



A Learning Method for Thai Phonetization of English Words

Ausdang Thangthai, Chai Wutiwiwatchai, Anocha Ragchatjaroen, Sittipong Saychum

Human Language Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC), Thailand
{ausdang.tha, chai.wut, anocha.rak, sittipong.say}@nectec.or.th

Abstract

This article tackles the problem of transcribing English words using Thai phonological system. The problem exists in Thai, where modern writing often composes of English orthography, and transcribing using English phonology results unnatural. The proposed model is totally data-driven, starting by automatic grapheme-phoneme alignment, modeling transduction rules and predicting Thai syllabic-tones using learning machines. Three specific issues are addresses. The first one is involving English transcription information in transduction once the input English word appears in an English pronunciation dictionary. Second, more precise transduction rules can be obtained by a constraint of Thai syllable-structure. Lastly, the ambiguity in assigning tones to Thai pronunciations of English words is alleviated by introducing a learning machine. The proposed model achieves acceptable results in both objective and text-to-speech synthesis subjective tests.

1. Introduction

Grapheme-to-phoneme conversion (G2P), also called letter-to-sound conversion or phonetization, is an important function of text-to-speech synthesis (TTS). Its benefit also expands over other applications such as transliteration in machine translation. Normally, G2P is language-dependent and hence its difficulty relies on the complexity of particular language. Three approaches have been proposed, including dictionary-based, rule-based, and statistical approaches. At present, a hybrid model that composes of two or more of the mentioned approaches yields the best solution. For example, a G2P module firstly consults a pronunciation dictionary, which of course provides the most accurate results, and then executes a rule-based and/or statistical model once the input word is missing from the dictionary.

In Thai, several G2P algorithms have been investigated [1][2]. However, another interesting issue in Thai TTS rises by the modern style of Thai literature, which usually writes English words in Thai text. Figure 1 illustrates an example of Thai writing. Synthesizing speech of such English words can be straightforward by introducing a parallel system of Thai and English TTS. This solution, however, produces unnatural speech since users would not expect to hear an English accent mixed within Thai speech. A better solution is to build a G2P module that is able to produce transcriptions of English words on the basis of Thai phonological system.

Some research papers have proposed language-independent G2P algorithms [3][4][5]. Such algorithms learned to convert a source string to a targeted string given a training set of word-transcription pairs in any language. The algorithms are mainly based on some kinds of look-up tables such as a simple look-up table itself [3] and the Information gained tree (IGTREE) [4]. The table or tree determines a targeted phoneme string (or a null string) given a source character and its context. Ordering of conversion rules in the table or tree is very crucial since some characters can be

transformed in many ways. In Thai, Aroonmanakun [6] has applied a similar approach to transcribe English words using Thai phonemes. A training set consisted of English words and their transcriptions, which were manually aligned from grapheme characters to phonemes. The most outstanding characteristic of Thai sound is its explicit syllabic tone, even in reading of loan words. This task is not trivial since assigning a tone to a syllable depends on many factors such as the number of syllables in the word, syllable attributes, and the position of syllable in the word. No perfect rule has been given so far.

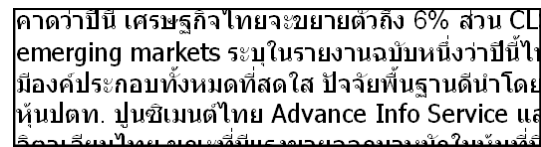


Figure 1. An example of English words mixed in Thai writing.

This article addresses the problem of Thai phonetization of English words and proposes another solution framework. We believe that the G2P engine of English has reached a nearly perfect result, at least comparing to other minor languages such as Thai. Furthermore, a lot of pronunciation dictionaries of English are available at present, for example the “CMUDICT” published by Carnegie Mellon University. For any English word existing in such dictionaries, using English transcriptions given by the dictionaries to assist in the English grapheme to Thai phoneme transduction problem is highly convincing. For unknown words, i.e. not in the dictionaries, we can apply the normal G2P model or conduct a two-step G2P process; finding an English transcription of the English word and converting the English transcription to a Thai transcription. Though the two-step process seems awkward, the transcription performance can be raised. This article applies the Dynamic time warping (DTW) algorithm in automatic English grapheme or phoneme to Thai phoneme alignment and introduces the Classification and regression tree (CART) for learning transduction rules. The same type of learning machine is also implemented for syllabic-tone prediction. The proposed model is evaluated both objectively and subjectively. For sake of explanation, the next section reviews transcription systems of both English and Thai. Section 3 then states the detail of the proposed framework and models. Experiments and conclusion are given respectively in Sect. 4 and 5.

2. Thai and English Phonology

There are currently a few phonological designs for Thai speech. The most fundamental is to separately define all single phonemes described in by the IPA. However, some empirical works showed the advantage of syllable-structure based design, where a syllable sound is represented by $/C_i V C_f^T/$. C_i and C_f denotes an initial and a final consonant, which can be either a single consonant or a consonant cluster. A vowel V includes either a single vowel or a diphthong and an

integer T varied from 0 to 4 indicates five Thai tones (Mid, Low, Falling, High, and Rising). Besides, some foreign phonemes are also defined in the modern Thai phonological system. All the defined phonemes and their computerized versions are shown in Table 1.

Table 1. A Thai phonological system.

Type		Symbol (IPA/Computerized)
Initial consonant	Single	p, t, c, k, ʔ/z, pʰ/ph, tʰ/th, cʰ/ch, kʰ/kh, h, b, d, m, n, ŋ/ ng, l, r, f, s, h, w, j
	Cluster	pr, pl, phr, pʰl/phl, tr, tʰr/thr, kr, kl, kw, kʰr/khr, kʰl/khl, kʰw/khw, fr, fl, br, bl, dr
Vowel	Single	i, i:/ii, i/v, i:/vv, u, u:/uu, e, e:/, e/q, e:/qq, a, a:/aa, e/x, e:/xx, o/@, o:/@@, o, o:/oo
	Diphthong	ia/ia, i:/a/ia, ia/va, i:/a/va, ua/ua, u:/a/ua
Final consonant		p/P, t/T, k/K, m/M, n/N, ŋ/NG, w/W, j/J, ʔ/Z, l/L, r/R, f/F, s/S
Tone		ē/0, è/1, ê/2, é/3, ǎ/4

There are unsurprisingly many designs of English transcription system such as TIMIT and CMU dictionary (CMUDICT). This article exploits for research the CMUDICT, which contains more than 129,000 English word entries with their transcriptions encoded by 39 phonemes (regardless of lexical stress variation). Table 2 illustrates a typical mapping between CMU and Thai phonemes.

Table 2. The definition of CMU phonemes and their Thai correspondences.

CMU	IPA	Thai	CMU	IPA	Thai	CMU	IPA	Thai
AA	ɑ	@	UH	u	u	N	n	n
AE	æ	x	UW	u	uu	NG	ŋ	ng
AH	ə	v	B	b	b	P	p	p, ph
AO	ɑ:	@@	CH	tʃ	ch	R	r	r
AW	au	aW	D	d	d	S	s	s
AY	ai	aJ	DH	ð	th	SH	ʃ	ch
EH	e	e	F	f	f	T	t	t, th
ER	ɜ	q	G	g	k	TH	θ	th
EY	e:	ee	HH	h	h	V	v	w
IH	i	i	JH	dʒ	c	W	w	w
IY	i:	ii	K	k	k, kh	Y	j	j
OW	o:	oo	L	l	l	Z	z	s
OY	ɔi	@J	M	m	m	ZH	ʒ	ch

To build a module for English grapheme to Thai phoneme transduction, an intuition is just to transform English transcriptions in the CMUDICT to Thai transcriptions. Transduction simply consults the transformed dictionary. Nevertheless, transforming English transcriptions to Thai transcriptions is yet not trivial. Let's see an example,

Word: "EXCLAMATION",
 CMU transcription: /EH K S K L AH M EY SH AH N/,
 Thai transcription: /z e K 3 kh l aa Z 0 m ee Z 0 ch a N 2/.

A major problem is that the CMU transcription provides no syllable boundary as sharing the final of a syllable with the initial of the successive syllable is common. When converting to Thai, syllables must be clearly separated. Initial and final consonants must be explicitly determined. Moreover, the glottal /z/ or /Z/ in Thai is not presented in the CMU transcription. Mapping the sound as described in the Table 2 is not always true, that is, an English phoneme can inherit

multiple Thai phonemes. For examples, an English sound /B/ can be either Thai /b/, /P/ or even /P b/. Assigning tones (0-4) to English words might be somehow systematic. The issue, however, remains unsolved.

3. Proposed Learning Models

As mentioned in the introduction, the simplest approach to G2P is to consult a pronunciation dictionary. This is also true for the English grapheme to Thai phoneme transduction task. According to a preliminary analysis on a 10-million word Thai text crawled from Thai websites, approximately 0.2% of the text was written by English alphabets. 2,137 words, out of 6,308 unique words in total, occurred at least twice. We found that 73.2% of 2,137 words appeared in the CMUDICT. Words missing from the dictionary were partly abbreviations or acronyms, proper nouns, and misspelling words.

By the vast coverage of the CMUDICT over English words appearing in Thai text, we propose a framework shown in Figure 2. Finding Thai transcriptions of an English word existing in the CMUDICT can be performed by using both the English grapheme and phoneme information. Otherwise, only the grapheme information is available. The tone prediction process is necessary in both cases.

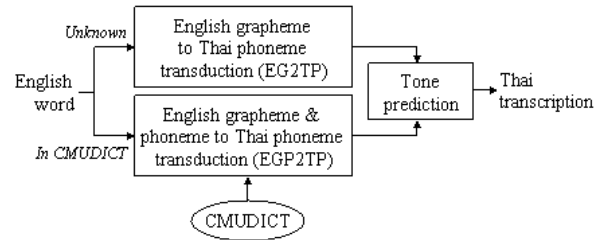


Figure 2. The proposed English grapheme to Thai phoneme transduction framework.

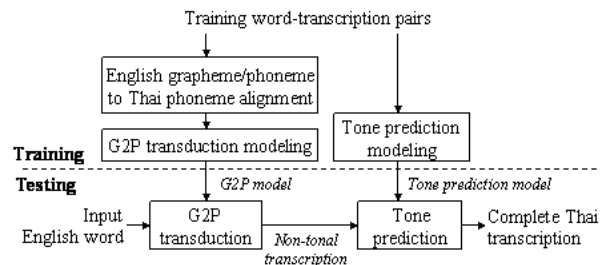


Figure 3. The overall structure of English grapheme to Thai phoneme transduction algorithm.

This article investigates on a G2P transduction algorithm based on learning machines. The overall model structure is illustrated in Figure 3. In the training step, graphemes and phonemes (if exist) of English words are aligned with corresponding Thai phonemes. Aligning results contain grapheme-phoneme mapping pairs, which are sources of features used to train the transduction model. Tone prediction modeling is also based on a learning machine over potential factors extracted from training data. In a real application, the learning machines are used to transform all English transcriptions in the CMUDICT to Thai transcriptions including predicted tones. The transformed dictionary is then applicable for the G2P part of Thai TTS systems. The learning machines are also embedded in the G2P module for transcribing any English words not found in the dictionary. The following subsections describe some important details of each sub-component.

3.1. Automatic grapheme/phoneme alignment

The first step in the training phase is to determine matching pairs of English grapheme (or phoneme) and Thai phoneme in a training word. In aligning, the DTW algorithm utilizes a cost function indicating the matching cost of an English grapheme (or phoneme) to a Thai phoneme. A basic idea to define the cost function is to give a cost of 0.0 to the closest pair such as from a character “k” to a phoneme /k/, 1.0 to the mapping of a consonantal character to its corresponding sound located as a final consonant, 2.0 to any consonantal character mapped as a part of consonant cluster or any vowel character mapped as a part of diphthong, and 5.0 to the mapping of any consonant to another unrelated consonant and any vowel to another unrelated vowel. The rest of mapping costs 10.0. In the case of English phoneme to Thai phoneme alignment, another cost function is constructed in a similar way. With these cost functions, DTW alignment can be performed effectively given an appropriate size of aligning window. Figure 4 shows an example of aligning result of the word “EXCLAMATION”.

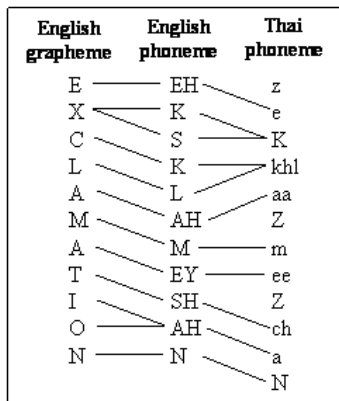


Figure 4. An example of automatic grapheme/phoneme alignment result.

3.2. Grapheme to phoneme transduction model

The aligning results indicate what Thai phonemes correspond to each English character in the input word. A learning machine then learns to convert each character to Thai phonemes using a feature set extracted from the aligning result. The feature set composes of the character itself and neighboring characters. In our model, the CART engine is chosen for the learning machine since it can deal with non-numerical inputs and is suited to sparse training data. Hereafter, we will denote this first model as “EG2TP” (English grapheme to Thai phoneme transduction).

A major problem found in the above transduction model is that the Thai phonological system relies on the syllable structure mentioned in Sect. 2. Based on the structure, C_i will only follow C_f or V , V can be placed behind C_i , and C_f only follow V . These constraints will greatly assist in transduction. We then try to incorporate such constraints into the model by feeding back the preceding predicted Thai phoneme to the current feature vector. Using a phoneme class, such as C_i , V , and C_f , instead of the exact phoneme in the fed-back feature can prevent the data sparseness problem.

Though the fed-back feature mentioned in the previous paragraph has been applied to the system, the transduction result might not comply with the Thai syllable structure. Therefore, a post-processing module is constructed to correct resulting phonemes one by one using heuristic rules, for instance, inserting a phoneme /Z/ if no C_f presents between V and C_i , inserting a phoneme /z/ between two consecutive V s,

and Henceforth, the improved model incorporating the fed-back feature and the post-processing module is called “EG2TP-Syl” (English grapheme to Thai phoneme transduction constrained by the Thai syllable structure).

As mentioned earlier, words existing in the CMUDICT can utilize their corresponding English transcriptions in the G2P process. Let’s consider a word “ABACUS” where English and Thai transcriptions are /AE B AH K AH S/ and /z a Z b aa Z k a S/ (tones are omitted) and a word “ABANDON” where English and Thai transcriptions are /AH B AE N D AH N/ and /z a Z b xx N d @ N/, the second /A/ character in both words best matches to different Thai phonemes /aa/ and /xx/. This ambiguity can be solved easily using the given English transcriptions where the character /A/ matches also to different English phonemes /AH/ and /AE/. Therefore, we enhance the model by introducing English transcription-based features to the learning machine if the input word exists in the CMUDICT. The features are simply the English phoneme aligned to the focused English character as well as the neighboring English phonemes. We call hereafter this last model “EGP2TP-Syl” (English grapheme and phoneme to Thai phoneme transduction constrained by the Thai syllable structure).

3.3. Tone prediction model

The last module functions to assign the most likely tone to each syllable of the word. Aroonmanakun [6] constructed a set of rules considering potential factors such as the number of syllables and the characteristic of syllables.

Expressing Thai tones of each syllable in English words has not yet systematically described. This article hence investigates on the use of a learning machine in tone prediction modeling. The learning machine is again the CART model. Effective features summarized in Table 3 are extracted from a training set and input to the CART model. It is noted that several features are Thai specific. More details can be found in many literatures such as http://en.wikipedia.org/wiki/Thai_language.

Table 3. Features used in syllabic-tone prediction.

Feature	Detail
C_i	- The initial consonant of the syllable
V	- The vowel of the syllable
C_f	- The final consonant of the syllable
C_i class	- Thai consonant group (High, Middle, or Low)
V class	- Short or Long vowel
No. of syllables	- The number of syllables in the word
Syllable type	- Dead (short vowel or stop final consonant) or Live (else) syllable
Syllable position	- The position of the syllable in the word (Begin, Middle, End)
Previous tone	- Predicted tone of the previous syllable

4. Experiments

Experimental data consists of 8,300 English words tagged with their transcriptions in both the CMU and Thai phonological systems. 10 cross-validation experiments are conducted, each using 90% of data for training and 10% for testing. Average results over the ten experiments are reported.

The first experiment aims to compare three models using different input features in the CART-based English grapheme to Thai phoneme system. The first model (EG2TP) employs only the grapheme information. The second model (EG2TP-Syl) incorporates the constraint of Thai phonology based on the syllable structure by introducing a fed-back feature

indicating the previous predicted Thai phoneme and a post-processing module to clean the transduction result. The last model (EGP2TP-Syl) utilizes in addition the English phoneme information. All modeling approaches have been explained in the Sect. 3. A transduction table-based model similar to the one proposed in [3] and [6] is also additionally evaluated for sake of comparison. Table 4(a) reports phoneme and word accuracies. According to the results, though constraining the transducer by the Thai syllable structure slightly reduces the phoneme accuracy, it strongly helps increasing the word accuracy by 4.8%. Further significant improvement of 6.8 and 11.0% phoneme and word accuracies is obtained by augmenting transduction input features by the English phoneme information. We observed that transduction errors came from two major reasons. Firstly, some English consonantal characters can be ambiguously transformed to a Thai final consonant, a Thai initial consonant, or both in sequence. Solving the problem requires much larger training data. Secondly, the English phonology has no explicit differentiation of short and long vowels. An English word, however, are read by Thai with a certain vowel length due to the explicit separation of short/long vowels in Thai. Determination of Thai vowel length in English words is based on several factors such as the lexical stress, the position of focused syllable in the word, and the number of syllables in the word. This issue has not yet included in the work reported in this article.

Table 4. English grapheme to Thai phoneme transduction and Syllabic-tone prediction results.

(a) G2P transduction		
Model	Phoneme accuracy (%)	Word accuracy (%)
Table	72.0	19.9
EG2TP	74.7	16.4
EG2TP-Syl	73.6	21.2
EGP2TP-Syl	80.4	32.2

(b) Tone prediction		
Model	Syllable accuracy (%)	Word accuracy (%)
Ideal inputs	88.4	71.7
Erroneous inputs	73.3	48.1

(c) Integration of G2P transduction and tone prediction		
Model	Syllable accuracy (%)	Word accuracy (%)
EG2TP-Syl +Tone	43.6	16.2
EGP2TP-Syl +Tone	53.3	24.5

The next experiment regards an evaluation of the tone prediction model based on the CART learning machine. Input features constitute potential factors for predicting syllabic tones as described in the Sect. 3.3. As effective features mostly depend on the Thai transcription produced by the G2P transducer, we will observe the prediction performance both for ideal inputs where perfect Thai transcriptions are provided, and for erroneous inputs where transcriptions are produced automatically by the EGP2TP-Syl transduction module. Resulting syllable and word accuracies of the tone prediction model are shown in Table 4(b). Although limited data are available, the results are somewhat promising. The most effective features are the exact phonemes in the syllable, i.e. the first three features in the Table 3. When combining the G2P and the tone prediction models, the entire G2P module yields 43.6 and 16.2% syllable and word accuracies when using only the English grapheme information, and increases by approximately 9% when the English phoneme information is additionally provided. Table 4(c) shows the exact results.

Finally, a subjective test is carried out. Forty English words, frequently written in Thai passages, with balance on the word length (one to five syllables) are selected. Speech synthesized from their correct transcriptions and automatically generated transcriptions are played back to seven subjects aged between 21 to 29. Three degrees of acceptability (Good, Fair, Poor) [7] is adopted in judgment. The test reveals that, in average, 28.6% are classified as Good, 33.6% as Fair, and 37.8% as Poor. This means that more than 62% of the test cases are acceptable.

5. Conclusions

This article proposed a novel approach to transcribe English words using a Thai phonological system. The problem is crucial for Thai where modern writing often mixes of Thai and English orthographies and transcribing such English words using English phonemes causes unnatural speech in TTS. The proposed method utilized two CART engines, one for G2P transduction and the other for syllabic-tone prediction. Significant enhancement of the model was obtained when the input English word was found in an English pronunciation dictionary. English transcriptions existing in the dictionary appeared to be very useful for the G2P process. Using Thai-specific phonological format based on the syllable structure additionally improved the transduction performance. Though the syllable and word accuracies produced by the proposed model were not high, the listening test showed the sufficiency of the proposed model to the task. The enhanced G2P model was then applied to construct a Thai-transcribed English dictionary which was then embedded in our Thai TTS engine.

To leverage the transduction performance, a specific treatment of converting consonants is required to alleviate the ambiguity in distinguishing Thai initial and final consonants. Lexical stress given in the English pronunciation dictionary is another source useful for more accurate prediction of vowel length and perhaps syllabic-tones.

6. References

- [1] Chotimongkol, A. and Black, A. W., Statistically trained orthographic to sound models for Thai, In Proc. of ICSLP 2000, Beijing, China October, 2000.
- [2] Tarsaku P., Sornlertlanvanich, V. and Thongprasirt, R., Thai grapheme-to-phoneme using probabilistic GLR parsers. In Proc. of Eurospeech 2001, Aalborg, Denmark, September, 2001.
- [3] van den Bosch, A., Daelemans, W., Data-oriented methods for grapheme-to-phoneme conversion. In Proc. of European Chapter of the ACL, Utrecht, April, 1993.
- [4] Daelemans, W., van den Bosch, A., Language-independent data-oriented grapheme-to-phoneme conversion. In van Santen et al. (Eds.), Progress in speech synthesis. New York, Springer, 1997.
- [5] Pagel, V., Lenzo, K., Black, A. W., Letter to sound rules for accented lexicon compression. In Proc. of ICSLP 1998, Sydney, Australia, 1998.
- [6] Aroonmanakun, W., Thapthong, N., Wattuya, P., Kasisopa, B., Luksaneeyanawin, S., Generating Thai transcriptions of English words. Presented at SEALS 14 Conf., May, 1994.
- [7] Coker, C., Church, K., Liberman, M., Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. In Proc. of the Conf. on Speech Synthesis, European Speech Communication Association, 1993.