Research Report on Phonetics and Phonology of Sinhala

Asanka Wasala and Kumudu Gamage
Language Technology Research Laboratory,
University of Colombo School of Computing,
Sri Lanka.
awasala@webmail.cmb.ac.lk,kgamage@webmail.cmb.ac.lk

Abstract

This report examines the major characteristics of Sinhala language related to Phonetics and Phonology. The main topics under study are Segmental and Supra-segmental sounds in Spoken Sinhala. The first part presents Sinhala Phonemic Inventory, which describes phonemes with their associated features and phonotactics of Sinhala. Supra-segmental features like Syllabification, Stress, Pitch and Intonation, followed by the procedure for letter to sound conversion are described in the latter part.

The research on Syllabification leads to the identification of set of rules for syllabifying a given word. The syllabication algorithm achieved an accuracy of 99.95%. Due to the phonetic nature of the Sinhala alphabet, the letter-to-phoneme mapping was easily done, however to produce the correct phonetized version of a word, some modifications were needed to be done. The Letter-to-sound conversion algorithm written for this purpose yields an accuracy of 98.21%. The study carried out on stress assignment revealed that word level stress is fixed in the first syllable. Generally, the long vowels are also stressed.

1. Introduction

The stream of speech is composed of two kinds of phonological units: Segmental sounds and Suprasegmental sounds. Segmental sounds are those which can be segmented into distinct, discrete units, such as vowels and consonant [5]. The features of speech like variations of pitch, stress, and accents are called suprasegmental sounds. The naturalness of the synthesized speech is largely characterized by the fact that how successfully the above mentioned phonological units i.e. both segmental and suprasegmental features exists in the language are modeled in the speech synthesizer.

The research is focused on identifying unique characteristics of Sinhala language in terms of phonetics and phonology for improving the naturalness and intelligibility for our Sinhala TTS system.

Many researches have been carried out in the linguistic domain regarding Sinhala phonetics and phonology. These researches cover both aspects of segmental and suprasegmental features of spoken Sinhala. Some of the themes elaborated in these researches are; Sinhala sound change-rules, Sinhala Prosody, Duration and Stress, and Syllabification.

In the domain of Sinhala Speech Processing, a very little number of researches were reported and most of them were carried out by undergraduates for their final year projects. Some restricted domain Sinhala TTS systems developed by undergraduates are already working in a number of domains. The study of such projects laid the foundation for our project. However, an extensive research was done in all-aspects of Sinhala Phonetics and Phonology; in order identify the areas to be deeply concerned in developing a natural sounding Sinhala text-to-speech synthesizer.

The rest of the paper is organized as follows, Section 2 describes the adopted research methodology, Section 3 presents the Segmental feature of Sinhala, Section 4 presents the Suprasegmental features of Sinhala, Section 5 discusses the Sinhala Letter-to-Sound Conversion, and paper concludes with the results in Section 6.

2. Methods

Methodology adapted in this research included a critical review of literature, and further clarification of issues and difficulties encountered with distinguish scholars in the fields of linguistics and Sinhala.

The next step involved was the identification of topics which are more relevant to our TTS. Further Studying of some phonological aspects were irrelevant to our TTS system since it is a prose type text synthesizer. Some of the sections eliminated from our studies are: the different regional variations of speech, colloquial/classical Sinhala, and Sound-Change-Rules.

An extensive research on topics like Letter-To-Sound conversion, Syllabification of words, and Prosody (intonation, stress, phoneme durations, pitch, and accent) were carried out since they are vital to our TTS.

3. Segmental features of Sinhala

3.1. Sinhala phonemic inventory

3.1.1. Background. The Sinhala language is a member of the Indo-Aryan subfamily, which is a member of a still larger family of languages known as Indo-European. Sinhala is the official language of Sri Lanka and the mother tongue of the majority of the people constituting about 74% of its population. Sinhala language is presented in two major varieties: the Spoken and the Literary. They differ not only in their form and structure but also in their typical uses and functions. Literary Sinhala is generally considered the 'higher' variety in that its structure is closer to the classical literary idiom. It is used in all forms of non-fictional writing, including news bulletins, and in electronic media. News is read rather than spoken. Different genres of fiction use a mixture of both: literary Sinhala for narration and spoken Sinhala for dialog. Spoken Sinhala is used in all face-to-face communication.

3.2.2. Phonemes in Sinhala. Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants, including a set of 4 pre-nasalized voiced stops peculiar to Sinhala, as classified below in Table 1 and Table 2. [4]

Table 1: Spoken Sinhala consonant classification

		Labial	Dental	Alveolar	Retroflex	Palatal	Velar	glottal
Stops	**Voiceless**	p	t		ʈ		k	
	Voiced	b	d		ɖ		g	
Affricates	**Voiceless**					c		
	Voiced					ɟ		
Pre-nasalized voiced stops		ᵬ	ⁿd		ᶯɖ		ᵑg	
Nasals		m		n		ɲ	ŋ	
Trill				r				
Lateral				l				
Spirants		f	s			ʂ		h
Semivowels		ʋ				y		

Table 2: Spoken Sinhala vowel classification

	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:		

3.3.3. Phonotactics of Sinhala.

Diphthongs: Spoken Sinhala has following Diphthongs. In such diphthongs, the second vowel is always a high vowel. As the first vowel, all vowels, except the central vowel **/ə/**, occur, **/iu/, /eu/, /æu/ , /ou/ , /au/ , /ui/ , /ei/ , /æi/ , /oi/** , and **/ai/** [5].

Example:

/siurə/- *robe*
/kæɖei/- *may break*

Nasals: There are two nasalized vowels which occur in two or three words as single entity in Sinhala. They are either short or long: /ã/, /ã:/, /æ̃/, and /æ̃:/ [4]. But when a vowel is preceded by pre-nasalized

voiced stops in the same syllable, the other vowels also tend to be nasalized.

Example:

/kũːɓi/ - *ants*

/vĩɖinəvaː/ - *to suffer*

Pre-nasalized voiced stops, which is peculiar to Sinhala, share features of both voiced stops and nasals. Phonetically, pre-nasalized voiced stops are shorter than the nasals in duration.

Example:

/kanə/ - *ear*

/kaɖə/ - *trunk*

The duration of the pre-nasalized voiced stop **/ɖ/**, is much shorter than that of the corresponding nasal **/n/** [6].

Special features: Sinhala writing system does not provide a separate symbol for **/ə/**.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)*. With contrasting this fact, though there exists a letter in Sinhala writing system as (ඍ), which symbolizes the consonant sound**/ʄ/**, this is not consider as a phoneme in Sinhala.

4. Suprasegmental Features of Sinhala

In this section, Study of Sinhala Syllabification, and phonological aspects of Sinhala Prosody (sp. stress assignment, Intonation) are described.

4.1. Sinhala syllable structure and syllabification

4.1.1. Background. Words in the Sinhala language can be divided in to three groups namely *Nishpanna*, *Thadbhava* and *Thathsama* as described below:

Nishpanna: Words that have the local origin.
Thathsama: Words borrowed from other languages as they are.
Thadbhava: Words derived from other languages, mainly from Sanskrit and Pali.

The high impact of Sanskrit to Sinhala vocabulary is due to the fact that, Sinhala and Sanskrit both languages belong to the same Indo-Aryan language family. As the vehicle of Buddhism, Pali also has influenced the vocabulary of Sinhala. Due to various cultural, historical and sociolinguistic factors such as Tamil, as well as modern European languages such as Portuguese, Dutch and English has had a significant impact on the structure and vocabulary of Sinhala.

It is important to note that *Thathsama*, and *Thadbhava*, categories of words are available in Sinhala language as same as *Nishpanna* Category. Besides, for *Thathsama*, and *Thadbhava*, categories of words, the legal syllable structures or examples were not found in literature. This reveals that, a researcher who deals with Sinhala syllabification should essentially, study *Thathsama*, and *Thadbhava*, categories of words as well. This fact leads the research to study *Nishpanna* category separately as well as *Thathsama* and *Thadbhava*.

4.1.2. Syllabification of words belonging to the category of Nishpanna. It has been identified that there are four legal syllable structures as V, VC, CV and CVC for words which belongs to the category of Nishpanna [5]. This can also be represented as (C)V(C). Though a large number of examples for syllabified words belonging to each of the above structures are presented in literature [5], the methodology or grammatical rules describing how to syllabify a given word is not presented. A word can be syllabified in many ways retaining the permitted structures, but only a single correct combination of structures is accepted as the properly syllabified word.

Example:
A word having VCVCVC in its underlying consonant-vowel structure can be syllabified in following different ways, retaining the valid syllable structures described in the literature. (V)(CVC)(VC), (VC)(VC)(VC), (VC)(V)(CVC), however only one form represents the properly syllabified word.

Therefore, the determination of a proper mechanism leading to identification of the correct combination and the sequence of syllable structures in syllabifying a given word was the major challenge in this research.

When the literature is critically reviewed, the following model with regard to syllabification is identified:

1. It is identified that correct syllabified form of a word can be obtained by formulating a set of rules.
2. The following set of rules was identified.

Syllabification Procedure for Nishpanna Category:

a. Reach the first vowel of the given word and then,

1. If the phoneme next to this vowel is a consonant followed by another vowel, mark the syllable boundary just after the first vowel, i.e., word having a consonant-vowel structure of xVCV.. should be syllabified as (xV)(CV..), where x denotes either a single consonant or zero consonant.

2. If the phoneme next to this vowel is consonant followed by another consonant and a vowel, mark the syllable boundary just after the first consonant, i.e., word having a consonant-vowel structure of xVCCV.., should be syllabified as (xVC)(CV..), where x denotes either a single consonant or zero consonant.

3. If the phoneme next to this vowel is another vowel, mark the syllable boundary just after the first vowel, i.e., word having a consonant-vowel structure of xVV.., should be syllabified as (xV)(V..), where x denotes either a single consonant or zero consonant. Only few words were found in Sinhala having two consecutive vowels in a single word. eg. "**giriulltə**" *(Name of city.)*, "**aa:və** "*(Alphabet)*. Thus, the syllable structure (V) mostly occurs at the beginning of the word, except for above listed words.

b. Having marked the first syllable boundary, continue the same procedure for the rest phonemes as a new word, i.e., repeat the step (a) to the rest of the word, until the whole word is syllabified.

The accuracy was tested by working out with the examples given in the literature. Since results were consistent, it was concluded that above set of rules describes the accurate syllabification procedure for the words belonging to the category of *Nishpanna*.

4.1.3. Syllabification of Words Belonging to the Category of Thathsama and Thadbhava. In Syllabification, words borrowed from Sanskrit play a major role than the words borrowed from other languages. Reason behind is, distribution of large number of Sanskrit originated words and the complexity of codas and onsets of those words. Defining proper syllabic structures for words borrowed or either derived from Sanskrit had been focused preliminarily due to this reason. For this purpose, a carefully chosen list of related words[1] was presented to the distinguished scholars of Sinhala and Sanskrit for syllabification and recommendations. By carefully analyzing the gathered information and views, a set of rules have been identified on how to syllabify the borrowed Sanskrit. Critical analysis of gathered information and views, lead to a valid syllable structures followed by a set of rules to syllabify the borrowed Sanskrit words. Syllabic structures for those words can be represented as (C)(C)(C)V(C)(C)(C). It is also noteworthy to mention that, syllabic structures for words belong to the category of *Nishpanna* i.e. (C)V(C) is actually a subset of these structures.

Languages are unique in syllable structures. Syllabification of words belongs to the categories of *Thathsama* and *Thadbhava* does not completely adhere to syllabification rules imposed in the language from which the word is originated. Syllabification of such words will naturally be altered according to the ease of pronunciation and the existing syllable structures in Sinhala. This view was expressed by 100 % of the distinguished scholars to whom the chosen list of word was presented. As a proof of this fact, it was evident that the syllabification of all most all the words borrowed from languages other than Sanskrit (Pali, Tamil and English) are also consistent with the above determined set of syllabification rules and the syllabic structures.

Syllabification procedure for Thathsama and Thadbhava category:

a. Reach the first vowel of the given word and then,

1. If the vowel is preceded by a consonant cluster of 3, followed by a vowel,

   - If the consonant preceded by the last vowel is **/r/** or **/y/** then mark the syllable boundary after the first consonant of the consonant cluster, i.e., word having a consonant-vowel structure of xVCC[**/r/** or **/y/**]V.., should be syllabified as (xVC)(C[**/r/** or **/y/**]V..), where x denotes zero or any number of consonants.

   - In the above rule, if the consonant preceded by the last vowel is phoneme other than **/r/** or **/y/** then,
     – If the first two consonants in the consonant cluster are both stop

[1] See Appendix A: Word-List.

consonants, mark the syllable boundary after the first consonant of the consonant cluster.

i.e., word having a consonant-vowel structure of xV[C-Stop][C-Stop]CV.., then it should be syllabified as (xVC)(CCV..), where x denotes zero or any number of consonants.

- At other situations mark the syllable boundary after the second consonant of the consonant cluster.

i.e., word having a consonant-vowel structure of xVCCCV.., should be syllabified as (xVCC)(CV..), where x denotes zero or any number of consonants.

2. If this vowel is preceded by a consonant cluster of more than 3 consonants,
   - If the consonant just before the last vowel is **/r/** or **/y/** then, mark the syllable boundary before 2 consonants from the last vowel.
     i.e., word having a consonant-vowel structure of xVCCC…[**/r/** or **/y/**]V.., then it should be syllabified as (xVCCC…)(C[**/r/** or **/y/**]V..), where x denotes zero or any number of consonants.
   - At other situations, mark the syllable boundary just after the minimum sonorant consonant phoneme of the consonant cluster, i.e., word having a consonant-vowel structure of xVCCC…CV.., then it should be syllabified just next to the minimum sonorant consonant in the consonant cluster.

b. Having marked the first syllable boundary, continue the same procedure for the rest phonemes as a new word, i.e., repeat the step (a) to the rest of the word, until the whole word is syllabified.

The words with consonant clusters of more than 3 consonants are rarely found in Sinhala, and to avoid the confusions of syllabification of words in those situations, the algorithm makes use of the universal sonority hierarchy [2] in deciding the proper position to mark the syllable boundary.

4.1.4. Ambisyllabic words in Sinhala. Some ambisyllabic words are also found in Sinhala. This situation arises due to the fact that some words with complex coda or onset can be syllabified in several ways.

Example:

In a word such as **/sampre:kʃənə/** *(transmit)* the **/p/** can be interpreted as a coda with respect to the preceding vowel -**/samp/re:k/ʃə/nə/**or as an onset with respect to the following vowel - **/sam/pre:k/ʃə/nə/.**

More examples for this kind of words includes, **/mats/yə/, /mat/syə/***(fish);*

/sank/ya:/, /san/kya:/ *(number)* and **/lakʃ/yə/,/lak/ʃyə/***(point).*

Among the Sanskrit loan words in Sinhala including word internal clusters ending in /r/ and preceding a vowel, can either be syllabified by reduplicating the first consonant sound of the cluster or by retaining the original word as follows.

/krə/mak/rə/mə/yə/→ /krə/mak/krə/mə/yə/ (*gra dually)*

/ap/rə/ma:/nə/→ /ap/prə/ma:/nə/ (*unlimited)*

/ja/yag/ra:/hi:/→ /ja/yag/gra:/hi:/ (*victory)*

This explanation shows that, the determined rules and procedures comply even with the ambisyllabic words as well. It is important to note that one of the syllabified forms of such word will be provided in accordance with the determined rules.

4.1.5. The syllabification algorithm. In this section, the identified Sinhala syllabification linguistic rules are presented as an algorithm. The function "*syllabify*" accepts an array of phonemes generated by our Letter-To-Sound module for a particular word, along with a variable called *current_position* which is used to determine the position of the given array which is currently being processed by the algorithm.

Initially the *current_position* variable will hold the value 0. This function is called recursively until all phonemes in the array are processed. The function *mark_syllable_boundary*(*postion)* will mark the syllable boundaries of accepted array of phonemes. The other functions used within the "*syllabify*" function are described below.

no_of_vowels(phonemes): accepts an array of phonemes and returns the number of vowels contained in that array.

is_a_vowel(phoneme): accepts a phoneme and return true if the given phoneme is a vowel.

count_no_of_consonants_upto_next_vowel(phonemes, position): accepts an array of phonemes and a starting position; and returns the count of consonants from the starting position of the given array until next vowel is found.

is_a_stop(phoneme): returns true if the accepted phoneme is a stop consonant.

find_min_sonority_position(phonemes, position): returns the minimum sonorant position of given array

of phonemes, by start finding from the given position.
Complete listening of the algorithm is provided in the Appendix B: Syllabification Algorithm.

4.2. Sinhala prosodic features

4.2.1. Stress. Stress is a supra-segmental feature of utterances. It applies not to individual vowels and consonants but to whole syllables-whatever they might be. A stressed syllable is pronounced with a greater amount of energy than an unstressed syllable [1].
In Spoken Sinhala, usually the initial syllable receives the stress [7].
Example:
/kə́rənəva/ - *to do*
/aḿma/ - *mother*
/pótak/ - *a book*
Even the words can have more than one stress per word. In words of long vowels, then the stress falls even on the long vowel.
Example:
/há:muduruvó:/ - *monk*
/má:ligá:və/ - *castle*

Spoken Sinhala also has phrasal stress, with one stress per syllable of an entire phrase, to emphasize words. For this purpose, any word except a verb can be chosen for emphasis. In general, the verb does not receive the stress [5].

For example, if the sentence:
/ape: ta:tta sikura:da havəsə pansal yanəva/ is pronounced without stressing any of the words, simply means that our father goes to the temple on Friday evening. It is also possible to pronounce this sentence by placing stress on one of the words, except on the verb **/yanəva/**. If the word **/apé:/** stressed, it brings "our" into contrast with other fathers. The word **/ta:tta/** can be stressed to bring "father" into contrast with others. By placing stress on each word in the same sentence, it can be conveyed different kinds of information.

4.2.2 Pitch. Another prosodic feature is Pitch, defined as the frequency of vibration of vocal cords. Variations of the frequency of vibration are heard by the listener as variations of Pitch: the more frequently the vocal folds open and close, the higher the pitch [2].

Pitch in Spoken Sinhala helps not only to declare differences of meaning but also to identify certain subclasses of words; mainly related to the verbs. Those are Finite & Nonfinite participle and Imperative & Infinitive verbs [5].

Finite & nonfinite participle: The difference of those subclasses is indicated by the differences of pitch with which they are associated.
Example:
/amma gedərə ævilla/ (Mother has come home)
The finite participle which is a finite verb, is pronounced with a falling intonation.
/amma gedərə ævilla nida: gatta/ (Mother having come home and slept)
The nonfinite participle, which is a non-finite verb, uses a level intonation.

Imperative & Infinitive Verbs: These two subclasses of the verbs are also distinguished by their distinctive pitch patterns.
The imperative verb, which is a finite verb, uses a falling intonation:
/amma gedərə enɖə/ (Mother please come home)
The infinitive verb, which is a nonfinite verb, uses a level intonation:
/amma gedərə enɖə hadənəva/ (Mother tries to come home)

4.2.3 Intonation. If pitch varies over an entire phrase or sentence, the different pitch curves by the term intonation. Intonation conveys the speaker's attitude or feelings. In other words, intonation has a deictic function in discourse: questions; or a connotative function: anger, sarcasm, or various emotions. Intonation can also convey purely syntactic information, as when it marks where a sentence ends. Three basic forms of intonation can be illustrated in Spoken Sinhala [5].

Falling intonation: This simply shows a fall from a higher pitch to a lower pitch. Expressions with falling Intonation convey the meaning of finality.
/amma pansal gihilla/ (Mother has gone to the temple)

Rising intonation: This indicates the rise from a lower pitch to a higher pitch which conveys the meaning of unexpectedness and surprise.
/amma pansal gihilla/ (Mother has gone to the temple? / are you sure or has she gone?)

Level intonation: The maintenance of the same pitch levels illustrate by Level intonation and convey the meaning of non-finality.

/amma pansal gihilla.....ba:vəna: kərənəva/
(Mother has gone to the temple ….and meditates)

5. Letter-to-sound conversion

A letter-to-sound module is typically used in TTS system to generate pronunciations for those words not found in the lexicon. For many languages, there is a systematic relationship between the written form of a word and its pronunciation [3]. The precise definition of this relationship would lead to a successful letter-to-sound conversion module.
Sinhala is an orthographically phonemic language. Therefore, a set of rules can be easily formulated for letter-to-sound conversion.

5.1 The Sinhala character set

The Sinhala character set has 20 vowels and 40 consonants. There are 18 diacritics, 17 which denote the vowels and are known as vowel modifiers, and one(*hal lakuna* ්) for representing the pure consonant form.(i.e. without *hal lakuna,* the consonant is pronounced either associating it with schwa or the vowel "a" eg. ක් = k and ක = kə or ka).
The composition of a Consonant and Vowel is represented by either inserting or attaching appropriate diacritics (vowel modifier) to the Consonant. Even this composition does not occur for all C-V combinations.

Vowels:
අ , ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, ඍ, ඎ, එ, ඒ, ඓ, ඔ, ඕ, ඖ

Consonants:
ක, ඛ, ග, ඝ, ඩ, ඟ, ච, ඡ, 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

Vowel modifiers: ං ඃ, ා, ැ, ෑ, ෘ, ෲ, ෟ, ෳ, ෳ ,ෘෘ, ෙ, ේ,ෙෙ, ෛ, ෝ, ෞ, ෟ, ෘෘ, ෲ

eg. Some composition of C-V for consonant "ක":ක්, කා, කි, කී, කු

5.2. Letter-to-sound rules

A table was prepared mapping each letter with associated phoneme. (See Appendix C). Among the whole set of consonants, 10 consonants have their aspirated form as a separate character for each. But in spoken Sinhala, the non-aspirated sounds are generally preferred. Thus, it was decided to use the same phoneme to represent both aspirated and non-aspirated sounds of those consonants.

After tokenizing a word, its phonetic representation is obtained with the aid of above table. Unicode characters in a word are obtained character-by-character and converted to its phonetic representation.

The next step involves with applying a predetermined set of rules in a specific order to the phonetized letters. These rules are used to predict more accurate pronunciation of a word. The identified rules are as follows:

In the absent of a diacritic (except *hal lakuna*) for a particular consonant, the consonant should be associated with either schwa or vowel "a". Generally, in Sinhala words, the tendency of associating a schwa is high; therefore, such consonants are associated with schwa.

If the nucleus of the first syllable is a schwa, the schwa should be replaced by vowel "a", except in the situations where syllable starts with "s" followed by "v" phoneme, and phoneme "k" is preceded by schwa and "r" is followed by schwa.

If a consonant follows the phoneme sequence "r", schwa, and "h", then schwa should be replaced by vowel "a". If schwa is preceded by any consonant other that "h" then it retained [4].

If schwa is followed by phoneme "h", and any phonemes in the set { }, is preceded by phoneme "h", then schwa should replaced by "a" [4].

If schwa is preceded by a consonant cluster, it should be replaced by "a" phoneme [4].

If schwa is preceded by the last consonant of the word, it should be replaced by phoneme "a", except in the situations where the word final consonant is "r", "b", "d" and "t".

At the end of a word, if schwa precedes the phoneme sequence "i" and "y" then, it should be replaced by "a".

If the phoneme "k" is preceded by schwa, and the phoneme sequence "r" and "u" are followed by schwa then schwa should be replace by phoneme "a".
If a word starting with the phoneme sequence "k", "a", "l", "e", the vowel "a" in this sequence is changed to schwa.

Some words are inaccurately phonetized even after applying the above set of rules. Therefore, it

was decided to maintain a lexicon to obtain the accurate phonetized form of a word.

6. Results and conclusion

The fundamental approach of our Sinhala TTS system is diphone concatenation. The research on Sinhala Phonemic Inventory leads to the identification and preparation of a list of diphones. All together 1,401 diphones were identified, and some of the diphones were listed for the purpose of reading the foreign words appear in the text.

The extensive study carried out on Syllabification of Sinhala words leads to the definition of set of rules which is capable of predicting the syllable boundaries of a given word accurately. The algorithm was tested on 30,000 distinct words extracted from "UCSC Sinhala Corpus BETA" corpus, and compared the results with correctly syllabified words. Text obtained from the category "News Paper > Feature Auricles > Other " was chosen for testing the algorithm, due to the heterogeneous nature of the text and better representation of corpus (i.e. approx. 2/3 of the entire corpus). A list of distinct words was obtained, and first 30,000 words with the highest frequency of occurrence were chosen for testing the algorithm. The algorithm achieved an accuracy of 99.95%.
Most of the errors were due to,

1. Words developed by composing two or more words. (ie. Single word is formed by combining 2 or more distinct words; like the word "thereafter"). In this case, syllabification should be carried out separately for each word that the word is comprised of and then it should be concatenated to form a single word.

2. Directly termed other language words.

Letter-to-Sound rules were also tested with the same set of words, and compared with the correct phonetized representation of each word. The algorithm achieved an accuracy of 98.21%.

It was noticed that, here also the words developed by composing two or more words caused most of the errors. Beside, the contribution for errors of the directly termed foreign words is also a significant factor. Few errors were occurred due to the words which have the same written form but with different pronunciation, i.e. the written form "කර", may pronounced as /karə/ or /kərə/ depending on the context. Considering all these issues, it was decided to maintain a lexicon consist of such words that are most frequently occurred in text with their proper pronunciation form.

Unlike in English, Sinhala does not make a very clear distinction between stressed and unstressed syllables. In Sinhala, the position of the stress is fixed in relation to the word. Sinhala words nearly always have the stress on the first syllable, irrespective of the number of syllables in the word. Also, the long vowels in the words are stressed. The phrase stress in Sinhala cannot be easily predicted since it depends on the context. Hence, it was decided to place the stress on the first syllable and long vowels of each word by the stress assignment algorithm. A research is underway to determine the prediction of phase stress of an utterance.

The available literature on Sinhala Intonation is inadequate, and an extensive research on Sinhala prosody is underway. The research criterion includes the observation and analysis of selected previous news, broadcasted on radio with their scripts. Since, our TTS engine is also prose type reader, it is expected that this kind of study will reveal the necessary information to model Sinhala Prosodic features. Also, the study will be useful in determining the phrase breaks in an utterance too.

7. References

[1] P. Ladeforged, *A Course In Phonetics*, 3rd edn., Harcourt Brace Jovanovich College Publishers, 301, Commerce Street, Suite 3700, Fort Worth TX 76102, 1993.

[2] C. Gussenhoven, and H. Jacobs, *Understanding Phonology*, Oxford University Press Inc, 198, Madison Avenue, New York, NY 10016, 1998.

[3] A.W. Black and K.A. Lenzo, *Building Synthetic Voices*, Carnegie Melon University, Edinburgh, 2000.

[4] W.S. Karunatillake, *An introduction to spoken Sinhala*, 3rd ed., M.D. Gunasena & Co. ltd., 217, Olcott Mawatha, Colombo 11, 2004.

[5] J.B. Disanayaka, *The structure of spoken Sinhala*, National Institute of Education, Maharagama, 1991.

[6] R.M.W. Rajapaksa, *Verb: A Prosodic analysis*, Department of Phonetics and Linguistics, School of Oriental and African studies, University of London, 1988.

[7] M. Inman, *Duration and Stress in Sinhala*, Stanford University, 1986.

Appendix A: Word-List

කූටෝපක්‍රම්	k u: t o: p a k k r ə m ə
පෘතග්ජන	p r u t a g j a n ə
ක්‍රම්ක්‍රම්යෙන්	k r ə m a k k r ə m ə y e n
ස්ත්‍රීන්	s t r i: n
විද්‍යාඥයා	v i d y a: j ɲ ə y a:
කෘමියා	k r u m i y a:
පිතෘන්	p i: t t r u: n
සෞභාග්‍ය	s a u b a: g y ə
කාව්‍යෝපදේශය	k a: v y o: p ə d e: ʃ ə y ə
අවිද්‍යාව	a v i d y a: v ə
ප්‍රඥාව	p r a j ɲ a: v ə
කොන්ස්තන්තිනෝපලය	k o ŋ s t a n t i n o: p ə l ə y ə
ස්වප්න	s v a p n ə
ශල්‍යකර්ම	ʃ a l y ə k a r m ə
පාර්ලිමේන්තුව	p a: r l i m e: n t u w ə
ද්වන්ද්ව	d v a n d v ə
හෘදස්පන්දනය	h ə r d ə s p a n d ə n ə y ə
සම්ප්‍රේක්ෂණ	s a m p r e: k ʃ ə n ə
ක්ෂේත්‍ර	ʃ e: ʃ t r ə
ප්‍රවෘත්ති	p r ə v u r t i
ප්‍රව්‍රජ්‍යා	p r ə v u r j y a:
ප්‍රශ්ඛ්ධිය	p r ə ʃ r a b d i y ə
සංස්කෘත	s a n s k r u t ə
springs	s p r iŋ g s
scratched	s k r æ c ɖ
straights	s t r e: i t s
strength	s t r e n t s
postscript	p o: s t s k r i p t
area	e: r i a:

Appendix B: Syllabification Algorithm

function syllabify(*phonemes*, *current position*)

if no_of_vowels(*phonemes*) is 1 **then**
    mark_syllable_boundary(at the end of *phonemes)*
else
    if is_a_vowel(*phonemes*[*current position*]) is true **then**

    no_of_consonants=count_no_of_consonants_upto_next_vowel
        (*phonemes*, *current position*)

```
if no_of_consonants is 0 then
        if is_a_vowel(phonemes[current_position+1]) is true then
                mark_syllable_boundary (current_position+1)
                syllabify(phonemes, current_position+1)
        end if

else
        if no_of_consonants is 1 then
                mark_syllable_boundary(current_position+1)
                syllabify(phonemes, current_position+2)
        end if

        if no_of_consonants are 2 then
                mark_syllable_boundary(current_position+1)
                syllabify(phonemes, current_position+3)
        end if

        if no_of_consonants are 3 then
                if phonemes[current_position+3] is( "r" or "y") then
                        mark_syllable_boundary(current_position+1)
                        syllabify(phonemes, current_position+4)
                else
                        if is_a_stop(phonemes[current_posi+1]) is true and
                        is_a_stop(phonemes[posi+2])) is true then
                                mark_syllable_boundary
                                (current_position+1)
                                syllabify(phonemes, current_position+4)
                        else
                                mark_syllable_boundary
                                (current_position+2)
                                syllabify(phonemes, current_position+4)
                        end if
                end if
        end if
```

Appendix B: Syllabification algorithm continued..

```
        if no_of_consonants are greater than 3 then
                if phonemes[current_position+no_of_consonants]
                is( "r" or "y") then
                        mark_syllable_boundary
                        (current_position+no_of_consonants-2)
                        Syllabify
                        (phonemes, current_position+no_of_consonants-1)
                else
                        syllable_boundary=find_min_sonority_position
                        (phonemes,current_postion)
                        mark_syllable_boundary(syllable_boundary+1)
                        Syllabify
                        (phonemes, current_position+ no_of_consonants-1)

                end if

        end if
```

```
            end if

        else
        temp=current_postion
        repeat
                temp = temp + 1;
        until ( is_a_vowel(phonemes[temp]) is true
                syllabify(phonemes,temp)
        end if
end if
```

Appendix C: Mapping between Sinhala Letters and Phonemes

Table 1: Mapping between Sinhala letters and phonemes

Sinhala Letter	Corresponding Phoneme(s)	
	IPA	**Festival Equivalent[2]**
අ	**a**	a
ආ	**a:**	aaz
ඇ	**æ**	aez
ඈ	**æ:**	awz
ඉ	**i**	i
ඊ	**i:**	iiz
උ	**u**	u
ඌ	**u:**	uuz
ඍ	**/iru/**	r i
ඎ	**/iru:/**	r iiz
ඏ	**/ilu/**	i l u
ඐ	**/ilu:/**	i l uuz
එ	**e**	e
ඒ	**e:**	eez
ඓ	**ai**	aiz
ඔ	**o**	o
ඕ	**o:**	ooz
ඖ	**au**	auz
ක, ඛ	**k**	k
ග, ඝ	**g**	g
ඞ (ංං)	**ŋ**	q
ඟ	**g̃**	ngz
ච, ඡ	**c**	c
ජ, ඣ	**j**	j
ඤ	**ɲ**	cnz
ඥ	**/jɲ/**	j cnz
ඦ	**/nj/**	njz
ට, ඨ	**ʈ**	t
ඩ, ඪ	**ɖ**	d
ඬ	**ɖ̃**	ndz
ත, ථ	**t**	thz
ද, ධ	**d**	dhz
න, ණ	**n**	n
ඳ	**d̃**	ndz

[2] Festival : The framework used in Sinhala Text-to-Speech synthesizer is incompatible with IPA scheme, therefore an ASCII based scheme was proposed for representing the phonemes.

ප, ඵ	**p**	p
බ, භ	**b**	b
ම	**m**	m
ඹ	**ƀ**	mbz
ය	**y**	y
ර	**r**	r
ල, ළ	**l**	I
ව	**v**	v
ශ, ෂ	**ṡ**	shz
ස	**s**	s
හ, (ඃ)	**h**	h
ෆ	**f**	f