J. Sci. Univ. Kelaniya **3** (2007): 53-62 **Back to Contents**

AUTOMATIC SEGMENTATION OF GIVEN SET OF SINHALA TEXT INTO SYLLABLES FOR SPEECH SYNTHESIS

K.H. KUMARA[1], N.G.J. DIAS*[2] AND H. SIRISENA[3]

[1]Department of Mathematical Sciences, University of Wayamba, Kuliyapitiya, Sri Lanka

[2]Department of Statistics & Computer Science, University of Kelaniya, Kelaniya, Sri Lanka

[3]Department of Modern Languages, University of Kelaniya, Kelaniya, Sri Lanka

***Corresponding author (E-mail: ngjdias@kln.ac.lk)**

ABSTRACT

A dictionary based automatic syllabification tool has been given for Speech Synthesis in Sinhala language. This tool is also capable of providing frequency distributions of Vowels, Consonants and Syllables for a given set of Sinhala text.

A method of determining syllable boundaries has also been shown. Detection of Syllable boundaries for a given Sinhala sentence is achieved by four main phases and those phases have been described with examples. Rules for the automatic segmentation of words into syllables have been derived based on a dictionary. An algorithm has been produced for the implementation of these rules which utilizes the dictionary together with an accurate mark up of the syllable boundaries.

Keywords: Speech Synthesis, Text to Speech (TTS), Phonetic, Vowels, Consonants, Syllables, algorithm, transcription.

INTRODUCTION

In the present era of human computer interaction, the educationally under privileged and the rural communities of Sri Lanka are being deprived of technologies that pervade the growing interconnected web of computers and communications. One good solution for this problem would be computers talking to the common man in the language he is comfortable to communicate in. Sri Lankan population has a significant percentage of people who are educationally under-privileged. On one hand we claim that to build an E-Sri Lanka or an E-Society in Sri Lanka on the other hand, the advances we make are totally inaccessible by a large number of people in Sri Lanka. Under such circumstances, we cannot expect rural/educationally under-privileged people to use computers and IT products unless we remove the need of being literate in English, which exists as a barrier between them and computers. However, the interaction between the computer and the user is largely through keyboard and screen-oriented systems. In the current Sri Lankan context, this restricts the usage of computers to a miniscule fraction of the population, who are both computer-literate and conversant with written English. In order to enable a wider proportion of population to benefit from Information technology, there is a dire need for an interface other than keyboard and screen-interface that is widely in use at present. Speech technologies promise to be the next generation user interface. Software applications having speech and voice recognition abilities have a better chance to communicate with a large percentage of population which include educationally under-privileged, visually challenged and computer illiterates, if these applications can speak and understand the native language.

Text to Speech (TTS) or Speech Synthesis can be considered as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter. TTS software can "read" text from a document, Web page or e-Book and generating synthesized speech through a computer's speaker. The three basic steps that a syllable based TTS is required can be described as (1). Parse the text breaking it up into sentences, words, and punctuations; (2). For each word; use a dictionary to determine the syllable used to pronounce that word; (3). Use those syllables to extract recorded voice from a database, and concatenate those together to produce speech. It is

evident that the transcription of orthographic words into syllables is one of the principal steps of a syllable based TTS system. In such a system suitable syllable pronunciation must be supplied, without human intervention, for every word in the text. No dictionary, however large, will contain all words, let alone proper names, technical terms and other textual items commonly found in unrestricted texts. Consequently, an automatic syllabification is usually considered essential.

Sinhala is a language of around 17 million people living in Sri Lanka. The Sinhala language can be consider as a phonetic language. There is a clear distinction between two language forms in Sinhala, namely literary (written) and colloquial (spoken). Traditional literary Sinhala alphabet consists of 62 letters (Sirisena, 2003), but the modern literary Sinhala language consists of 12 vowels, 2 diphthongs and 39 consonants letters. To denote sound /ɸ/, the grapheme ඵ is used in the modern written Sinhala. The technique proposed for syllabification is based on the general sound system of the Sinhala language. It is done by two-step conversion process. First the orthographic word is divided into its syllables, and secondly the syllable sequence is converted to phonetic strings. The detail description of the algorithm is given in the section 4 bellow.

In this work first we were able to develop a $C_0^n V$ dictionary which represent the $C_0^n V$ definition of a single grapheme of the Sinhala alphabet and secondly based on the $C_0^n V$ dictionary , we were able to develop a Sinhala text transcriber which is capable of automatically generate the syllable sequence of a given set of Sinhala text according to the $C_0^n V\ C_0^n$ definition.

GENERAL INFORMATION ABOUT THE SINHALA SOUND SYSTEM

There is a clear difference between the two language forms in Sinhala: literary (written) Sinhala and colloquial (spoken) Sinhala. Traditional written Sinhala alphabet consists of 62 letters. To denote vowels in writing Sinhala, 14 letters are used. They include 12 monophthongs and 2 diphthongs (letters/ඍ/and/ඎ/, once designating vowels, are not used in the modern literary language). There are 39 letters for consonant single phonemes among which three phonemes/ව/, /ළ/, /ෆ/are written with

two letters. The letter ණ historically corresponds to phoneme /ɳ/ which is described as a retroflex nasal consonant. However in the modern Sinhala language this phoneme is pronounced as alveolar, i.e., it has merged with phoneme /ν / which is designated by the letter න. Thus, two different letters are now used to denote the same consonant (alveolar nasal).

The letter ළ historically corresponds to phoneme /→/ and is described as retroflex post alveolar. This phoneme has also merged with phoneme /l/ which is designated by letter ල. Letter ං corresponds to phoneme /□/ which is described as a backward consonant. The same phoneme/□/, is designated by letter ඏ which is not practically used in the modern literary language. Two phonemes /□◌/, /∝/ are written with combinations of letters කා, කෑ, and combinations of phonemes/□∝/and/к□/are written accordingly by combinations of letters කෑ, කෟ 44 letters are used for consonants (Sirisena, 2003)

The traditional alphabet had no grapheme for sound/f/and grapheme ෆ is used in modern written Sinhala. A particular feature of Sinhala is its spelling with no capital letters.

SYLLABLE DEFINITION

$C_0^n\ V\ C_0^n$ definition

Before embarking on the task of automatic syllable detection one has to decide what constitutes a syllable in the first place. Although there is no universal agreement on a rigorous definition of the syllable but one which has wide acceptance is consonants and vowels combine to make a syllable. In this work our syllable definition can be expressed as $C_0^n\ V\ C_0^n$ where C_0^n signifies 0 to n consonants (see Table 1) and V signifies a vowel (see Table 2) including two diphthongs ඔෟ /□ʊ/ and ඓ /□◌ෛ/. According to the above definition we can roughly say that a syllable can be considered as a vowel like sound together with some of the surrounding consonants that are most closely associated with it. Also it can be considered a syllable as having onset (an optional initial consonant or set of consonants C_0^n), followed by a vowel V, and followed by a coda (an optional final consonant or set of consonants C_0^n). Thus ස් [τ] is

the onset of syllable තුන් [τυν], while the න් [ν] is the coda. The task of breaking up a word into syllable is called syllabification.

Table 1 : Set of consonant sounds of the modern Sinhala alphabet, which is used to assemble $C_0^n V$ dictionary for the single graphemes of the Sinhala alphabet

ක් /κ/	ඛ /κ^{ɳ}/	ග් /ɣ/	ඝ් /ɣ^{ˡ}/	ඞ් /ŋ/	ඟ /^{ˡ}ɣ/
ච් /χ/	ඡ් /χ^{ɳ}/	ජ් /ʝ/	ඣ් /ʝ^{ˡ}/	ඤ් /ɲ/	ඦ් /^{n}ʝ/
ට් /□/	ඨ් /□^{ɳ}/	ඩ් /ɖ/	ඪ් /ɖ^{ˡ}/	ණ් /n/	ඬ් /^{n}ɖ/
ත් /τ/	ථ් /τ^{ɳ}/	ද් /δ/	ධ් /δ^{ˡ}/	න් /ν/	ඳ් /^{n}δ/
ප් /π/	ඵ් /π^{ɳ}/	බ් /β/	භ් /β^{ˡ}/	ම් /μ/	
ය් /φ/	ර් /ρ/	ල් /λ/	ව් /ν/	ළ් /l/	
ශ් /♣/	ෂ් /ʃ/	ස් /σ/	හ් /η/	ෆ් /⊕/	
ක්ෂ් /κʃ/	ඥ් /ɲɲ/	ෆ් /ɸ/	ං /ŋ/	ඔ /^{μ}β/	

Table 2 : Set of vowel sounds of the modern Sinhala alphabet, which is used to assemble $C_0^n V$ dictionary for the single graphemes of the Sinhala alphabet

ඊ /ιɔ/	ඉ /ɪ/	ඒ /ɛɔ/	එ /ɛ/	ඈ /—ɔ/	ඇ /—/
ඇ /□ɔ/	අ /□/	ඕ /oɔ/	ඔ /o/	ඌ /υɔ/	උ /υ/
ඓ /□ɪ/	ඖ /□υ/				

Although automatic syllabification algorithm exist, the problem is hard partly because there is no agreed-upon definition of syllable boundaries. It leads to have different number of syllables for a given word. As an example the Sinhala word අක්කා is usually divided as /ak.ka/ , according to our definition it as /□κ.κ□ɔ/ but the acoustic graph is as shown in Figure 1. According to the acoustic graph it is clear that the Sinhala word අක්කා can't be divided as /ak.ka/ or /□κ.κ□ɔ/.

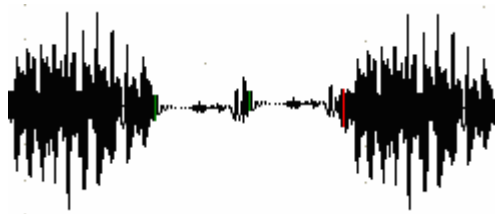

Figure 1. The acoustic graph of the Sinhala word /akka/

Why $C_0^n V C_0^n$ definition?

Using $C_0^n V C_0^n$ definition instead of $C_0^n V$ definition as syllable boundary will maximize the naturalness of the concatenated wave form for speech synthesis. The acoustic graph of $C_0^n V$ pattern syllable is shown in Figure 2 while the $C_0^n V C_0^n$ pattern syllable shown in Figure 3. It is clear that the both waveforms start with low pitch values but Figure 3 end with high pitch value and Figure 4 with low pitch value. Hence concatenating of two $C_0^n V$ pattern syllables generate more discontinuity in the resulting concatenated waveform than the discontinuity occurred in concatenating of two $C_0^n V C_0^n$ pattern syllables. It is clear that the $C_0^n V C_0^n$ definition provides more naturalness to resulting concatenated waveform than the $C_0^n V$ concatenation in the syllable base speech synthesis. On the other hand getting the $C_0^n V C_0^n$ definition as the syllable boundary in the automatic segmentation, gives much more similar syllable sequence which is produced by an expert manually. The n value of $C_0^n V$ units usually goes from 0 to 2 within a word when $C_0^n V$ appears in the left hand side of the vowel, in the right hand side n is 0 to 3 for Sinhala.

Figure 2. An example of acoustic graph of $C_0^n V$ pattern syllable, when n=1

Figure 3. An example of acoustic graph of $C_0^n V C_0^n$ pattern syllable, when n=1

AUTOMATIC SYLLABLE DETECTION

Generally, a TTS system uses a pronunciation dictionary for its pronunciation phase; one of the advantages of such a system being that the syllabification of words is recorded in a dictionary for speech synthesis. Additionally, use of a certain algorithm for the imposition of the required intonation and rhythm onto the concatenated waveform requires precise marking of the periods.

The $C_0^n V$ Dictionary

In this research the automatic syllable detection module uses a dictionary which represents the $C_0^n V$ definition of single graphemes of the alphabet for its automatic syllable identification phase. The implementation of $C_0^n V$ dictionary is based on the simplicity and consistency of ASCII but goes far beyond ASCII's limited ability to encode only the Latin alphabet. Hence it is impossible to represent/process some of International Phonetic Alphabet (IPA) symbols with some databases management systems (e.g. MS Access) and/or some integrated development environments (IDEs). Hence we provide the capacity to encode all of the IPA characters relevant to the Sinhala sound system by using a coding scheme. To keep character coding simple and efficient, we assign each sound (only for vowel sounds and consonant sounds in Sinhala) a unique numeric value and relevant IPA symbol. It has laid out provisions for encoding all scripts in the Sinhala.

One of the advantages of such a dictionary is that imposition of the required intonation and rhythm can be added into the dictionary to some extent. Developing such an algorithm is left for future research. Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) method might be trying out. One of another advantages of this approach in Sinhala is we can ignore the morphemic analysis phase in the speech synthesis. A Part of the dictionary which represents the $C_0^n V$ definition of single graphemes of the alphabet is given in Figure 4.

	ඇ /□/	ඇා /□ɔ/	ඇෘ /—/	ඇෲ /—ɔ/	ඉ /ɪ/	ඕ /ɪɔ/	

ක්/k/	ක	කා	කෑ	කෑ	කි	කී
ඛ්/kʰ/	ඛ	ඛා	ඛෑ	ඛෑ	ඛී	ඛී
ග්/ූ/	ග	ගා	ගෑ	ගෑ	ගි	ගී
ඝ්/ූ්/	ඝ	ඝා	ඝෑ	ඝෑ	ඝි	ඝී
ඞ්/ූ/	ඞ	ඞා				

Figure 4. A Part of the Dictionary which represents the $C_0^n V$ definition of single graphemes of the alphabet

Since Unicode is already playing a significant part with respect to localization and internationalization, using Unicode will be a better solution for the some of the above problems. Although it reduces the complexities of algorithms and the complexity of the $C_0^n V$ dictionary in the development phase, Unicoded Sinhala font hadn't been freely available in Sri Lanka at the beginning of this research. The conversion of existing algorithms and dictionary to the Unicode representation will not be a difficult task. Another advantage of using Unicode representation is, the ability to enable the TTS engine to recognize characters belongs to different languages (especially English alphabet and Latin alphabet) in the same document, because using separate encoding formats for two different languages would not support each other. This ability should be integrated with TTS engine to overcome the code switching problem generally occurred in the speech synthesis process.

Rules for detection of Syllable boundaries for a given sentence

It is assumed that an ASCII representation of each input Sinhala sentence is available as input to the text analysis module of the algorithm. First the input is analyzed in such a way as to reformat everything encountered (e.g., digits, abbreviations) into words and punctuation. Then parse the sentence into the automatic syllable detection module to detect the relevant set of syllables for a given word. Syllable detection module works according to the following algorithm. First it finds the $C_0^n V$ units for a given word and concatenates them. Then reformat $C_0^n V$ sequence

in order to obtain $C_0^n\ V\ C_0^n$ units by considering the consecutive syllable boundaries for a given word. In this phase it is possible to integrate the required prosody features by using statistical or rule based method. As an initial step we would like to proposed that the well known Bay's theorem (Sharman, 1994) and then some complex stochastic models.

As an example, the input ASCII string for a typical input sentence, shown below, was processed by the text analysis module of the algorithm to derive set of syllables together with their phonemes representation.

Input Text	ඕ තම පෞද්ගලික ප්‍රශ්න 3 නොතකා, ස්වකීය සංස්කෘතියට අනුව දූ පුතුන් වැඩුවා ය
Reformatted into words	ඕ තම පෞද්ගලික ප්‍රශ්න තුන නොතකා, ස්වකීය සංස්කෘතියට අනුව දූ පුතුන් වැඩුවා ය
Phonemic conversion of $C_0^n\ V$ Definition	ο⊃ϟ–τ□ϟ μ□ϟ–π□υϟ δ≀□ϟ λιϟ κ□ϟ–πραϟ♣ νϟ–τυϟ ν□ϟ–νοϟ τ□ϟ κ□⊃ϟ–σϖ□ϟ κι⊃ϟ ψ□ϟ–σ□ϟ |σκρ″ϟ τιϟ ψ□ϟ □□ϟ–□ϟ νυϟ ϖ□ϟ–δυ⊃ϟ–πυϟ τυνϟ–ϖ—ϟ⌊υϟ ϖ□⊃ϟ–ψ□ϟ
Phonemic conversion of $C_0^n\ V\ C_0^n$ Definition	ο⊃ϟ–τ□ϟ μ□ϟ–π□υδϟ ≀□ϟ λιϟ κ□ϟ–πρα♣ϟ ν□ϟ–τυϟ ν□ϟ–νοϟ τ□ϟ κ□⊃ϟ–σϖ□ϟ κι⊃ϟ ψ□ϟ–σ□ |σϟ κρ″ϟ τιϟ ψ□ϟ □□ϟ–□ϟ νυϟ ϖ□ϟ–δυ⊃ϟ–πυϟ τυνϟ–ϖ—ϟ⌊υϟ ϖ□⊃ϟ–ψ□ϟ

Note: ϟ Represent the syllable boundary and – represent the word boundary.

First, the word-formatting module transformed the numerals "3" into the Sinhala words "තුන". Next, each grapheme is compared with entries in the $C_0^n\ V$ dictionary. Then extract the consonants and vowel sound for each grapheme in the word. In this sentence, grapheme ක in the word පෞද්ගලික consists of consonant ක්/k/ and vowel sound ඇ/□/ which produce the syllable /k□/.

CONCLUSIONS

A method of determining syllable boundaries has been shown. Rules for the automatic segmentation of words into syllables have been derived based on a dictionary. An algorithm has been produced for the implementation of these rules which utilizes the dictionary together with an accurate mark up of the syllable

boundaries. Although not capable of doing a complete automatic segmentation of all words in the Sinhala language, we believe that about 95% of the words can now be automated (we have tested it in a limited domain). The method can be further improved by adding prosody features into the transcription, finding the semantically determined locations of contrastive and emphatic stress and assigning a (lexical) stress pattern to each word in the transcription at the runtime. The method leaves unsolved the treatment of unusual or idiosyncratic textual conversions, notations, and numeric information.

REFERENCES

Jurafsky D. & J. H. Martin 2004. Speech and Language Processing, Pearson Education Series.

Dutoit T. 1993. High Quality text to Speech Synthesis of the French Language, PhD dissertation, Faculte Polytechnique de Mons, TCTS Lab, 31 bvd Dolez, B-7000 Mons (Belgium).

Sirisena H. 2003. Phonetic properties of sound system of Sinhala Language as a basis for automatic transcriber, PhD Thesis 2003, Saint Petersburg State University, Russia.

Sharman R.A. 1994. Syllable-based Phonetic transcription by Maximum Likelihood Methods, International Conference On Computational Linguistics, Proceedings of the 15th conference on Computational linguistics - Volume 2, Kyoto, Japan Pages: 1279 – 1283.