3.3 A tool for automatic segmentation of a given Sinhala text into Syllables for Speech synthesis and Speech recognition

K.H.Kumara,
Department of Mathematical Sciences,Faculty of Applied Sciences,
Wayamba University of Sri Lanka
N.G.J.Dias
Department of Statistics and Computer Science,Faculty of Science
University of Kelaniya
H.Sirisena
Department of Modern Languages,Faculty of Humanities
University of Kelaniya

ABSTRACT

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-Government 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, 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. It is well known that the transcription of orthographic words into syllables is one of the principal steps of a syllable based Speech synthesis and Speech recognition. Hence we put forward a dictionary based automatic syllabification tool for Speech Synthesis and Automatic Speech Recognition in Sinhala language. Also it is capable to provide the frequency distributions of Vowels, Consonants and Syllables of given Sinhala text. Although there is no universal agreement for syllable definition, in this research our syllable definition can be considered as $C_0^n\ V_1^n\ C_0^n$ where C_0^n signifies 0 to n consonants and V_1^n signifies 1 to n vowels. In this tool, detection of Syllable boundaries for a given Sinhala sentence is achieved by four main phases: (1) Reformat everything encountered (e.g. digits, abbreviations) into words and punctuation.

(2) Derive a phonemic representation for each word. (3) Determine the V_1^n units for a given word. (4) Reformat above $C_0^n V_1^n$ units according to the $C_0^n V_1^n C_0^n$ definition in order to obtain the syllable boundaries. Following example will give a better explanation of the algorithm.

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 $C_0^n V_1^n$ Definition	o:‖-ta‖ma‖-pau‖dga‖li‖ka‖-pra‖ʃna‖-tu‖na‖-no‖ta‖ka:‖ -sva‖ki:‖ya‖-sa‖ŋskɾ‖ti‖ya‖ʈa‖-a‖nu‖va‖-du:‖-pu‖tun‖- væ‖ɖu‖va:‖-ya‖
Phonemic conversion of $C_0^n V_1^n C_0^n$ Definition	o:‖-ta‖ma‖-paud‖ga‖li‖ka‖-praʃ‖na‖-tu‖na‖-no‖ta‖ka:‖ -sva‖ki:‖ya‖-saŋs‖kɾ‖ti‖ya‖ʈa‖-a‖nu‖va‖-du:‖-pu‖tun‖- væ‖ɖu‖va:‖-ya‖

Note: '‖' indicates the syllable boundary while '–'indicates the word boundary

Title: A tool for automatic segmentation of a given Sinhala text into Syllables for Speech synthesis and Speech recognition