Sinhala G2P Conversion for Speech Processing

Thilini Nadungodage[1], *Chamila Liyanage*[2], *Amathri Hansinie Perera*[3], *Randil Pushpananda*[4], *Ruvan Weerasinghe*[5]

Language Technology Research Laboratory, University of Colombo School of Computing, Colombo, Sri Lanka

{hnd[1],cml[2],rpn[4],arw[5]}@ucsc.cmb.ac.lk, amathriperera@gmail.com[3]

Abstract

Grapheme-to-phoneme (G2P) conversion plays an important role in speech processing applications and other fields of computational linguistics. Sinhala must have a grapheme-to-phoneme conversion for speech processing because Sinhala writing system does not always reflect its actual pronunciations. This paper describes a rule basedG2P conversion method to convert Sinhala text strings into phonemic representations. We use a previously defined rule set and enhance it to get a more accurate G2P conversion. The performance of our rule-based system shows that the rule-based sound patterns are effective on Sinhala G2P conversion.

Index Terms: G2P Conversion, Speech Processing, Phonology of Sinhala, Sinhala Language

1. Introduction

In the past few decades, automatic speech processing has become a very popular concept for most of the languages in world. Speech processing tasks such as Automatic Speech Recognition (ASR), Text To Speech (TTS) conversion, etc. have shown huge success rates for many languages, especially for English and other European languages which are rich with resources. However, there are many under-resourced languages which are new to the fields of speech processing and are in need of creating relevant resources. Grapheme to Phoneme (G2P) conversion is an essential part in automatic speech processing and creating a G2P conversion scheme is a major task. Sinhala is one of the under-resourced languages and in this paper, we describe the process of G2P conversion and G2P scheme creation for Sinhala language.

2. Related Work

G2P conversion approaches can roughly be divided into two categories as Data driven [1] approaches and Rule based approaches [2] The data driven approach requires a readily available data set such as a pronunciation dictionary or a corpus of transcribed words for looking up required inputs. In [3], the authors present two instances of data driven G2P approach for Dutch language. Three approaches using Artificial Neural Networks, Decision Trees and an Information Gain Tree, in converting G2P for English is presented in [4]. The use of Long Short-Term Memory Recurrent Neural Network for G2P conversion is presented in [5]. A knowledge-based G2P conversion system for Swedish is presented in [6]. While the data driven approach is more popular and can be language independent, it requires a huge amount of pre-defined data which is difficult to acquire in under-resourced languages.

On the contrary, the rule based approach only requires a smaller set of rules that describe how to convert a grapheme into one or several phonemes. Although it requires a through professional knowledge of the given language, this approach is much more suitable for under-resources languages. [7], [8] and [9] respectively presents Rule-based G2P conversion systems for Malayalam, Greek and Korean languages. There is also a previously built G2P conversion system for Sinhala language using the Rule based approach [10]. We have used most of the rules defined by them and there were some rules that needed to be enhanced to get more accurate output. In this paper we provide a review on the previously defined rule set and points on how they should be enhanced.

3. Sinhala Writing System

As most Brahmi-derived scripts, Sinhala is an alphasyllabary writing system and written from left to right. Sinhala script is used to write Sinhala language, which is one of the official languages of Sri Lanka, spoken by 74% of its population. In addition, Sinhala language is spoken by the Sinhala Diaspora communities in Middle East (Saudi Arabia, Kuwait, Qatar, and UAE), Britain, USA, Australia and Canada.

3.1. Sinhala Phonology and G2P Distinctions

Although we can define phonemic symbols for Sinhala consonants and vowels, G2P conversion for Sinhala cannot be performed as one to one mapping of grapheme to phoneme. In Sinhala, each consonant grapheme is associated with an inherent vowel either /a/ or /ə/ - schwa. All the vowels except the inherent vowel have separate symbols which are occurred after consonants. Absence of the inherent vowel is marked by adding *hal kirima* (remover of the inherent vowel) to the consonant; thus 'ක' /ka/ but 'ක්' /k/, and 'ව' /va/ but 'ව්' /v/.

Table 1: *Sinhala vowel classification* [10]

	Front		Central		Back	
	Short	Long	Short	Long	Short	Long
High	i	i:			u	u:
Mid	e	e:	ə	ə:	o	o:
Low	æ	æ:	a	a:		

There are 60 characters in the Sinhala alphabet including 18 vowels and 42 consonants. Some of the characters in the alphabet are borrowings from Sanskrit (Devanagari script) and not occurred in contemporary Sinhala. They include 'ඓ' (0D8F), 'ඖ' (0D90) with their corresponding modifiers 'ෟ' (0DDF), 'ෳ' (0DF3) and 'ඍෘ' (0D8E), and they were not considered in this work. There are 40 phonemes in spoken Sinhala with 14 vowels and 26 as given in table1 and 2.

10.21437/SLTU.2018-24

Table 2: *Sinhala consonant classification*[10]

		Lab.	Den.	Alv.	Ret.	Pal.	Vel.	Glo.
Stops	Voiceless	p	t		ʈ		k	
	Voiced	b	d		ɖ		g	
Affricates	Voiceless					c		
	Voiced					ɟ		
Pre-nasalized voiced stops		ᵬ	ᵭ		ᶑ		ᶢ	
Nasals		m		n		ɲ	ŋ	
Trill				r				
Lateral				l				
Spirants		f	s			ʃ		h
Semivowels		w				j		

3.1.1. ඍwith its distinctions

This is a specific letter in the Sinhala alphabet. Though 'ඍ' (0D8D) occurred as a vowel, with its modifier '◌ෘ' (0DD8) and '◌ෲ' (0DF2 - modifier of 'ඍෘ') pronounced as Consonant-Vowel Sequences (CVS) in spoken Sinhala. 'ඍ'is occurred in the initial position only for few words, and has two pronunciations as /ri/ for 'ඍණ' /rinə/ and /ir/ for 'ඍතු' /irtu/. Its modifier '◌ෘ' is used with consonants and pronounced different ways. Frequently it is pronounced as /ru/ with preceding consonants, thus 'මෘදු' and 'විස්තෘත' as /mrudu/ and /vistrutə/. For number of words it is pronounced as /ur/ for 'විකෘති' /vikurti/ and 'වෘකයා' /vurkəja/ but in 'ප්‍රවෘත්ති' /prəurti/, preceding consonant is in silent. When '◌ෘ' with h in the initial consonant, the '◌ෘ' is pronounced as /ər/ for 'හෘදේ' /hərdə/. But in the middle of words it sounds /ur/ as in 'සහෘදේ' /sahurdə/. Thus, there is a set of words, when '◌ෘ' is occurred the preceding consonant duplicates. For instance, 'පිතෘ'and 'ප්‍රාකෘත'pronounced as /pittru/ and /prakkrutə/. As a special case in 'කර්තෘ' /r/ is replaced with /t/ and pronounced /kattru/.

3.1.2. Anusvara and Visarga

Both *anusvara* '◌ං' (0D82) and *visarga* '◌ඃ' (0D83) are specific modifiers in Sinhala which can be preceded by any sign except *halanta* (0DCA). The *anusvara* pronounced /ŋ/ and frequently used in spoken Sinhala. *Visarga* /h/ is occurred in the borrowings from Sanskrit and rarely used.

3.1.3. Rakaranshaya and Yanshaya

Rakaranshaya (*RSY*) and *yanshaya* (*YSY*) are representations of 'ර' /r/ and 'ය' /j/ in Sinhala writing system. They are occurred after consonants (where the inherent vowel is removed) thus 'ක්‍රම' /krəmə/ and 'වාක්‍ය' /va:kjə/ are formed. In speaking, the preceding consonant of *RSY* or *YSY* is doubled if they occurred not in the followings; at the initial syllable of a word (ශ්‍රම /ʃrəmə/ and ධ්‍යාන /dja:nə/), two consonants together (චන්ද්‍ර /candrə/ and අචින්ත්‍ය /acintjə/) or with *anusvara* (ඉංග්‍රීසි /iŋgri:si/ and සංඛ්‍යා /saŋkja:/). However, in the other places it is getting doubled; 'අග්‍ර' /aggrə/, 'චිත්‍ර' /cittrə/ and 'අන්‍ය' /annjə/, 'සත්‍ය' /sattjə/ etc.

3.1.4. Diphthongs in Sinhala

There are 19 diphthongs in Spoken Sinhala. Two of them; i.e. /ai/ (ඓ) and /au/ (ඖ) has two separate letters in the character alphabet with corresponding modifiers. All the other diphthongs are made from 3 phonetic sequences of /vu/ 'වු', /v/ 'ව්' and /ji/ 'යි' which are formed with two semi vowels; /v/ and /j/. G2P diphthong mapping for developing TTS is not much complicated. However. Phoneme to grapheme mapping which is needed for automatic speech recognition is little complicated since some of the diphthongs have two or three written forms. Ex: diphthong /au/ has three forms as 'ඖ' for 'ඖෂධ' /auʃədə/, 'අවු' for 'අවුකන' /aukənə/ and 'අව්' for 'අව්ව' /auvə/. Diphthong mapping is given in table 5.

4. G2P Conversion

In this section we describe the rules we implemented for converting G2P for Sinhala. As mentioned above Sinhala G2P conversion cannot be done as a one to one mapping. There are several rules to adhere when converting both consonants and vowels in Sinhala.

4.1. Phonetic Scheme

The first step of G2P conversion is defining a symbol set for Sinhala phonemes. There are several options we can use to represent Sinhala phonemes. One way is to use IPA symbols which is the standard way to represent speech sounds. However, IPA symbols have non-ASCII characters and most of the speech processing engines do not support those characters. Hence, we need to define a symbol set using ASCII characters only. Typical way of defining phonetic symbols is to use the English alphabet to represent the speech sounds of the target language. Since English alphabet has only 26 characters and Sinhala has more characters (and speech sounds) than that, we should either use both lower-case and upper-case English characters or, use multiple characters to represent one speech sound. The choice between these two representations vary with the used speech processing engine as some engines only supports the use of single case (upper or lower) – where we should use multiple characters, and some engines does not support the use of multiple characters – where we should use both cases.

Table 3: *G2P Mapping for Vowels & Modifiers*

Independent Vowels	Dependent Vowels	Pronunciation
අ		a / ə
ආ	◌ා	a:
ඇ	◌ැ	æ
ඈ	◌ෑ	æ:
ඉ	◌ි	i
ඊ	◌ී	i:
උ	◌ු	u
ඌ	◌ූ	u:
ඍ	◌ෘ	ri / ru
	◌ෲ	ru:
එ	◌ෙ	e
ඒ	◌ේ	e:
ඓ	◌ෛ	ai
ඔ	◌ො	o
ඕ	◌ෝ	o:
ඖ	◌ෞ	au

In Sinhala we encounter three types of graphemes as vowels, consonants, and vowel modifiers. Each modifier is pronounced as same as their corresponding vowel. Hence, a vowel and its corresponding vowel modifier are mapped into same phonetic symbol (Table 3).

Also, Sinhala language has different orthographic representations for aspirated and non-aspirated versions of the same consonant. However, in colloquial Sinhala, the aspirated sounds are not pronounced, and they are pronounced same as their non-aspirated sound. Therefore, representing aspirated and non-aspirated sounds either with same symbol or two different symbols depends on the type of application. Table 4 shows the IPA mapping for Sinhala consonants.

Table 4: *G2P Mapping for Consonant Characters*

Consonant Character(s)	Pronunciation		Consonant Character(s)	Pronunciation
ක/ඛ	k		ඳ	d̆
ග/ඝ	g		ප/ඵ	p
ඞ/ං	ŋ		බ/භ	b
හ	ɡ̆		ම	m
ච/ඡ	c		ඹ	b̆
ජ/ඣ	ɟ		ය	j
ඤ/ඥ	ɲ		ර	r
ට/ඨ	ʈ		ල/ළ	l
ඩ/ඪ	ɖ		ව	v
ණ/න	n		ශ/ෂ	ʃ
ඬ	ɖ̆		ස	s
ත/ථ	t		හ/ඃ	h
ද/ධ	d		ෆ	F

4.2. G2P Rules

After defining a phonetic scheme, before applying any rules, we initially convert the given text (input) into its rough phonetic representation. In this step we combine the inherent vowel - schwa (/ə/) sound to each consonant which is without any modifier (For ex: මම->məmə). Then, using this initial representation, we apply the following rules to get an accurate phonetic representation of Sinhala text.

4.2.1. Consonants

In Sinhala G2P conversion, most of the consonants has a one to one mapping with the addition of inherent vowel sound or the corresponding modifier sound. However, there are some special cases when the pronunciation of the consonant differs from the normal.

Consonant_Rule_01

If the word contains a *RSY* or *YSY* - (The presence of these two characters are identified by the presence of the Zero-Width-Joiner (ZWJ) – a Unicode control character which is used to combine Sinhala consonants with /r/ or /j/), then the consonant immediately before these two characters should be duplicated.

E.g. 'චිත්‍ර' -> /citrə/ -> /cittrə/
'අන්‍ය'-> /anjə/ -> /annjə/

There are few situations that this rule does not apply.

a) If the *RSY* or *YSY* occurs in the first syllable of the word.

E.g.ශ්‍රම -> /ʃrəmə/ -> /ʃrəmə/

b) If a consonant cluster is followed by a RSY or YSY.

E.g.චන්ද්‍ර -> /cəndrə/ -> /cəndrə/

Consonant_Rule_02

If the word contains the consonant 'ඤ - /ɲ/', and /ɲ/ is followed by a vowel sound, then the sound /k/ should be added immediately before the /ɲ/ sound.

E.g.ප්‍රඥා -> /prəɲa:/ -> /prəkɲa:/
ඥාන-> /ɲa:nə/ -> /ɲa:nə/ => no difference

4.2.2. Vowels

In Sinhala vowels conversion, there are only two types of vowels that did not have one to one G2P mapping. First type is the inherent vowel and the second is Sinhala diphthongs. Inherent vowel in Sinhala is the only phoneme with two allophones as /a/ and /ə/ occurred in complementary distribution. The occurrence of diphthongs in Sinhala language is discussed in 3.1.4.

Rules of the inherent vowel

Most of the rules applied in converting Sinhala inherit vowel is covered in [10]. However, we have identified that those rules needed some modifications to get an accurate G2P conversion for Sinhala.

Vowel_Rule_01

If the nucleus of the first syllable is a schwa, the schwa should be replaced by vowel /a/, except in the following situations;

a) If the syllable starts with the CC structure.

-In the previous definition this was given as if the syllable starts only with consonant pair /s/ followed by /v/. This is not the only case as there are many other consonant pairs which does not follow the above rule. (i.e. /tv/, /st/, /vj/, /[consonant]r/)

b) If the first syllable starts with /k/ whereas, /k/ is followed by /ə/ and subsequently /ə/ is followed by /r/. (i.e. /kər/)

-In the previous definition the last part of this is given as, /ə/ is preceded by /r/, which gives the wrong order.

c) The word consists of a single syllable having CV structure (E.g. /də/)

Vowel_Rule_02

If /r/ is preceded by any consonant, followed by /ə/ and subsequently followed by /h/, then /ə/ should be replaced by /a/.(/[consonant]rəh/->/[consonant]rah/)

-In the previous definition this was given in 4 steps as:

a) (/[consonant]rəh/->/[consonant]rah/)
b) (/[consonant]rə[!h]/->/[consonant]ra[!h]/)
c) (/[consonant]ra[!h]/->/[consonant]rə!h]/)
d) (/[consonant]ra[h]/->/[consonant]ra[h]/)

where, it is sufficient to implement only the first option (a) as the next two options (b & c) counters each other and option d does nothing.

Vowel_Rule_03

If any vowel in the set {/a/, /e/, /æ/, /o/, /ə/} is followed by /h/ and subsequently /h/ is followed by schwa, then schwa should be replaced by vowel /a/.

(/[a, e, æ, o, ə]hə/) -> ([a, e, æ, o, ə]ha/)

-In previous definition the last part of the condition is given as, /h/ is preceded by schwa, which gives the wrong order.

Vowel_Rule_04

If schwa is followed by a consonant cluster, the schwa should be replaced by /a/.

(/ə[consonant][consonant]/)->(/a[consonant][consonant]/)

Vowel_Rule_05

If /ə/ is followed by the word final consonant (without inherent vowel or modifiers), it should be replaced by /a/, except in the situations where the word final consonant is /r/, /b/, /ɖ/ or /ʈ/. Further, if a word contains either of these consonant-vowel clusters at the end, the word final position is retained; i.e. /ʃən/, /rəl/, /nəl/, /bəl/, /jəs/, /ʃəl/, /ɖən/, /jəl/, /gən/, /rəs/, /pəl/, /ləs/, nəs//, /kəl/, and /bən/.

(/ə![r, b, ɖ,ʈ]#) ->(/a![r, b, ɖ,ʈ]#)

Vowel_Rule_06

At the end of a word, if schwa precedes the phoneme sequence /ji/ or /vu/, the schwa should be replaced by /a/

(/ə[ji, vu]/#) ->(/a[ji, vu]/#)

-In the previous definition the rule is defined only for the /ji/ pair. However, it is true to the /vu/ pair also.

Vowel_Rule_07

If phoneme /k/ is followed by schwa, and subsequent phonemes are /r/ or /l/ followed by /u/, then schwa should be replaced by phoneme /a/.

(/kə[r, l]u/) -> (/ka[r, l]u/)

Vowel_Rule_08

Within the given context of following words, /a/ found in phoneme sequence /kal/, (the left-hand side of the arrow) should be changed to /ə/ as shown in the right-hand side.

- /kal(a:|e:|o:)j/->/kəl(a:|e:|o:)j/
- /kale(m|h)(u|i)/->/kəle(m|h)(u|i)/
- /kaləh(u|i)/->/kəleh(u|i)/
- /kalə/->/kələ/

Table 5: *Sinhala Diphthongs Mapping with Examples*

Phoneme sequences	Diphthong	Example
/ivu/ /iv/	/iu/	‘කිවුවා’‘කිව්වා’
/i:vu/ /i:v/	/i:u/	‘රජිව්’
/evu/ /ev/	/eu/	‘පෙවුවා’‘පෙව්වා’
/e:vu/ /e:v/	/e:u/	‘සේව්’
/ævu/ /æv/	/æu/	‘නැවුවා’ ‘නැව්වා’
/æ:vu/ /æ:v/	/æ:u/	‘බෑවුම’ ‘ගැව්වා’
/ovu/ /ov/	/ou/	‘ඔවුන්’ ‘පොව්ව’
/avu/ /av/	/au/	‘කවුද’ ‘කව්පි’
/a:vu/ /a:v/	/a:u/	‘සාවුරුද්දක්’
/uyi/	/ui/	‘බකුයි’
/u:yi/	/u:i/	‘දූයි’
/oyi/	/oi/	‘පූසොයි’
/o:yi/	/o:i/	‘රෝයි’
/ayi/	/ai/	‘කයි’
/a:yi/	/a:i/	‘මායි’
/eyi/	/ei/	‘බෑලෙයි’
/e:yi/	/e:i/	‘ගේයි’
/æyi/	/æi/	‘ඇයි’
/æ:yi/	/æ:i/	‘ඇයි’

Rules of the diphthongs

After the above rules were applied, we get a fairly accurate conversion of Sinhala G2P. However, there are some situations where we get a sequence of vowel + semi vowel together. In Sinhala these vowel + semi vowel sequences are converted into their corresponding diphthongs. Table 5 shows the occurrences of vowel + semi vowel sequences and the corresponding diphthongs which are used to replace the relevant vowel + semi vowel pair with examples.

5. Evaluation& Discussion

To evaluate our system, we extracted the most frequent 15,000 words from UCSC 10M Sinhala corpus. The converted word list was given to a linguist to identify errors in the conversion. From the 15,000 words, 428 words were identified as incorrectly converted. A breakdown of the incorrectly identified words is illustrated in table 6.

Table 6: *Error analysis*

Error type	# of words
Compound words	202
Complication with '◌ා' and inherent vowel	78
Foreign words	69
Homographs	22
Exceptions for the G2P Rules	17
In-Script Variants	5
Other	35

After analyzing the results, we identified that there are some occurrences that cannot be handled with the G2P rules. Hence, we used a lexicon with direct mapping for a finite set of words to resolve the issues discussed in 3.1.1 and exceptions discussed in G2P rules. Issues that cannot be dealt with a lexicon are discussed below.

As a morphologically rich and productive language, Sinhala uses more inflectional and derivational suffixes forming new words. Also in Sinhala, compound words are formed very frequently. When closed compounds are formed with two words containing consonants with inherent vowel, it changes the contexts. However, they are pronounced as separate words. E.g. ‘නව’ /navə/ and ‘කථා’ /kata:/ are two words which can be used together to yield the compound word ‘නවකතා’ /navəkata:/ in which pronunciation of inherent vowel has not been changed and it is a challenge to identify these occurrences.

In some situations, the modifier ‘◌ා’ /a:/ at the end of a word is pronounced as its short representation /a/ in Sinhala and sometime the ‘◌ා’ modifier is omitted from written form also. This misleads the rules to apply the /ə/ inherent. E.g. ‘ගත්තා’ is pronounced as /gatta/ and it is incorrectly written as ‘ගත්ත’ which is converted to /gattə/ according to the G2P rules.

6. Conclusion

In this paper we have discussed how we apply rules in Sinhala G2P conversion. We have enhanced a previously defined set of G2P rules. Our evaluation shows approximately 98% accuracy. To get a better accuracy, we have used a predefined lexicon to convert exceptional cases. This work is available as an online resource at: http://transliteration.sinhala.subasa.lk/

7. Acknowledgements

The authors of this paper would like to thank Language Technology Research Laboratory – University of Colombo School of Computing for the support given to make this work a success.

8. References

[1] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication,* vol. 50, no. 5, pp. 434-451, 2008.

[2] M. Choudhury, "Rule-based grapheme to phoneme mapping for hindi speech synthesis," in *90th Indian Science Congress of the International Speech Communication Association (ISCA)*, Bangalore, India, 2003.

[3] A. Van Den Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, 1993.

[4] U. Reichel, H. R. Pfitzinger and H. U. Hain, "English grapheme-to-phoneme conversion and evaluation," *Speech and Language Technology,* vol. 11, pp. 159-166, 2008.

[5] K. Rao, F. Peng, H. Sak and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, 2015.

[6] N. Torstensson, "Grapheme-to-phoneme conversion, a knowledge-based approach," *Speech Music and Hearing TMH-QPSR-Fonetik,* vol. 44, pp. 117-120, 2002.

[7] S. S. Nair, C. R. Rechitha and C. S. Kumar, "Rule-Based Grapheme to Phoneme Converter for Malayalam," *International Journal of Computational Linguistics and Natural Language Processing,* vol. 2, no. 7, pp. 417-420, 2013.

[8] A. Chalamandaris, S. Raptis and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] Y. C. Wang and R. T. H. Tsai, "Rule-based korean grapheme to phoneme conversion using sound patterns," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, 2009.

[10] A. Wasala, R. Weerasinghe and K. and Gamage, "Sinhala grapheme-to-phoneme conversion and rules for schwa epenthesis," in *Proceedings of the COLING/ACL on Main conference poster sessions*, 2006.