Sinhala Grapheme-to-Phoneme Conversion and Rules for Schwa Epenthesis

Asanka Wasala, Ruvan Weerasinghe and Kumudu Gamage
Language Technology Research Laboratory
University of Colombo School of Computing
35, Reid Avenue, Colombo 07, Sri Lanka
{awasala,kgamage}@webmail.cmb.ac.lk, arw@ucsc.cmb.ac.lk

Abstract

This paper describes an architecture to convert Sinhala Unicode text into phonemic specification of pronunciation. The study was mainly focused on disambiguating schwa-/ə/ and /a/ vowel epenthesis for consonants, which is one of the significant problems found in Sinhala. This problem has been addressed by formulating a set of rules. The proposed set of rules was tested using 30,000 distinct words obtained from a corpus and compared with the same words manually transcribed to phonemes by an expert. The Grapheme-to-Phoneme (G2P) conversion model achieves 98 % accuracy.

1 Introduction

The conversion of Text-to-Speech (TTS) involves many important processes. These processes can be divided mainly in to three parts; text analysis, linguistic analysis and waveform generation (Black and Lenzo, 2003). The text analysis process is responsible for converting the non-textual content into text. This process also involves tokenization and normalization of the text. The identification of words or chunks of text is called text-tokenization. Text normalization establishes the correct interpretation of the input text by expanding the abbreviations and acronyms. This is done by replacing the non-alphabetic characters, numbers, and punctuation with appropriate text strings depending on the context. The linguistic analysis process involves finding the correct pronunciation of words, and assigning prosodic features (eg. phrasing, intonation, stress) to the phonemic string to be spoken. The final process of a TTS system is waveform generation which involves the production of an acoustic digital signal using a particular synthesis approach such as formant synthesis, articulatory synthesis or waveform concatenation (Lemmetty, 1999). The text analysis and linguistic analysis processes together are known as the Natural Language Processing (NLP) component, while the waveform generation process is known as the Digital Signal Processing (DSP) component of a TTS System (Dutoit, 1997).

Finding correct pronunciation for a given word is one of the first and most significant tasks in the linguistic analysis process. The component which is responsible for this task in a TTS system is often named the Grapheme-To-Phoneme (G2P), Text-to-Phone or Letter-To-Sound (LTS) conversion module. This module accepts a word and generates the corresponding phonemic transcription. Further, this phonemic transcription can be annotated with appropriate prosodic markers (Syllables, Accents, Stress etc) as well.

In this paper, we describe the implementation and evaluation of a G2P conversion model for a Sinhala TTS system. A Sinhala TTS system is being developed based on *Festival*, the open source speech synthesis framework. Letter to sound conversion for Sinhala usually has simple one to one mapping between orthography and phonemic transcription for most Sinhala letters. However some G2P conversion rules are proposed in this paper to complement the generation of more accurate phonemic transcription.

The rest of this paper is organized as follows: Section 2 gives an overview of the Sinhala phonemic inventory and the Sinhala writing system, Section 3 briefly discusses G2P conversion approaches. Section 4 describes the schwa epenthesis issue peculiar to Sinhala and Section 5 explains the Sinhala G2P conversion architecture.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 890–897,
Sydney, July 2006.

Section 6 gives experimental results and our discussion on it. The work is summarized in the final section.

2 Sinhala Phonemic Inventory and Writing System

2.1 The Sinhala Phonemic Inventory

Sinhala is the official language of Sri Lanka and the mother tongue of the majority - 74% of its population. Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants as classified below in Table 1 and Table 2 (Karunatillake, 2004).

There are two nasalized vowels occurring in two or three words in Sinhala. They are /ã/, /ã:/, /æ̃/ and /æ̃:/ (Karunatillake, 2004). Spoken Sinhala also has following Diphthongs; /iu/, /eu/, /æu/, /ou/, /au/, /ui/, /ei/, /æi/, /oi/ and /ai/ (Disanayaka, 1991).

	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:		

Table 1. Spoken Sinhala Vowel Classification.

		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		ɓ	ɗ		ᶑ		g̃	
Nasals		m		n		ɲ	ŋ	
Trill				r				
Lateral				l				
Spirants		f	s			ʃ		h
Semivowels		w				j		

Table 2*. Spoken Sinhala Consonant Classification.

A separate sign for vowel /ə/ is not provided by the Sinhala writing system. In terms of distribution, the vowel /ə/ does not occur at the beginning of a syllable except in the conjugational variants of verbs formed from the verbal stem /kərə/ *(to do)*. In contrast to this, though the letter "ඥ", which symbolizes the consonant sound /ɲ/ exists, it is not considered a phoneme in Sinhala.

2.2 The Sinhala Writing System

The Sinhala character set has 18 vowels, and 42 consonants as shown in Table 3.

Vowels and corresponding vowel modifiers (within brackets):
අ ආ(ා) ඇ(ැ) ඈ(ෑ) ඉ(ි) ඊ(ී) උ(ු) ඌ(ූ) ඍ(ෘ) ඎ(ෲ) ඏ(ෟ) ඐ(෠) එ(ෙ) ඒ(ේ) ඓ(ෛ) ඔ(ො) ඕ (ෝ) ඖ(ෞ)

Consonants:
ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඥ ඦ ට ඨ ඩ ඪ ණ ඬ ත ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ ං ඃ

Special symbols: ් ා ෘ ෲ
Inherent vowel remover (*Hal* marker): ්

Table 3. Sinhala Character Set.

Sinhala characters are written left to right in horizontal lines. Words are delimited by a space in general. Vowels have corresponding full-character forms when they appear in an absolute initial position of a word. In other positions, they appear as 'strokes' and, are used with consonants to denote vowel modifiers. All vowels except "ඍ" /iru:/, are able to occur in word initial positions (Disanayaka, 1995). The vowel /ə/ and /ə:/ occurs only in loan words of English origin. Since there are no special symbols to represent them, frequently the "අ" vowel is used to symbolize them (Karunatillake, 2004).

All consonants occur in word initial position except /ŋ/ and nasals (Disanayaka, 1995). The symbols "ණ", and "ළ" represent the retroflex nasal /ɳ/ and the retroflex lateral /ɭ/ respectively. But they are pronounced as their respective alveolar counterparts "න"-/n/ and "ල"-/l/. Similarly, the symbol "ෂ" representing the retroflex sibilant /ʂ/, is pronounced as the palatal sibilant "ශ"-/ʃ/. The corresponding aspirated symbols of letters ක, ග, ච, ජ, ට, ඩ, ත, ද, ප, බ namely ඛ, ඝ, ඡ, ඣ, ඨ, ඪ, ථ, ඵ, භ respectively are pronounced like the corresponding unaspirates (Karunatillake, 2004). When consonants are combined with /r/ or /j/, special conjunct symbols are used. "ර්"-/r/ immediately following a consonant can be marked by the symbol "්ර" added to the bottom of the consonant preceding it. Similarly, "ය්"-/j/, immediately following consonant can be marked by the symbol "්ය"

* Lab. – Labial, Den. – Dental, Alv. – Alveolar, Ret. – Retroflex, Pal. – Palatal, Vel. – Velar and Glo. – Glottal.

added to the right-hand side of the consonant preceding it (Karunatillake, 2004). "ඔ" /ilu/ and "ඕ" /ilu:/ do not occur in contemporary Sinhala (Disanayaka, 1995). Though there are 60 symbols in Sinhala (Disanayaka, 1995), only 42 symbols are necessary to represent Spoken Sinhala (Karunatillake, 2004).

3 G2P Conversion Approaches

The issue of mapping textual content into phonemic content is highly language dependent. Three main approaches of G2P conversion are; use of a pronunciation dictionary, use of well defined language-dependent rules and data-driven methods (El-Imam and Don, 2005).

One of the easiest ways of G2P conversion is the use of a lexicon or pronunciation dictionary. A lexicon consists of a large list of words together with their pronunciation. There are several limitations to the use of lexicons. It is practically impossible to construct such to cover the whole vocabulary of a language owing to Zipfian phenomena. Though a large lexicon is constructed, one would face other limitations such as efficient access, memory storage etc. Most lexicons often do not include many proper names, and only very few provide pronunciations for abbreviations and acronyms. Only a few lexicons provide distinct entries for morphological productions of words. In addition, pronunciations of some words differ based on the context and their parts-of-speech. Further, an enormous effort has to be made to develop a comprehensive lexicon. In practical scenarios, speech synthesizers as well as speech recognizers need to be able to produce the pronunciation of words that are not in the lexicon. Names, morphological productivity and numbers are the three most important cases that cause the use of lexica to be impractical (Jurafsky and Martin, 2000).

To overcome these difficulties, rules can be specified on how letters can be mapped to phonemes. In this way, the size of the lexicon can be reduced as only to contain exceptions to the rules. In contrast to the above fact, some systems rely on using very large lexicons, together with a set of letter-to-sound conversion rules to deal with words which are not found in the lexicon (Black and Lenzo, 2003).

These language and context dependent rules are formulated using phonetic and linguistic knowledge of a particular language. The complexity of devising a set of rules for a particular language is dependent on the degree of correspondence between graphemes and phonemes. For some languages such as English and French, the relationship is complex and require large numbers of rules (El-Imam and Don, 2005; Damper et al., 1998), while some languages such as Urdu (Hussain, 2004), and Hindi (Ramakishnan et al., 2004; Choudhury, 2003) show regular behavior and thus pronunciation can be modeled by defining fairly regular simple rules.

Data-driven methods are widely used to avoid tedious manual work involving the above approaches. In these methods, G2P rules are captured by means of various machine learning techniques based on a large amount of training data. Most previous data-driven approaches have been used for English. Widely used data-driven approaches include, Pronunciation by Analogy (PbA), Neural Networks (Damper et al., 1998), and Finite-State-Machines (Jurafsky and Martin, 2000). Black et al. (1998) discussed a method for building general letter-to-sound rules suitable for any language, based on training a CART – decision tree.

4 Schwa Epenthesis in Sinhala

G2P conversion problems encountered in Sinhala are similar to those encountered in the Hindi language (Ramakishnan et al., 2004). All consonant graphemes in Sinhala are associated with an inherent vowel schwa-/ə/ or /a/ which is not represented in orthography. Vowels other than /ə/ and /a/ are represented in orthographic text by placing specific vowel modifier diacritics around the consonant grapheme. In the absence of any vowel modifier for a particular consonant grapheme, there is an ambiguity of associating /ə/ or /a/ as the vowel modifier. The inherent vowel association in Sinhala can be distinguished from Hindi. In Hindi the only possible association is schwa vowel where as in Sinhala either of vowel-/a/ or schwa-/ə/ can be associated with a consonant. Native Sinhala speakers are naturally capable of choosing the association of the appropriate vowel (/ə/ or /a/) in context. Moreover, linguistic rules describing the transformation of G2P, is rarely found in literature, with available literature not providing any precise procedure suitable for G2P conversion of contemporary Sinhala. Automating the G2P conversion process is a difficult task due to the ambiguity of choosing between /ə/ and /a/.

A similar phenomenon is observed in Hindi and Malay as well. In Hindi, the "deletion of the schwa vowel *(in some cases)*" is successfully

solved by using rule based algorithms (Choudhury 2003; Ramakishnan et al., 2004). In Malay, the character 'e' can be pronounced as either vowel /e/ or /ə/, and rule based algorithms are used to address this ambiguity (El-Imam and Don, 2005).

In our research, a set of rules is proposed to disambiguate epenthesis of /a/ and /ə/, when associating with consonants. Unlike in Hindi, in Sinhala, the schwa is not deleted, instead always inserted. Hence, this process is named "Schwa Epenthesis" in this paper.

5 Sinhala G2P Conversion Architecture

An architecture is proposed to convert Sinhala Unicode text into phonemes encompassing a set of rules to handle schwa epenthesis. The G2P architecture developed for Sinhala is identical to the Hindi G2P architecture (Ramakishnan et al., 2004). The input to the system is normalized Sinhala Unicode text. The G2P engine first maps all characters in the input word into corresponding phonemes by using the letter-to-phoneme mapping table below (Table 4).

අ	/a/	ඔ ,ොෟ	/o/	ඩ	/ɖ/	ෆ	/f/
ආ,ා	/a:/	ඕ,ෝ	/o:/	ත,ථ	/t/	ාෘ	/ru:/
ඇ,ැ	/æ/	ඖ,ෞ	/ou/	ද,ධ	/d/		
ඈ,ෑ	/æ:/	ක,ඛ	/k/	ඳ	/ð/		
ඉ ,ි	/i/	ග,ඝ	/g/	ප,ඵ	/p/		
ඊ,ී	/i:/	ඞ,ං	/ŋ/	බ,භ	/b/		
උ,ු	/u/	ඟ	/ɠ/	ම	/m/		
ඌ,ූ	/u:/	ච,ඡ	/c/	ඹ	/ƀ/		
සෘ	/ri/	ජ,ඣ	/ɟ/	ය	/j/		
ෘ	/ru/	ඤ	/ɲ/	ර	/r/		
ඵ	/ilu/	ඥ	/jɲ/	ල,ළ	/l/		
ඹෟ	/ilu:/	ඡ	/ʄ/	ව	/w/		
එ ,ෙ	/e/	ට,ඨ	/ʈ/	ශ,ෂ	/ʃ/		
ඒ,ේ	/e:/	ඩ,ඪ	/ɖ/	ස	/s/		
ඓ,ෛ	/ai/	න,ණ	/n/	හ,ඃ	/h/		

Table 4. G2P Mapping Table

The mapping procedure is given in section 5.1. Then, a set of rules are applied to this phonemic string in a specific order to obtain a more accurate version. This phonemic string is then compared with the entries in the exception lexicon. If a matching entry is found, the correct pronunciation form of the text is obtained from the lexicon, otherwise the resultant phonemic string is returned. Hence, the final output of G2P model is the phonemic transcription of the input text.

5.1 G2P Mapping Procedure

Each tokenized word represented by Unicode normalization form is analyzed by individual graphemes from left to right. By using the G2P mapping table (Table 4), corresponding phonemes are obtained. As in the given example Figure 1, no mappings are required for the Zero-Width-Joiner and diacritic *Hal* marker "්" *(Halant)* which is used to remove the inherent vowel in a consonant.

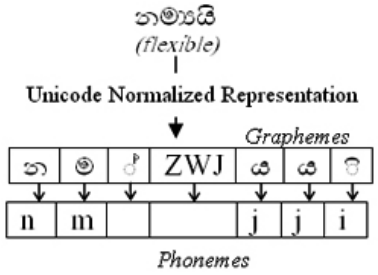

Figure 1. G2P Mapping (Example).

The next step is epenthesis of schwa-/ə/ for consonants. In Sinhala, the tendency of associating a /ə/ with consonant is very much higher than associating vowel /a/. Therefore, initially, all plausible consonants are associated with /ə/. To obtain the accurate pronunciation, the assigned /ə/ is altered to /a/ or vice versa by applying the set of rules given in next section. However, when associating /ə/ with consonants, /ə/ should associate only with consonant graphemes excluding the graphemes "ං", "ඞ" and "ඃ", which do not contain any vowel modifier or diacritic *Hal* marker. In the above example, only /n/ and first /j/ are associated with schwa, because other consonants violate the above principle. When schwa is associated with appropriate consonants, the resultant phonemic string for the given example (section 5.1) is; /nəmjəji/.

5.2 G2P Conversion Rules

It is observed that resultant phoneme strings from the above procedure should undergo several modifications in terms of schwa assignments into vowel /a/ or vice versa, in order to obtain the accurate pronunciation of a particular word. Guided by the literature (Karunatillake, 2004), it was noticed that these modifications can be carried out by formulating a set of rules.

The G2P rules were formulated with the aid of phonological rules described in the linguistic literature (Karunatillake, 2004) and by a comprehensive word search analysis using the *UCSC*

Sinhala corpus BETA (2005). Some of these existing phonological rules were altered in order to reflect the observations made in the corpus word analysis and to achieve more accurate results. The proposed new set of rules is empirically shown to be effective and can be conveniently implemented using regular expressions.

Each rule given below is applied from left to right, and the presented order of the rules is to be preserved. Except for rule #1, rule #5, rule #6 and rule #8, all other rules are applied repeatedly many times to a single word until the conditions presented in the rules are satisfied.

Rule #1: If the nucleus of the first syllable is a schwa, the schwa should be replaced by vowel /a/ (Karunatillake, 2004), except in the following situations;

(a) The syllable starts with /s/ followed by /v/. (ie. /sv/)

(b) The first syllable starts with /k/ where as, /k/ is followed by /ə/ and subsequently /ə/ is preceded by /r/. (ie. /kər/)

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

Rule #2:

(a) 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/)

(b) If /r/ is preceded by any consonant, followed by /ə/ and subsequently followed by any consonant other than /h/, then /ə/ should be replaced by /a/.

(/[consonant]rə[!h]/->/[consonant]ra[!h]/)

(c) If /r/ is preceded by any consonant, followed by /a/ and subsequently followed by any consonant other than /h/, then /a/ should be replaced by /ə/.

(/[consonant]ra[!h]/->/[consonant]rə!h]/)

(d) If /r/ is preceded by any consonant, followed by /a/ and subsequently followed by /h/, then /a/ is retained.

(/[consonant]ra[h]/->/[consonant]ra[h]/)

Rule #3: If any vowel in the set {/a/, /e/, /æ/, /o/, /ə/} is followed by /h/ and subsequently /h/ is preceded by schwa, then schwa should replaced by vowel /a/.

Rule #4: If schwa is followed by a consonant cluster, the schwa should be replaced by /a/ (Karunatillake, 2004).

Rule #5: If /ə/ is followed by the word final consonant, it should be replaced by /a/, except in the situations where the word final consonant is /r/, /b/, /ɖ/ or /ʈ/.

Rule #6: At the end of a word, if schwa precedes the phoneme sequence /ji/, the schwa should be replaced by /a/ (Karunatillake, 2004).

Rule #7: If the /k/ is followed by schwa, and subsequent phonemes are /r/ or /l/ followed by /u/, then schwa should be replaced by phoneme /a/. (ie. /kə(r|l)u/->/ka(r|l)u/)

Rule #8: 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:)y/->/kəl(a:|e:|o:)y/
- /kale(m|h)(u|i)/->/kəle(m|h)(u|i)/
- /kaləh(u|i)/->/kəleh(u|i)/
- /kalə/->/kələ/

The above rules handle the schwa epenthesis problem. The corresponding diphthongs (refer section 2) are then obtained by processing the resultant phonetized string. This string is again analyzed from left to right, and the phoneme sequences given in the first column of Table 5 are replaced by the diphthong, represented in the second column.

Phoneme Sequence	Diphthong
/i/ /w/ /u/	/iu/
/e/ /w/ /u/	/eu/
/æ/ /w/ /u/	/æu/
/o/ /w/ /u/	/ou/
/a/ /w/ /u/	/au/
/u/ /j/ /i/	/ui/
/e/ /j/ /i/	/ei/
/æ/ /j/ /i/	/æi/
/o/ /j/ /i/	/oi/
/a/ /j/ /i/	/ai/

Table 5. Diphthong Mapping Table.

The application of the above rules for the given example (section 5.1) is illustrated in Figure 2.

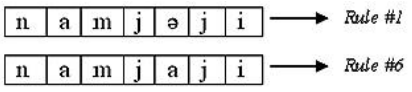

Figure 2. Application of G2P Rules – An Example.

6 Results and Discussion

Text obtained from the category *"News Paper> Feature Articles > Other"* of the *UCSC Sinhala* corpus was chosen for testing due to the heterogeneous nature of these texts and hence perceived better representation of the language in this part of the corpus[*]. A list of distinct words was first extracted, and the 30,000 most frequently occurring words chosen for testing.

The overall accuracy of our G2P module was calculated at 98%, in comparison with the same words correctly transcribed by an expert.

Since this is the first known documented work on implementing a G2P scheme for Sinhala, its contribution to the existing body of knowledge is difficult to evaluate. However, an experiment was conducted in order to arrive at an approximation of the scale of this contribution.

It was first necessary, to define a baseline against which this work could be measured. While this could be done by giving a single default letter-to-sound mapping for any Sinhala letter, owing to the near universal application of rule #1 in Sinhala words (22766 of the 30000 words used in testing), the baseline was defined by the application of this rule in addition to the 'default mapping'. This baseline gives us an error of approximately 24%. Since the proposed solution reduces this error to 2%, this work can claim to have improved performance by 22%.

An error analysis revealed the following types of errors (Table 6):

Error description	**# of words**
Compound words- (ie. Single words formed by combining 2 or more distinct words; such as in the case of the English word "thereafter").	382
Foreign (mainly English) words directly encoded in Sinhala. eg. ෆැෂන් - fashion, කැම්පස් - campus.	116
Other	118

Table 6. Types of Errors.

The errors categorized as "Other" are given below with clarifications:

- The modifier used to denote long vowel "ආ" /a:/ is "ා" which is known as "Aela-pilla". eg. consonant "ක්" /k/ associates with "ා" /a:/ to produce grapheme "කා" is pronounced as /ka:/. The above exercise revealed some 37 words end without vowel modifier "ා", but are usually pronounced with the associated long vowel /a:/. In the following examples, each input word is listed first, followed by the erroneous output of G2P conversion, and correct transcription.

  "අම්ම"(mother) -> /ammə/ -> /amma:/
  "අක්ක"*(sister)* -> /akkə/ -> /akka:/
  "ගත්ත"*(taken)*-> /gattə/ -> /gatta:/

- There were 27 words associated with erroneous conversion of words having the letter "හ", which corresponds to phoneme /h/. The study revealed this letter shows an unusual behavior in G2P conversion.

- The modifier used to denote vowel "ඍ" - "ෘ" is known as *"Geta-pilla"*. When this vowel appears as the initial letter of a word, it is pronounced as /ri/ as in "ඍණ" /rinə/ *(minus)*. When the corresponding vowel modifier appears in a middle of a word most of the time it is pronounced as /ru/ (Disanayaka, 2000). eg. "කෘතිය" *(book)*is pronounced as /krutijə/, "පෘෂ්ඨය" *(surface)* - /pruʃʈəjə/, "උත්කෘෂ්ට" *(excellent)*-/utkruʃʈə/. But 13 words were found as exceptions of this general rule. In those words, the "ෘ" is pronounced as /ur/ rather than /ru/. eg. "ප්‍රවෘත්ති" *(news)*-/prəwurti/,"සමෘද්ධි"*(prosperity)*-/samurdi/, "විවෘත" *(opened)* - /wiwurtə/.

- In general, vowel modifiers "ැ" *(Adha-pilla)*, "ෑ" *(Diga Adha-pilla)* symbolizes the vowel "ඇ" /æ/ and "ඈ" /æ:/ respectively. eg. consonant "ක්" /k/ combines with vowel modifier "ැ" to create "කැ" which is pronounced as /kæ/. Few words were found where this rule is violated. In such words, the vowel modifiers "ැ" and "ෑ" represent vowels "උ"- /u/, and "ඌ"- /u:/ respectively. eg. "ජනශ්‍රැති" *(legend)* - /ʄanəʃruti/, "ක්‍රෑර" *(cruel)* - /kru:rə/.

- The verbal stem "කර" *(to do)* is pronounced as /kərə/. Though there are many words starting with the same verbal stem, there are a few other words differently pronounced as /karə/ or /kara/. eg. "කරත්තය" *(cart)* /karattəyə/, "කරවල" *(dried fish)* /karəvələ/.

[*] This accounts for almost two-thirds of the size of this version of the corpus.

- A few of the remaining errors are due to homographs; “වන” - /vanə/, /vənə/; “කල” -/kalə/, /kələ/; “කර” - /karə/, /kərə/.

The above error analysis itself shows that the model can be extended. Failures in the current model are mostly due to compound words and foreign words directly encoded in Sinhala (1.66%). The accuracy of the G2P model can be increased significantly by incorporating a method to identify compound words and transcribe them accurately. If the constituent words of a compound word can be identified and separated, the same set of rules can be applied for each constituent word, and the resultant phonetized strings combined to obtain the correct pronunciation. The same problem is observed in the Hindi language too. Ramakishnan et al. (2004) proposed a procedure for extracting compound words from a Hindi corpus. The utilization of compound word lexicon in their rule-based G2P conversion module improved the accuracy of G2P conversion by 1.6% (Ramakishnan et al., 2004). In our architecture, the most frequently occurring compound words and foreign words are dealt with the aid of an exceptions lexicon. Homographs are also disambiguated using the most frequently occurring words in Sinhala. Future improvements of the architecture will include incorporation of a compound word identification and phonetization module.

7 Conclusion

In this paper, the problem of Sinhala grapheme-to-phoneme conversion is addressed with a special focus on dealing with the schwa epenthesis. The proposed G2P conversion mechanism will be useful in various applications in the speech domain. To the best of our knowledge no other documented evidence has been reported for Sinhala grapheme-to-phoneme conversion in the literature. There are no other approaches available for the transcription of Sinhala text that provides a platform for comparison of the proposed rule-based method. The empirical evidence from a wide spectrum Sinhala corpus indicates that the proposed model can account for nearly 98% of cases accurately.

The proposed G2P module is fully implemented in Sinhala TTS being developed at Language Technology Research Lab, UCSC. A demonstration tool of the proposed G2P module integrated with Sinhala syllabification algorithm proposed by Weerasinghe et al. (2005) is available for download from:

http://www.ucsc.cmb.ac.lk/ltrl/downloads.html

Acknowledgement

This work has been supported through the PAN Localization Project, (http://www.PANL10n.net) grant from the International Development Research Center (IDRC), Ottawa, Canada, administered through the Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan. The authors would like to thank Sinhala Language scholars Prof. R.M.W. Rajapaksha, and Prof. J.B. Dissanayake for their invaluable support and advice throughout the study. Special thanks to Dr. Sarmad Hussain (NUCES, Pakistan) for his guidance and advices. We also wish to acknowledge the contribution of Mr. Viraj Welgama, Mr. Dulip Herath, and Mr. Nishantha Medagoda of Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka.

References

Alan W. Black and Kevin A. Lenzo. 2003. *Building Synthetic Voices*, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from http://festvox.org/bsv/

Alan W. Black, Kevin Lenzo, and Vincent Pagel. 1998. Issues in Building General Letter to Sound Rules. *In Proc. of the 3rd ESCA Workshop on Speech Synthesis*, pages 77–80.

Monojit Choudhury. 2003. Rule-Based Grapheme to Phoneme Mapping for Hindi Speech Synthesis, *presented at the 90th Indian Science Congress of the International Speech Communication Association (ISCA)*, Bangalore.

R.I. Damper, Y. Marchand, M.J. Adamson and K. Gustafson. 1998. Comparative Evaluation of Letter-to-Sound Conversion Techniques for English Text-to-Speech Synthesis. *In Proc. Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 53- 58, Blue Mountains, NSW, Australia.

J.B. Disanayaka. 1991. *The Structure of Spoken Sinhala*, National Institute of Education, Maharagama.

J.B. Disanayaka. 2000. *Basaka Mahima: 2, Akuru ha pili*, S. Godage & Bros., 661, P. D. S. Kularathna Mawatha, Colombo 10.

J.B. Disanayaka. 1995. *Grammar of Contemporary Literary Sinhala - Introduction to Grammar,*

Structure of Spoken Sinhala, S. Godage & Bros., 661, P. D. S. Kularathna Mawatha, Colombo 10.

T. Dutoit. 1997. *An Introduction to Text-to-Speech Synthesis,* Kluwer Academic Publishers, Dordrecht, Netherlands.

Yousif A. El-Imam and Zuraidah M. Don. 2005. Rules and Algorithms for Phonetic Transcription of Standard Malay**,** *IEICE Trans Inf & Syst*, E88-D 2354-2372.

Sarmad Hussain. 2004. Letter-to-Sound Conversion for Urdu Text-to-Speech System, *Proceedings of Workshop on "Computational Approaches to Arabic Script-based Languages," COLING 2004,* p. 74-49, Geneva, Switzerland.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education (Singapore) Pte. Ltd, Indian Branch, 482 F.I.E. Patparganj, Delhi 110 092, India.

W.S. Karunatillake. 2004. *An Introduction to Spoken Sinhala*, 3rd edn., M.D. Gunasena & Co. ltd., 217, Olcott Mawatha, Colombo 11.

Sami Lemmetty. 1999. *Review of Speech Synthesis Technology*, MSc. thesis, Helsinki University of Technology.

A.G. Ramakishnan, Kalika Bali, Partha Pratim Talukdar N. and Sridhar Krishna. 2004. Tools for the Development of a Hindi Speech Synthesis System, *In 5th ISCA Speech Synthesis Workshop*, Pittsburgh. pages 109-114.

Ruvan Weerasinghe, Asanka Wasala and Kumudu Gamage. 2005. A Rule Based Syllabification Algorithm for Sinhala, *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, p. 438-449, Jeju Island, Korea.

*UCSC Sinhala Corpus BETA***.** 2005. Retrieved August 30, 2005, from University of Colombo School of Computing, Language Technology Research Laboratory Web site:

http://www.ucsc.cmb.ac.lk/ltrl/downloads.html