-length distortion in Thai often occurs, we present experimental results in two cases: the case considering vowel-length correction and the case ignoring vowel-length distortion. We will denote two cases as "ExactV" and "IgnoreV" respectively. Figure 3 shows transcription accuracies for different sets of n-best results. Transcription is considered to be correct if the correct phone-sequence is one of n-best sequences ranked by the n-gram model. Dash lines in

Figure 3 present results of the experiment A described in Table 2. 81.1% and 88.9% is achieved at the 1-best result of the ExactV and IgnoreV case.

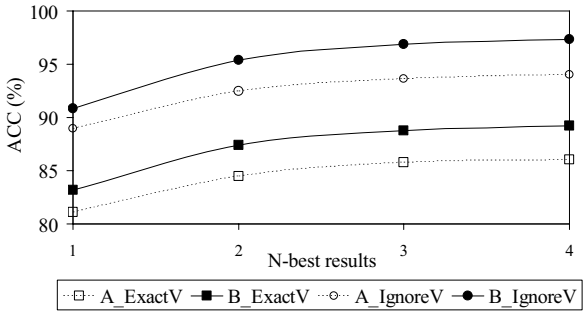

Figure 3 *Text-to-phone transcription accuracies.*

Table 3. *Comparison with previous systems.*

System	Accuracy (%)	
	ExactV	IgnoreV
Our model	83.0	90.9
PGLR	72.9	90.4
Decision tree	68.8	86.9

According to an error analysis, a significant problem is incorrect syllabification caused by unseen syllable-patterns. Adding new syllable patterns in previous works [5] and [6] was performed manually by expert linguists who understand the existing rule set. In contrast to our system, syllable patterns can be induced automatically given a large non-annotated text. Results after applying the large text for syllable-pattern induction are shown in the experiment B, solid lines in Figure 3. Approximately 2% improvement from the first experiment is achieved in both ExactV and IgnoreV cases. The number of syllable patterns discovered in the experiment A and B is 3304 and 3991 respectively. Table 3 presents a comparison between our best result and results obtained in the previous work using a PGLR parser [5] and a decision tree [4]. It is noted that two previous systems have been trained and evaluated using only the Lexitron set. Adding Webtext in our training set obviously enhances system accuracy without difficulty of text preparation.

There are still two important classes of errors. The first one comes from vowel-length distortion. Some Thai words are pronounced with unusual vowel length, i.e. the duration of vowel does not correspond to its grapheme. Consequently, transcribing these special words is often incorrect. These words should be treated specially in order to gain correct transcriptions.

The other important error comes from missing of linking-syllables in Pali-Sansakrit loan words. According to an analysis of linking-syllable errors, there are 61.1%, 21.5% and 7.2% of errors where the correct transcription is in the second, third, and forth rank. Up to 97.3% transcription accuracy can be obtained if the correct transcription is re-arranged to top. Therefore, a better scoring technique is required to solve this problem. Better scoring can be achieved by estimating $P(G|S)$ described in Equation 1 and, hence, Equation 2 is changed to

$$\hat{S} = \arg\max_{S} P(S \mid G) = \arg\max_{S} P(G \mid S)P(S) \qquad (4)$$

For the case of linking-syllables, P(“กาล”|/kā:/lá/) is likely to be higher that P(“กาล”|/kā:n’/), where the linking-syllable /lá/ in the former term is often missing. Incorporating the term $P(G|S)$ is therefore expected to help solving the case.

5. Conclusions

An efficient technique of syllable-pattern induction was proposed for a statistical Thai text-to-phone transcription system. The benefit of the proposed technique was that it required only a large raw text without any annotation. According to experiments, approximately 2% improvement was achieved after applying this technique given a large text. Our best configured system produced 83% accuracy, which was more than 10% higher than the previous system using PGLR.

Two major errors will be explored in the near future. First, vowel-length distortion will be solved by treating unusual words as special cases. Second, a pronunciation model, i.e. the probability that a character sequence represents a given syllable sequence, will be involved for better statistical scoring. The improved scoring algorithm will be used to solve especially the problem of missing linking syllables.

6. References

[1] Luksaneeyanawin, S., A Thai Text to Speech System. In Proceeding of the Conference of the Region Workshops on Computer Processing of Asian Language, pp. 305-315, Asian Institue of Technology, 1989.

[2] Khamya, A., Narupiyakul, L. and Sirinaovakul, B., SATTS : Syllable Analysis for Text-To-Speech System, In The 4th Symposium on Natural Language Processing (SNLP 2000), pp. 336-340, Chiangmai Plaza Hotel, Chiangmai, 2000.

[3] Narupiyakul, L, Khamya, A. and Sirinaovakul, B., The Phonetic Transcription of Thai Word. In Proceeding of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp. 789-792, Phuket Thailand, 1999.

[4] Chotimongkol, A. and Black, A. W., Statistically trained orthographic to sound Models for Thai, In Proceedings of ICSLP 2000, Beijing, China October, 2000.

[5] Tarsaku P., Sornlertlanvanich, V. and Thongprasirt, R., Thai Grapheme-to-Phoneme Using Probabilistic GLR Parsers. In Proceedings of Eurospeech 2001, Aalborg, Denmark, September, 2001.

[6] Aroonmanakun, W., and Rivepiboon, W., A Unified Model of Thai Word Segmentation and Romanization. In Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation, Dec 8-10, 2004, Tokyo, Japan, 2004.

[7] Meknavin, S., Charoenpronsawat, P., and Kijsirikul, B., Feature-based Thai Word Segmentation. In Proceeding of the National Language Processing Pacific Rim Symposiu, pp. 41-46, 1997.

[8] Thonglo, K., Thai Grammar (in Thai), Bangkok, Thailand, 1952.