Proceedings of the International Multiconference on Computer Science and Information Technology pp. 429–433
ISBN 978-83-60810-27-9
ISSN 1896-7094

LEXiTRON-Pro Editor: an Integrated Tool for developing Thai Pronunciation Dictionary

Supon Klaithin, Patcharika Chootrakool, Krit Kosawat
Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
112 Thailand Science Park, Phahon Yothin Rd., Pathumthani 12120, Thailand
Email: {supon.klaithin, patcharika.chootrakool, krit.kosawat}@nectec.or.th

***Abstract*—Pronunciation dictionary is a crucial part for both Text-To-Speech and Automatic Speech Recognition systems. In this paper, we propose a tool to easily create and edit Thai pronunciation dictionary, called LEXiTRON-Pro Editor. This tool integrates Thai word segmentation, Thai Grapheme-to-Phoneme (G2P) conversion, and database system with statistics. It automatically proposes a word's pronunciation to users by 1 of the 3 options in the successive order: the pronunciation from LEXiTRON-Pro database, the pronunciation combined from syllables with highest probability, and the pronunciation from Thai G2P. However, users can switch to another option or even directly input their own pronunciation with an easy interface editor. Our LEXiTRON-Pro database contains initially 105,129 unique words and 24,736 unique syllables with pronunciations. Compared to the previous version, our new program can reduce the process of dictionary development from 5 to only 1 step and the number of tools used by linguists from 3 to only 1. Moreover, our experiment shows that the time consumption and the number of ungenerable words are significantly reduced while the pronunciation accuracy is considerably improved.**

I. Introduction

PRONUNCIATION dictionary is an essential component for both Thai Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems. However, the procedure of developing a pronunciation dictionary is quite complicated and requires a lot of man-hours from linguists. To help them, many recent studies have focused on the Grapheme-to-Phoneme (G2P) conversion system to automatically generate phonemes of words and create easier the pronunciation dictionary. Several approaches have been proposed in the development of G2P to improve the phonetic transcriptions, such as Decision Tree, Statistical method, Pronunciation-by-analogy, and Rule-based approach. Some systems even combine various approaches together to increase efficiency. For example, Tarsaku *et al.* [1] had developed Probabilistic Generalized LR (PGLR) by combining Ruled-based and Decision Tree approaches which can achieve 72.87% of transcription accuracy. The Example-based Grapheme-to-Phoneme (EBG2P) conversion approach, developed by Paisarn Charoenpornsawat and Tanja Schultz [2], which generates pronunciations from syllables found in the training corpus, can reach 80.99% of transcription accuracy.

However, it is found that using only G2P to create the pronunciation dictionary has caused several problems. Firstly, Thai G2P cannot correctly deal with ambiguous strings that contain more than one possible segmentation. For example, "ตากลม" has two patterns of segmentation and accordingly two pronunciations, i.e. "ตาก|ลม|" (t-aa-k^-1|l-o-m^-0| : to be exposed to the wind) and "ตา|กลม|" (t-aa-z^-0|kl-o-m^-0| : round eyes). Unfortunately, Thai G2P always gives only one answer which is not always the right one. Secondly, words having more than one possible pronunciation could not be handled correctly. For example, "ประวัติศาสตร์" (history) can be pronounced "pr-a-z^-1|w-a-t^-1 t-i-z^-1|s-aa-t^-1|" or "pr-a-z^-1|w-a-t^-1|s-aa-t^-1|" but Thai G2P shows only one pronunciation. Lastly, Thai G2P could not correctly transcribe Thai Named Entities that are not typically found in the database, such as person names, acronyms, road names, points of interest, etc.

Given the above difficulties of Thai G2P, linguists or phoneticians are still indispensable in the development of the pronunciation dictionary. However, the previous version of Thai Pronunciation Dictionary, called LEXiTRON-Pro Version 1.0, required too many man-hours of linguists because there were several steps of manual tasks. Therefore, it is quite difficult and inconvenient to improve our LEXiTRON-Pro dictionary or create another one.

In this paper, we propose LEXiTRON-Pro Editor. It is an integrated tool to create a pronunciation dictionary more easily. This tool combines together Thai word segmentation, Thai G2P and Database with statistics, to enhance the accuracy of the phonetic transcription.

This paper is organized as follows. In the next section, we review basic concepts of the Thai language, the pronunciation dictionary and Thai word segmentation. In Section III, we summarize the previous version of LEXiTRON-Pro dictionary. Section IV explains how we reuse the old data. We propose our new application and explain the program interface in Sections V and VI. An experimental evaluation to compare our new program with the pre-

This work was supported by BEST Project under KET Platform of NECTEC.

vious system is done in Section VII and we conclude our paper in the last section.

II. Basic Concepts

A. Thai Language

The Thai language has 44 consonants and 24 vowels including 9 short vowels, 9 long vowels, and 6 diphthongs. Thai sound system can be derived in the format of /Ci-V-(Cf)-T/, where Ci denotes an initial consonant, V a vowel, Cf a final consonant which is optional, and T a tone [3].

Table I.
Thai consonant mapping to phonetic symbol

Consonant	Phoneme		Consonant	Phoneme	
	Initial (Ci)	Final (Cf)		Initial (Ci)	Final (Cf)
ก	k	k^	บ	b	p^
ข,ค,ฆ	kh	k^	ป	p	p^
ง	ng	ng^	ผ,พ,ภ	ph	p^
จ	c	t^	ร	r	n^
ฉ,ฌ,ช	ch	t^	ล,ฬ	l	n^
ซ,ศ,ษ,ส	s	t^	ว	w	w^
ญ,ย	j	j^	ห,ฮ	h	-
ฏ,ด	d	t^	ฝ,ฟ	f	p^
ฎ,ต	t	t^	ม	m	m^
ฐ,ฑ,ฒ,ถ,ท,ธ	th	t^	อ	z	-
ณ,น	n	n^			

Table II.
Thai vowel and mapped phoneme

Tongue Height	Tongue Advancement		
	Front (short, long)	Central (short, long)	Back (short, long)
Close	i, ii (อิ, อี)	v, vv (อึ, อือ)	u, uu (อุ, อู)
Mid	e, ee (เอะ, เอ)	q, qq (เออะ, เออ)	o, oo (โอะ, โอ)
Open	x, xx (แอะ, แอ)	a, aa (อะ, อา)	@, @@ (เอาะ, ออ)
Diphthongs	ia, iia (เอียะ, เอีย)	va, vva (เอือะ, เอือ)	ua, uua (อัวะ, อัว)

B. Pronunciation Dictionary

The pronunciation dictionary is a collection of words associated with their pronunciations in the form of phoneme sequences. Phonemes could be derived from standard sound representatives such as International Phonetic Alphabets (IPA) or Speech Assessment Methods Phonetic Alphabets (SAMPA). A letter-to-sound conversion (LTS) module takes the pronunciation dictionary as a primary source of knowledge to convert any textual word to its corresponding phoneme sequence. The LTS module plays an important role in building the phonetic transcription of speech corpus given speech orthographies. The output from the LTS tool is shown in Fig. 1. The first column is word list with syllable segmentation by "|" symbol and the second column is word's pronunciation. The pronunciation represents the word in the form of phoneme sequence with syllable segmentation and syllabic tone marked.

```
ไฮเปอร์นีเอา|      h-a-j^-0|p-qq-z^-0|n-ii-z^-0|z-aa-z^-0|
ดับเบิ้ลยูทีโอ|    d-a-p^-1|b-qq-n^-2|j-uu-z^-0|th-ii-z^-0|z-oo-z^-0|
เซโครไซแอทติค|   s-ee-z^-0|khr-oo-z^-0|s-a-j^-0|z-xx-t^-3|t-i-k^-1|
ฟาริงโกเปล้าสติ้|  f-aa-z^-0|r-i-ng^-0|k-oo-z^-0|pl-aa-t^-3|t-ii-z^-2|
บาเลนเซีย|        b-aa-z^-0|l-ee-n^-0|s-iia-z^-0|
อิคเทโรเยนนิค|    z-i-k^-1|th-ee-z^-0|r-oo-z^-0|j-ee-n^-0|n-i-k^-3|
นูเรอทติเซชั่น|   n-uu-z^-0|r-@@-t^-1|t-i-z^-1|s-ee-z^-0|ch-a-n^-2|
แบ็งคอก|          b-x-ng^-0|kh-@@-k^-1|
คอนทราเซปชั่น|    kh-@@-n^-0|thr-aa-z^-0|s-ee-p^-2|ch-a-n^-2|
พรอคเท็อปโตสิส|   phr-@@-k^-3|th-@-p^-3|t-oo-z^-0|s-i-s^-1|
```

Fig 1. LTS output in LEXiTRON-Pro format

C. Thai Word Segmentation

Since Thai has no word boundary, word segmentation is the first thing to do. At present, there are two main approaches on Thai word segmentation. The first approach is machine learning based (MLB) approach which is a technique that learns from a tagged corpus in which word boundaries are explicitly marked with special annotations. This algorithm creates statistical models based on the features of characters surrounding the boundaries (e.g., n-gram of a candidate word boundary). The other approach is dictionary-based (DCB) approach which is based on string parsing technique – i.e., input characters are scanned and matched with word set from dictionary [4].

LongLEXTO, the Thai word segmentation tool used in our application, was developed by Human Language Technology Laboratory (HLT). This tool is dictionary-based approach using the longest matching (LM) technique. The longest matching is a selection algorithm for solving the ambiguity problem by scanning the input text from left to right and then selecting the longest match with a word in dictionary [5]. The main dictionary is extracted from LEXiTRON, English-Thai online dictionary, containing 35,936 words.

III. Previous Version of LEXiTRON-Pro dictionary

In the previous version, most work was done by hand, which can be summarized as follows:

- Extract a word list from variety of text articles
- Compare the word list with the original pronunciation dictionary (if exist) and extract only new words
- Generate pronunciations of new words with letter-to-sound conversion (LTS)
- Manually check and correct the results of LTS
- Add new word entries to the pronunciation dictionary

105,129 words were extracted by linguists. There were quite a variety of sources ranging from general words, road names, person names, and abbreviations. All words were then converted to phoneme sequences using LTS conversion. All entries were verified by hand before importing to the LEXiTRON-Pro dictionary. This dictionary will be used as a primary resource in our LEXiTRON-Pro Editor.

IV. Data Preparation

Before using the new program, we had to prepare some more data. We analyzed all syllables with consistent pronunciation in LEXiTRON-Pro dictionary and found 364,707 syllable entries. After that, we counted the frequency of the same syllables and found 24,736 unique entries. Then, all unique entries were imported to the LEXiTRON-Pro database.

The probability of each syllable's pronunciation is calculated by the frequency of each syllable's pronunciation compared with the total frequency of all possible pronunciations for that syllable. We use the following equation:

$$(\%) = \frac{f}{\sum_{i=1}^{n} (fi)} \times 100$$

When % is the probability of syllable's pronunciation
f is the frequency of syllable's pronunciation
n is the number of possible pronunciations for the same syllable

For example, the term "เศรษฐกิจ" (/set-tha-kit/: economy) contains syllables "เศรษฐ" and "กิจ". The syllable "เศรษฐ" has 4 possible pronunciations: "s-ee-t^-1|", "s-ee-t^-1 th-a-z^-0|", "s-ee-t^-1 th-a-z^-1|", and "s-ee-t^-1 th-a-z^-3|" with the frequencies of 16, 4, 54, and 2, respectively. Their calculated probabilities are then 21.05%, 5.26%, 71.05%, and 2.63%, respectively. All values are recorded in LEXiTRON-Pro database to help users to decide which pronunciation they want to use.

V. LEXiTRON-Pro Editor

A. Input Format

LEXiTRON-Pro Editor accepts the input file in plain text with TIS-620 or UTF-8 encoding. This file can contain word lists or paragraphs and may contain foreign words, punctuations, or numbers, all of which will be ignored.

B. Pronunciation Generation

Normally, the pronunciation dictionary is manually created, checked, and rechecked by phoneticians or skilled linguists. However, this method is inconvenient and time-consuming [6], [7]. Therefore, we have developed LEXiTRON-Pro Editor that generate pronunciation automatically, based on database system with statistics, incorporated with Thai word segmentation tool and Thai G2P. The procedure to generate pronunciation of LEXiTRON-Pro Editor is shown in Fig. 2.

In the first step, the input file, which may contain multiple paragraphs, is read by LEXiTRON-Pro Editor and then segmented into words using LongLEXTO. After that, these words will be further segmented into syllables by Thai G2P. The results from G2P algorithm are syllable-segmented words with symbol "|" and their pronunciations that will be used in the modification process. Next, each word from the previous process will be compared to LEXiTRON-Pro database. There are, at this step, three possible scenarios:

- If that word matches a word in the database, then retrieve its pronunciation from the database and display the result.
- If not, then use its syllables from Thai G2P instead to compare again to LEXiTRON-Pro database, syllable by syllable. if all the syllables are found in the database but there may be more than one possible pronunciation, our system will choose the most probable one and then combine all syllables' pronunciations together before displaying the result which is supposed to be the pronunciation of that word.
- If not all of the syllables are found, then display the word's pronunciation from Thai G2P instead.

For the first case, the results will not be checked again because they come from the validated database. For the second and third cases, the results must be checked and corrected by linguists.

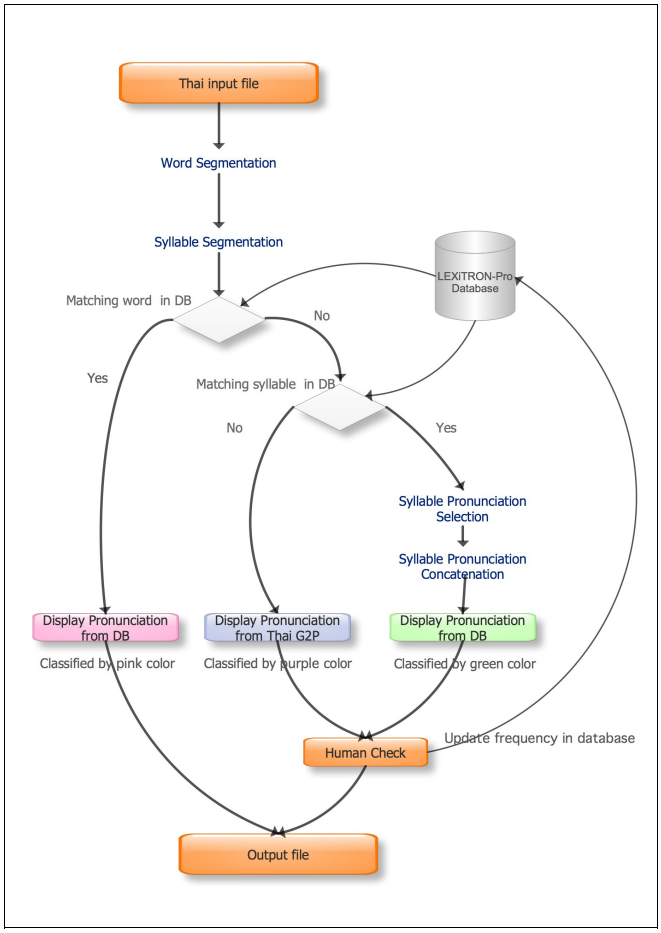

Fig 2. Work Flow in LEXiTRON-Pro Editor

C. Pronunciation Verification

Although the automatic generation of pronunciation is more convenient and can help linguists to reduce time, the results may contain some errors. Therefore, it is necessary to modify some results. We propose an easy interface to modify the pronunciation which will be explained in the section VI.B.

After verification, all data will be sent back to update the LEXiTRON-Pro database.

D. Output Format

The output file from LEXiTRON-Pro Editor has two options: word list format or paragraph format. The first format contains alphabetically ordered word list, divided into three columns: original words, words' syllables, and words' pronunciations. The other format contains alternate lines between word segmented text and words' pronunciation. The order of words' appearance is the same as the input file.

VI. Program Interface

LEXiTRON-Pro Editor has two major user-interfaces:

A. Main Interface

LEXiTRON-Pro Editor automatically generates pronunciations from files and displays results in three columns as shown in Fig. 3. Words in the left table are sorted according to the order of appearance in the input file. Each line contains ordinal number, original word, word's syllables, and word's pronunciation. When a line is selected, every syllable and pronunciation of that word will be displayed in the right table, one syllable per line. It is possible to edit each pronunciation by double-clicking on it to open another window, which will be explained in the next subsection.

The main interface includes two other useful features:

- Word Combination: since LongLEXTO may commit some errors by segmenting one word into two or more, it is necessary to combine them to make a single word by selecting them and clicking on this button.
- LongLEXTO's dictionary update: after combining words in the previous step, we can update the LongLEXTO's dictionary with this button. In addition, we can add a new word by ourselves to the LongLEXTO dictionary. The modification will take effect immediately.

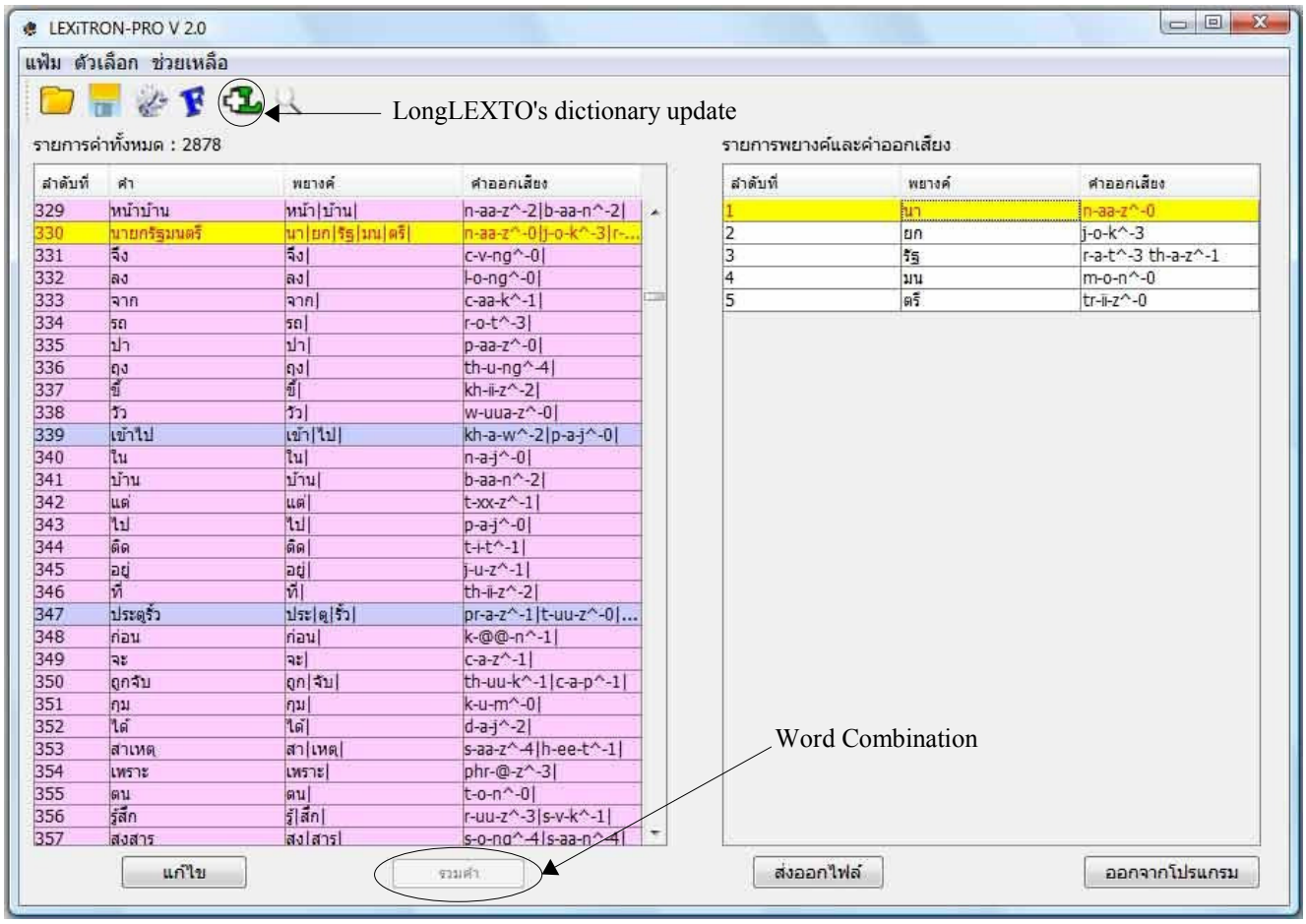

Fig 3. Main Interface

B. *Pronunciation Editor*

Since the automatic generation of pronunciation may be erroneous, especially when dealing with compound words, homographs, allophones and foreign words transliterated into Thai, we developed Pronunciation Editor for users to modify the syllable's pronunciation of the words with ease.

As shown in Fig. 4, users can edit any pronunciation by three methods:

- First, users can enter a new pronunciation by themselves in the top-left box.
- Second, users can choose the pronunciation proposed by Thai G2P in the bottom-right box.
- Last, users can choose one of the pronunciations from LEXiTRON-Pro database, as seen in the bottom-left box. If there are many possible pronunciations, they will be sorted by their calculated probability, as described in the section IV.

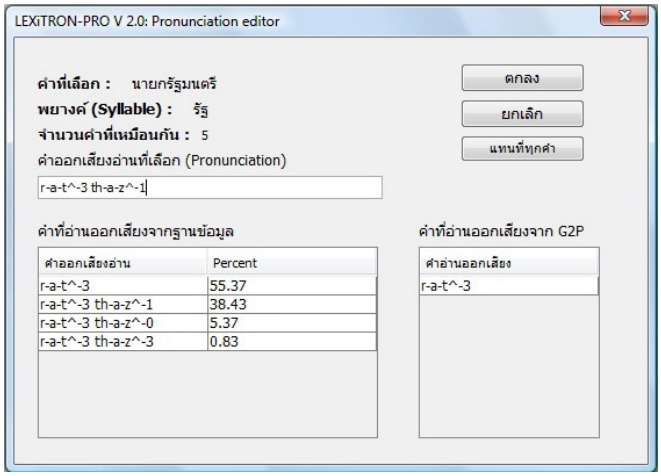

Fig 4. Pronunciation Editor

VII. Experimental evaluation

To evaluate the performance of LEXiTRON-Pro Editor compared with the previous system, we performed an experiment on both systems to generate a pronunciation dictionary with a test set of 1,072 Thai named entities, including names of universities/colleges, mosques, temples, hospitals, police stations, train stations, government offices, monuments, etc. The performance was evaluated in three aspects: generation time, pronunciation accuracy, and number of ungenerable words. The results are presented in the Table III.

Table III.
Systems' evaluation

Aspect	Previous system	LEXiTRON-Pro Editor
Generation time	5 min.	45 sec.
Pronunciation accuracy	18.1%	73.6%
Ungenerable word	60 words	0 word

The results show that LEXiTRON-Pro Editor can reduce time consumption from 5 minutes to approximately 45 seconds in pronunciation generating process, while the accuracy is largely improved from 18.1% to 73.6%. Lastly, the number of ungenerable words, found 60 words in the previous system, becomes zero.

VIII. Conclusion

The goal of this application is to assist linguists to reduce time and errors from manual work by simplifying several steps of development process, changing some tasks to automatic methods and proposing an easy interface to users.

Compared to the previous version, our new program can reduce the process of dictionary development from 5 to only 1 step and can reduce the number of tools used by linguists from 3 to only 1 program. In addition, LEXiTRON-Pro Editor can automatically propose a word's pronunciation to users by 1 of the 3 options in the successive order: pronunciation from LEXiTRON-Pro database, pronunciation combined from syllables with highest probability, and pronunciation from Thai G2P. However, users can switch to another option or even input directly the pronunciation they want by themselves with our easy interface editor.

Our experiment shows that, with LEXiTRON-Pro Editor, the time consumption and the number of ungenerable word are significantly reduced while the pronunciation accuracy is considerably improved as well.

We plan to use this program to develop LEXiTRON-Pro Dictionary Version 2.0 by increasing its size to more than 130,000 words in the near future.

Since Thai pronunciation dictionary is a crucial component in other speech processing applications such as Thai TTS and ASR, the more completed dictionary means an opportunity to the more successful speech applications too.

References

[1] P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, "Thai Grapheme-to-Phoneme using probabilistic GLR parser," in *Proceeding of EUROSPEECH 2001*, Aalborg, Denmark, 2001, pp. 1057–1060.

[2] P. Charoenpornsawat and T. Schultz, "Example-based Grapheme-to-Phoneme conversion for Thai," in *Proceeding of INTERSPEECH 2006–ICSLP*, Pittsburgh, PA, USA, 2006, pp. 1268–1271.

[3] P. Chootrakool, C. Wuttiwiwatchai, and K. Kosawat, "A large pronunciation dictionary for Thai speech processing," presented at the ASIALEX 2009, Bangkok, Thailand, August 20–22, 2009, Paper P013.

[4] C. Haruechaiyasak, S. Kongyoung, and M. N. Dailey, "A comparative study on Thai word segmentation approaches," in *Proceeding of ECTI-CON 2008*, Krabi, Thailand, 2008, pp. 125–128.

[5] Y. Poovarawan and W. Imarrom, "Dictionary-based Thai syllable separation," in *Proceeding of the 9th Annual Meeting on Electrical Engineering of the Thai Universities*, Khonkaen, Thailand, 1986.

[6] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary, continuous speech recognition," in *Proceeding of ICSLP 96*, Philadelphia, PA, USA, 1996, pp. 6–9.

[7] P. Pollák and V. Hanžl, "Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool," in *Proceeding of LREC 2002*, Las Palmas de Gran Canaria, Spain, 2002, pp. 1264–1269.