Building WFST based Grapheme to Phoneme Conversion for Khmer

Kak Soky[†‡] Xugang Lu[†] Peng Shen[†] Hiroaki Kato[†]
Hisashi Kawai[†] Chuon Vanna[‡] Vichet Chea[‡]

[†]National Institute of Information and Communications Technology,
Kyoto, Japan
[‡]National Institute of Posts, Telecommunications and Information Communication Technology,
Phnom Penh, Cambodia
soky.kak@niptict.edu.kh

Abstract

Building pronunciation lexicon is one of the most important steps for building automatic speech recognition (ASR) systems and text-to-speech (TTS) systems. For a low-resource language, there is no or lack of such a dictionary (lexicon) cracked by experts. Even there is such a dictionary, the out-of-vocabulary (OOV) problem is inevitable. In this paper, we introduce our work on grapheme to phoneme (G2P) conversion for Khmer language based on weighted finite state transducer (WFST) technique. In building and improving G2P for Khmer, we trained a Khmer G2P model based on WFST with a manually transcribed pronunciation lexicon set. We tested the model on a set of new lexicons for pronunciation predictions. Our results showed that 12.98% in word error rate (WER) or 3.49% in phoneme error rate (PER) was obtained for a test set.

Keywords: Grapheme-to-Phoneme, G2P, ASR, WFST, TTS, Khmer

1 Introduction

This paper will show our recent research aimed at developing an automatic speech recognition (ASR) system and text-to-speech (TTS) system for Khmer language. Pronunciations are in the middle-layer between the acoustic model and the language model and the performance of the overall system relies on the coverage and quality of the pronunciation component [1]. Knowing how Khmer words are pronounced is an essential ingredient in both of these systems. Different from rich-resource languages, e.g., English, Chinese, and Japanese, there is no standard pronunciation dictionary for Khmer cracked by experts for our ASR purpose. Manually transcribing such a dictionary is time consuming and tedious. It is better to find a way to automatically transcribe new words to their pronunciations based on already transcribed words.

A Khmer G2P converts a Khmer word, as a series of characters or graphemes, to a pronunciation, as a series of phones. A pronunciation system of Khmer language typically comprises a static word-pronunciation dictionary which is needed and written by linguistics or may even be generated using a data-driven approach. However, a static list will never cover all Khmer words. Therefore, to generate pronunciations that can cover all the possible Khmer words is usually complemented with a Khmer G2P engine. In this study, we build a G2P for Khmer which will be used for ASR in our future study.

The remaining part of the paper is structured as follows. The overview of the characteristics of the Khmer language is presented in Section 2. Section 3 summarizes some of the related research in this area. Section 4 presents process of construction of Khmer G2P conversion tool. Section 5 shows the experiments and results for testing the G2P with Rule-based method and WFST based technique. Section 6 discusses about the way of using these methods with Khmer language. Section 7 concludes this paper.

2 Khmer Language

Khmer /kʰmaɛ/ or Cambodian (ភាសាខ្មែរ /pʰiɜsaːkʰmaɛ/, or more formally ខេមរភាសា /kʰaɛmaʔraʔpʰiɜsaː/) is the national language of the Khmer people and the official language of Cambodia. Around 90% of Cambodian populations speak this language in Cambodia and also some speakers live in Vietnam, Thailand, U.S.A,

France, Australia, and Canada [2; 3].

Khmer language (Cambodian) is one of the under-resourced Southeast Asian languages for natural language processing (NLP). It is a SVO (Subject, Verb and Object) language. Syntactically it is quite similar to Chinese and English, and also it is similar to Japanese, Chinese, and Myanmar in the word composition. Each Khmer word is composed of single or multiple syllables which are usually not separated by white spaces. Although spaces are used for separating phrases for easy reading, it is not strictly necessary. In addition, these spaces are rarely used in short sentences, and there is no exact rule if they are used.

There are three main word groups in modern Khmer: (1) original Khmer words, (2) Sanskrit and Pali which have been influenced by the royal and religious registers, through Hinduism and Buddhism, and (3) loanwords from French and English, i.e., many words were borrowed and have become a part of the colloquial language, as well as medical and technical terms. There is also a smattering of Chinese and neighbor countries' loanwords in colloquial speech.

Unlike Thai, Vietnamese, and Lao, Khmer is non-tonal and has a high percentage of disyllabic words which are derived from monosyllabic bases by prefixation, and suffixation [4].

2.1 Writing System

The Cambodian script (Khmer letters) has symbols of consonants, dependent vowels, independent vowels, and several diacritic symbols. Most consonants change into reduced or modified forms, called sub-consonants, when they occur as the second member of a consonant cluster. Dependent vowels may be written before, after, over, or under a consonant symbol.

A two-volume dictionary prepared under the direction of the Venerable Chuon Nath of the Buddhist Institute in Phnom Penh is the standard work on Khmer lexicography and now this dictionary has been developed to be an electronic dictionary [5].

2.2 Consonants

There are 33 symbols of letter in the Cambodian writing system. They are arranged in five groups according to the place of articulation, proceeding from the velar to the labial and the sixth group is labelled as miscellaneous [5]. They are separated into two groups or series of consonants according to their unmarked dependent vowels, i.e., /ɑː/ series and /ɔː/ series, and indicate in total 21 phonemes as shown in Table 1. Some of the phonemes have only one consonant symbol belongin to one of the two series. In order to fulfil the blank series and complete the Khmer consonantal inventory, two specialized diacritical symbols are used, ie., /◌̈/, to change to /ɑː/ series and /◌̈/, to change to /ɔː/ series [5; 6; 7].

Table 1. Consonant inventory of modern Khmer

No.	IPA	ɑː series		ɔː series	
		c	sub-c	c	sub-c
1	k	ក	◌្ក	គ	◌្គ
2	k^h	ខ	◌្ខ	ឃ	◌្ឃ
3	ŋ	ង់		ង	◌្ង
4	c	ច	◌្ច	ជ	◌្ជ
5	c^h	ឆ	◌្ឆ	ឈ	◌្ឈ
6	ɲ	ញ៉		ញ	◌្ញ
7	d	ដ	◌្ដ	ឌ	◌្ឌ
8	t^h	ឋ, ថ	◌្ឋ, ◌្ថ	ឍ, ធ	◌្ឍ, ◌្ធ
9	n	ណ	◌្ណ	ន	◌្ន
10	t	ត	◌្ត	ទ	◌្ទ
11	b	ប	◌្ប	ប៊	
12	p^h	ផ	◌្ផ	ភ	◌្ភ
13	p	ប៉		ព	◌្ព
14	m	ម៉		ម	◌្ម
15	j	យ៉		យ	◌្យ
16	r	រ៉		រ	◌្រ
17	l	ឡ	◌្ឡ	ល	◌្ល
18	ʋ	វ៉		វ	◌្វ
19	s	ស	◌្ស	ស៊	
20	h	ហ	◌្ហ	ហ៊	
21	ʔ	អ	◌្អ	អ៊	

2.3 Sub-Consonants

A sub-consonant always follows a consonant in the pronunciation. The form of a sub-consonant

is in most cases a smaller version of its full-size counterpart with several exceptions which look completely different from the full-size letters. The list is provided in Table 1 [5].

2.4 Dependent Vowels

A dependent vowel is indicated using one of 27 diacritical symbols or combinations of diacritical symbols [5]. These symbols and combinations of symbols may indicate monophthongs, diphthongs, or consonantal codas, e.g., /m/ or /h/. The pronunciation of a dependent vowel in Khmer is determined by the series of its initial consonant. In total, 36 dependent vowel phonemes are found in modern Khmer as shown in Table 2.

Table 2. Vowel inventory of modern Khmer

No.	Phone	IPA	No.	Phone	IPA
1	អ	ɑː	19	អើ	oː
2	អ់	ɑ	20	អឺ	ɜi
3	អា	aː	21	អូ	oʊ
4	អ៉	a	22	អឺ	ɜɵ
5	អិ	e	23	អេ	ɛe
6	អឹ	ɜ	24	អើ	aɜ
7	អុ	o	25	អឿ	ɛă
8	អឺ	ɔː	26	ឣែ	aɛ
9	អ៊	ɔ	27	ឣៃ	ai
10	អុិ	i	28	អោ	aɔ
11	អុី	iː	29	អៅ	aɨ
12	អុឹ	ɵ	30	អឿ	iɜ
13	អុឺ	ɨː	31	អ័រ	oa
14	អុិ	u	32	អ័រ់	oă
15	អូ	uː	33	ឣើ	ɵi
16	អើ	ɵː	34	អឿ	ɵɨ
17	អើ	eː	35	អូ	uɜ
18	ឣើ	ɛː	36	អឿ	ɨɜ

2.5 Independent Vowels

Independent vowels are known as complete vowels. They are used in a few words to represent certain combinations of the initial glottal stop /ʔ/ or a liquid (/l/, /r/) and a vowel.

3 Related Work for G2P

Grapheme-to-phoneme conversion is the term applied to the process of automatically generating pronunciation hypotheses given an input orthography. It is an important issue for both automatic speech recognition and speech synthesis, and also plays a role in Spoken Dialog System and NLP for many languages. According to these reasons, some research papers and research reports related to Khmer G2P have been published.

In [8] the authors basically used their own grapheme-to-phoneme lexicon. A rule-baed method was applied only for out-of-vocabulary words that were not included in the lexicon (about 72% of word accuracy was obtained). Many rules have been applied and also many exceptional cases have been used but the performance is still not well with the group change of sound and Pali / Sanskrit words. The 20 linguistic rules in [9] have been applied to detect monosyllable words and 400 rules are used to recognize the bisyllable words to generate Khmer pronunciation dictionary based on grapheme-to-phoneme correspondence. But Pali/ Sanskrit and some exceptional spellings can not be detected by these rules. Both [8] and [9] still can't solve the problems of generating the pronunciation of Pali/Sanskrit and group sound change words well.

Statistic-based methods can learn pronunciation rules automatically from a given training data set. They are preferred for the lack of language expert for manually designing the rules. Presently, many tools of statistic-based method have been published and some researches have been experimented on these tools. In [10], the authors compared G2P methods on the large pronunciation dictionary and Large Vocabulary Continuous Speech Recognition tasks. On the other hand, six methods have been evaluated on Myanmar pronunciation dictionary in [11]. Their results showed that WFST based G2P is one of the best choices.

WFST based G2P obtained state of the art performance. A modified WFST-based G2P algorithm has been introduced [12]. In the algorithm, multiple to multiple Expectation Maximization (EM) driven alignment algorithm was used for

G2P conversion. The algorithm successfully operate for English and Japanese. In our study, the algorithm was adapted to a Khmer G2P task.

4 Khmer G2P Conversion

To build WFST based Khmer G2P conversion model, three steps are carried out:

1. Preparation of lexicon
2. Manually building pronunciation dictionary for a training set
3. Learn the G2P model based on the WFST technique which is used for automatic generating pronunciations of new words.

4.1 Preparation of Lexicon

Manually transcribing an accurate pronunciation dictionary is the base of the WFST based G2P. There are three main resources of our lexicon dictionary (1) Chuon Nat Dictionary (`https://code.google.com/archive/p/khmer-dictionary-tools/downloads`), (2) lexicons from other websites, and (3) Basic Expressions Travel Corpus (BTEC) of National Institute of Information and Communications Technology (NICT)[13]. We used Khmer word segmentation tools [14] to split sentences or compound words to single words. Finally, we separated each single word as one line. In total, we got around 20K unique words. An example is given in Table 3 to show the process.

Table 3. An example of keyword generation

Original text	ខ្ញុំទៅសាលារៀន។
Segmented text	ខ្ញុំ ទៅ សាលា_រៀន ។
Unique lexicon	ខ្ញុំ ទៅ សាលា រៀន

4.2 Manually building pronunciation dictionary

This step is tough and time consuming. To reduce the work load we have done several steps:

Step 1. Choosing lexicon: We randomly shuffled and selected 50% of our existed keywords.

Step 2. Mapping International Phonetic Alphabet (IPA) [15]: IPA symbols are the special characters. They are not easy to type in any application. To solve this problem, we have mapped each IPA symbol of Khmer phoneme to Latin alphabet which is explained in Table 4.

Table 4. Mapping IPA to Latin alphabet

IPA-Latin		IPA-Latin		IPA-Latin	
k	k	t^h	th	j	j
k^h	kh	n	n	r	r
ŋ	ng	t	t	l	l
c	c	b	b	β̞	v
c^h	ch	p^h	ph	s	s
ɲ	nh	p	p	h	h
d	d	m	m	ʔ	?
ɑː	or	ɨː	eu	ɛă	eak
ɑ	ok	u	u	aɛ	ae
aː	a	uː	uː	ai	ai
a	ak	ɘː	uer	aɔ	ao
e	e	eː	eː	aɨ	au
ɜ	oe	ɛː	@	iɜ	ie
o	o	oː	oː	oa	oa
ɔː	ur	ɜi	ei	oă	oar
ɔ	uok	oʊ	ou	ɘi	ey
i	i	ɜɘ	eo	ɘɨ	av
iː	iː	ɛe	ee	uɜ	uo
ɘ	ue	aɜ	oeu	ɨɜ	oeur

Step 3. Setting rule: The actual pronunciation of a Khmer consonant or vowel has context-dependent variations, some of which can be described by simple rules [16].

1. Final Consonant: The pronunciation of a consonant may differ from its normal pronunciation at the syllable-final position as shown in Table 5.

   Examples:
   ចាស ⇒ c aː h
   ប្រយ៊ុទ្ធ ⇒ b r ɑː j u t

2. No vowel: As some of Khmer words occur without using any vowel symbol, either of the two unmarked vowels /ɑː/ or /ɔː/ have

Table 5. Final Consonants and pronunciation

No	Script	Pronunciation
1	ក ខ គ ឃ /kɑː/ /k^hɑː/ /kɔː/ /k^hɔː/	/k/
2	ច ជ ឆ ឈ /cɑː/ /c^hɑː/ /cɔː/ /c^hɔː/	/c/
3	ដ ឋ ឌ ឍ /dɑː/ /t^hɑː/ /dɔː/ /t^hɔː/ ត ថ ទ ធ /tɑː/ /t^hɑː/ /tɔː/ /t^hɔː/	/t/
4	ប ផ ព ភ /bɑː/ /p^hɑː/ /pɔː/ /p^hɔː/	/p/
5	ស /sɑː/	/h/
6	Silence	/ʔ/

to be assigned according to the series of the consonant in Table 1:
ថ ឃ ⇒ t^h ɑː j

3. The sign /◌̍/ over the consonant indicates that the consonant is not pronounced.
ស្រែរ៉ ⇒ s aɨ

4. Final consonants /k/ and /ŋ/ have to change to /c/ and /ɲ/ after the following vowels /i iː e ɛe aɛ/.
រីក ⇒ r iː c

5. Final consonant /k/ is pronounced /ʔ/ after /ɑ a ɔ ɛă/
ចាក់ ⇒ c a ʔ

6. រ /r/ is not pronounced in the word final position.
ការ ⇒ k aː
In some cases, it is pronounced as /l/:
សំបុរ ⇒ s ɑ m b o l

7. The pronunciation of vowels in a multi-syllable word changes if the initial consonants of the first and second syllables belong to different series.

C1 + C2 → C1 + C1 if C1 is stronger than C2.
សាលា ⇒ s aː l iə⇒ s aː l aː

Table 6. Strength scale of consonants

Stronger	p t c k p^h t^h c^h k^h b d ʔ s m n ŋ ɲ
Weaker	l r β̞ h j

8. Consonant clusters: in Khmer words, the cluster's second consonant is written with a symbol usually called "consonant foot" or "sub-consonant".

C1C2 → C1, if C1 is stronger than C2.
ស្នាន ⇒ s m iə n ⇒ s m aː n

4.3 Building G2P model based on WFST technique

The open-source Phonetisaurus, WFST-based Grapheme-to-Phoneme tools are used in our study to build the G2P model. Phonetisaurus is a WFST-driven G2P converter [12], (`https://github.com/AdolfVonKleist/Phonetisaurus`). The EM based many-to-many alignment process on grapheme and phoneme sequences of training data was done prior to build a G2P model. In the updated version of Phonetisaurus, dictionary alignment is done with OpenFst (`https://github.com/AdolfVonKleist/Phonetisaurus/tree/openfst-1.5.3`).
In order to estimate n-gram language model in model building, any language model toolkit such MITLM (`https://github.com/mitlm/mitlm`) or SRILM (`http://www.speech.sri.com/projects/srilm/download.html`) can be used. In our case, We used MITLM toolkit for this work.

5 Experiments and Results

In this paper, we report G2P performance on Khmer language pronunciation dictionary by comparing the two experimental results from Rule-based [8] and WFST based techniques. We evaluate performance using phoneme error rate (PER) and word error rate (WER) metrics. PER is defined as the number of insertions, deletions, and substitutions divided by the number of true phonemes, while WER is the number of word errors divided by the total number of words. The Khmer pronunciation dictionary is composed of

32740 words. We randomly split it into 10 sets (3274 words for each). Nine sets are picked up for training the G2P model, and the left one set is used for testing. By this cross set selection, we have ten rounds of experiments.

5.1 Rule-based G2P

To evaluate the performance of a Rule-based G2P for Khmer language, we first has defined three main steps according to [8]: (1) setting syllable boundary of each keyword, (2) cleaning silence or non-pronounce symbols, and (3) converting each syllable to the corresponding phonemes. We have modified script of [8] by adding our rules in Subsection 4.2 in step 3 to generate the phoneme of each keyword and then compared with the same reference as the WFST based techniques. The results based on the rule-based G2P model are shown in Tables 7 and 8. The average WER is about 39.59%. There are some errors occur by using this method: (1) syllable boundary of preprocessing technique is not yet good enough. In Khmer language, the pronunciation will be different if we change the syllable boundary. For example: កកររឿយ can be segmented into កក|ររឿយ (k ɑː k|r ɔː l iɜ j) or កករ|ឿយ (k ɑː k ɑː |l iɜ j). (2) many words are pronounced out of the rule: អារុធ (ʔ aː β o t) but the correct one is (ʔ aː β u t).

Table 7. Performance (WER) of rule-based G2P

No.	Words	
	Accuracy (%)	Error (%)
1	62.20%	39.80%
2	59.90%	40.10%
3	61.85%	38.15%
4	61.09%	38.91%
5	58.80%	41.20%
6	60.87%	39.13%
7	59.90%	40.10%
8	60.26%	39.74%
9	61.70%	38.70%
10	59.90%	40.10%
Total	60.41%	39.59%

5.2 WFST based G2P

In WFST based G2P, nine selected sets with well transcribed pronunciations were used for training (the test set is not included in training the G2P model). The results are shown in Tables 9 and 10. From these tables, we can see that 12.98% in WER or 3.49% in PER was obtained.

Table 8. Performance (PER) of Rule-based G2P

No.	Phonemes	
	Accuracy (%)	Error (%)
1	85.42%	14.58%
2	85.58%	14.42%
3	87.10%	12.90%
4	86.75%	13.25%
5	85.68%	14.32%
6	85.92%	14.08%
7	85.88%	14.12%
8	86.48%	13.52%
9	86.51%	13.49%
10	85.71%	14.29%
Total	86.10%	13.90%

Table 9. Performance (WER) of WFST based G2P

No.	Words	
	Accuracy (%)	Error (%)
1	86.44%	13.56%
2	87.05%	12.95%
3	86.99%	13.01%
4	87.81%	12.19%
5	86.56%	13.44%
6	88.36%	11.64%
7	86.56%	13.44%
8	86.87%	13.13%
9	87.75%	12.25%
10	85.83%	14.17%
Total	87.02%	12.98%

6 Discussion

As we showed in Subsections 5.1 and 5.2, the performance of the rule-based G2P is much worse than the WFST based G2P technique. Although a rule-based method does not need any training data, the complexity of the rules grows substantially as the system takes into account irregular spellings or pronunciations. Furthermore, if we use rule-based method with no space or no word boundary of each word of language like Khmer, we will have a problem of word

Table 10. Performance (PER) of WFST based G2P

No.	Phonemes	
	Accuracy (%)	Error (%)
1	96.50%	3.50%
2	96.31%	3.69%
3	96.43%	3.57%
4	96.77%	3.23%
5	96.54%	3.46%
6	96.80%	3.20%
7	96.36%	3.64%
8	96.37%	3.63%
9	96.74%	3.26%
10	96.28%	3.72%
Total	96.51%	3.49%

boundary. In order to improve accuracy with this method, accurate syllable segmentation is needed. In addition, more rules should be added in the G2P conversion model. The WFST based G2P technique can provide a better result of Khmer G2P conversion with adding new words according to the training dataset. To obtain a well generalized model, a large training data set is required to learn statistic rules for new keywords.

7 Conclusion

Building a G2P engine to generate the grapheme to phoneme of Khmer language is very important in Khmer language processing. In our study, we constructed a WFST based G2P model for automatic G2P conversion. We tried to prepare unique words as our lexicon dictionary to cover the large scale of any Khmer words. However, the pronunciation of some compound words are different from the combination of unique words. Therefore, some more compound words have been added to our lexicon to improve pronunciation accuracy.

In our future work, the combination of rule-based and statistic based methods will be investigated. In addition, WFST based syllable alignment will also be tested to explore the best way to improve the Khmer G2P conversion tool.

Acknowledgement

We would like to express our gratitude to Mr. Chhan Kimsoeun, Research and Innovation Center, National Institute of Posts, Telecommunications and Information Communication Technology, Cambodia for his help providing some useful related documents for this work and we are also grateful to colleagues in NICT for their suggestions, discussions, and support in building the G2P conversion tool for Khmer language.

References

[1] Kanishka Rao, Fuchun Peng, Hasim Sak, and Francoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neurl networks. Translation and Compiling, 2012.

[2] PAN. Research report on phonetic and phonological analysis of khmer. 2009.

[3] Jonh Haiman. A Cambodian (Khmer) Grammar. Jonh Benjamins B.V, 2011.

[4] Franklin E. Huffman, Chhom Rak, Thong Lambert, and Im Proum. Cambodian System of Writing and Beginning Reader. Yale University Press, 1970.

[5] Chuon Nat. Khmer Electronic Dictionary (Grammar Part). Buddhist Institute Dictionary, 1967.

[6] Ministry of Education Youth and Sport. Grammar for Grade 5. MoEY, 1983.

[7] Royal University of Phnom Penh. Phonetics and Linguistics. Royal University of Phnom Penh, 2014.

[8] T.R. Annanda, S.M. Long, S. Heng, N. Long, and K.H. Sok. Complexity of letter to sound conversion (lts) in khmer language: under the context of khmer text-to-speech (tts). International Workshop on Spoken Languages Technologies for Under-Ressourced Languages, 2009.

[9] S. Seng, S. Sam, V.-B. Le, B. Bigi, and L. Besacier. Which units for acoustic and language modeling for khmer automatic speech recognition? International Workshop on Spoken Languages Technologies for Under-Ressourced Languages, 2008.

[10] Stefan Hahn, Paul Vozila, and Maximilian Bisani. Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and lvcsr tasks. Interspeech, 2012.

[11] Ye Kyaw Thu and Win Pa Pa. Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary. WSSANLP, 2016.

[12] Josef R. Novak, Paul R. Dixon, Nobuaki Minematsu, Keikichi Hirose, Chiori Hori, and Hideki Kashioka. Improving wfst-based g2p conversion with alignment constraints and rnnlm n-best rescoring. Interspeech, 2012.

[13] Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. EUROSPEECH, 2003.

[14] Vichet Chea, Ye Kyaw Thu, Chenchen Ding, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Khmer word segmentation using conditional random fields. KNLP, 2015.

[15] International Phonetic Association. International phonetic alphabet (ipa).

[16] Jean Michel Filippi, Hiep Chan Vicheth, Srin Sereyrat, Chan Somnoble, Kit Calineat, Chhun Kun Bopha, and Mao Bonna. EVERYDAY KHMER. Funan, 2004.