A Rule Based Syllabification Algorithm for Sinhala

Ruvan Weerasinghe, Asanka Wasala, and Kumudu Gamage

Language Technology Research Laboratory, University of Colombo School of Computing, 35, Reid Avenue, Colombo 7, Sri Lanka
arw@ucsc.cmb.ac.lk, {awasala, kgamage}@webmail.cmb.ac.lk

Abstract. This paper presents a study of Sinhala syllable structure and an algorithm for identifying syllables in Sinhala words. After a thorough study of the Syllable structure and linguistic rules for syllabification of Sinhala words and a survey of the relevant literature, a set of rules was identified and implemented as a simple, easy-to-implement algorithm. The algorithm was tested using 30,000 distinct words obtained from a corpus and compared with the same words manually syllabified. The algorithm performs with 99.95 % accuracy.

1 Introduction

Syllabification algorithms are mainly used in text-to-speech (TTS) systems in producing natural sounding speech, and in speech recognizers in detecting out-of-vocabulary words. The key objectives of this study are to identify the syllable structures in modern Sinhala language and to define an algorithm to syllabify a given Sinhala word to be used in our TTS system. Syllabification algorithms have been proposed for different languages including English, German, Spanish and Hindi, among others. Although a few researchers have documented attempts at syllabifying modern Sinhala words in the Linguistics literature, this is the first known documented algorithm for Sinhala syllabification and certainly the first evaluation of any syllabification scheme for Sinhala.

Languages differ considerably in the syllable structures that they permit. For most languages, syllabification can be achieved by writing a set of declarative grammatical rules which explain the location of syllable boundaries of words step-by-step. It has been identified that most of these rules adhere to well known theories such as the Maximum Onset Principle and the Sonority Profile. The association of consonants with the syllable nucleus is derived by the Maximum Onset Principle (MOP).

Maximum Onset Principal: First make the onset as long as it legitimately can be; then form a legitimate coda [2].

Sonority Profile: The sonority of a syllable increases from the beginning of the syllable onward, and decreases from the beginning of the peak onwards [2].

Sonority is related to the acoustic intensity of a sound. Thus, by measuring the acoustic intensities of sounds, the sonority of a sound can be estimated [1]. The classes of vowel and consonant sounds (segments) that are usually distinguished along this dimension are listed in the order of increasing sonority, and this list is referred to as the Sonority Scale [2].
Sonority Scale:
Obstruents – Nasals – Liquids ([l, r] etc.) – Glides ([w, j]) etc. – Vowels

Many syllabification algorithms have been developed based on these two theories. For example, the Festival Speech Synthesis System (the framework we use in our Sinhala TTS) by default syllabifies words by finding the minimum sonorant position between vowels [3]. Another sonority scale based syllabification algorithm is presented in detail in [4]. The sonority theory of the syllable does not, however, account for all the phenomena observed in language. Many examples have been provided in the literature to demonstrate this [1], [2]. To avoid the difficulties encountered when using the sonority profile, most of the language specific syllabication schemes are modeled by using finite state machines or neural networks. A multilingual syllabification algorithm using weighted finite-state transducers has been proposed in [5].

In this research, an algorithm to divide Sinhala words into syllables is proposed. The algorithm was tested by using a text corpus containing representative words for each grammatical rule, and its performance was then measured in terms of the percentage of correctly syllabified words. The rest of this paper is organized as follows: Section 2 gives an overview of the Sinhala Phonemic Inventory and Section 3 briefly reviews the linguistic background of modern Sinhala word syllabification including issues we identified and our proposed solutions. Section 4 describes the implementation of the algorithm. The paper concludes with the results & discussion in Section 5.

2 The Sinhala Phonemic Inventory

Sinhala is one of the official languages of Sri Lanka and the mother tongue of the majority (about 74%) of its population. Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants as classified below in Table 1 and Table 2 [6].

There are two nasalized vowels occurring in two or three words in Sinhala. They are /a/, /a:/, /æ/, and /æ:/ [6]. Spoken Sinhala also contains the following Diphthongs, /iu/, /eu/, /œu/, /ou/, /au/, /ui/, /ei/, /æi/, /oi/, and /ai/ [7].

A separate letter for vowel /ə/ has not been 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 /kəə/ (to do.). In contrast to this, though the letter “ə” exists in Sinhala writing system (corresponding to the consonant sound /j/), it is not considered a phoneme in Sinhala.

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>High</td>
<td>i</td>
<td>i:</td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td>e</td>
<td>e:</td>
<td>ø</td>
</tr>
<tr>
<td>Low</td>
<td>æ</td>
<td>æ:</td>
<td>a</td>
</tr>
</tbody>
</table>

Table 1. Spoken Sinhala vowel classification