(5) category 1

Unicode Sinhala and Phonetic English Bi-directional Conversion for Sinhala Speech Recognizer

M. Punchimudiyanse
Department of Mathematics and Computer Science
The Open University of Sri Lanka
Nawala, Sri Lanka
malinda@ou.ac.lk

R. G. N. Meegama
Department of Computer Science
University of Sri Jayawardenapura
Gangodawila, Sri Lanka
rgn@sci.sjp.ac.lk

***Abstract*—An automated speech recognizer (ASR) having a large vocabulary is yet to be developed for the Sinhala language because of the time consuming nature of gathering the training data to build a language model. The dictionary and building the language model require non-English text, in our case, Sinhala Unicode, to be transcribed in phonetic English text. Unlike text to speech conversions which only require transcribing the non-English text to phonetic English text, an ASR needs correct reproduction of the original language text when the phonetic English text is produced as the output of the speech recognizer. In the present research, newspaper articles are used to gather a large set of sentences to build a language model having thousands of words for the Sphinx ASR. We present a decoder algorithm that produces phonetic English text from Sinhala Unicode text and an encoder algorithm that produces the correct reproduction of Unicode Sinhala text from phonetic English. For a near phonetic tag set for Sinhala alphabet, results indicate 100% accuracy for the decoder algorithm while for numberless text, accuracy of the encoder algorithm stands at 98.61% for distinct phonetic English words.**

***Keywords*—Sinhala to Phonetic English, Phonetic English to Sinhala, Sinhala ASR, Sinhala Phonetic Tag set, Sphinx**

I. Introduction

Commercial and open source products to automatically recognize spoken English is available in the market today. In terms of performance and the accuracy, ASR systems that are based on acoustic and language models utilize hidden markov model (HMM) or Gaussian mixture models (GMM) for recognition. Most popular such open source ASR systems include Sphinx, HTK and Julius [1-3]. HMM based ASR systems typically have two phases, namely, the training phase and the recognition phase [16]. Onetime training phase is used in which a single or multiple user's voice is utilized to train a proper speech model called an acoustic model. In the recognition phase, the speech model built in the previous phase is recurrently used to recognize spoken words of the speaker.

The process of training a speech model requires a collection of acoustic data, a recorded voice sample of a person, and the correct phonetic text transcription of that voice sample. Adopting this technique for systems built for non-English languages require a production of non-English words in phonetic English text for the speech trainer. In order to build the phonetic text transcription from the original Unicode text, a proper phonetic tag per character is necessary.

Sinhala Unicode text to phonetic English tagging (syllabification) is not a straightforward task because Sinhala characters have several modifiers that are typed after the actual base character but appears before, after, top and the bottom of the actual base character. In addition, a non printable zero-width joiner (Unicode U+200D) is inserted between the characters to convert certain character combinations to one of the modifiers named yansaya, rakaranshaya or repaya which are not present in Sinhala Unicode.

For an ASR application to work, both Sinhala Unicode to phonetic text conversion and proper reproduction of Sinhala Unicode from phonetic English output of the speech recognizer is necessary. Phonetic tags for Sinhala alphabet has to be searched and replaced in a specific order to prevent invalid construction of words from phonetic English. For example, if phonetic tag for the characters ඉ is i and ී is ii with search order i, ii is used to encode a phonetic English word giithaya (meaning song), the word is constructed as (ග් + ඉ + ඉ + ත + ය) ගිඉතය which is misspelled. Correct spelling of the word ගීතය will be constructed if the search order ii,i is followed.

The ASR applications use N-gram based techniques to construct word combinations to build a language model in phonetic English. Therefore, it is necessary for the researchers to identify suitable phonetic tag set to develop a Sinhala to phonetic text decoder algorithm and phonetic English to Sinhala Unicode encoding algorithm.

II. Background and Related Work

Representation of Sinhala characters in digital form commenced in early 1980s to display Sinhala subtitles in television. The Wijesekara keyboard used in the type writers are extended to Sinhala fonts and keyboard drivers used in the computers. The Sinhala alphabet is first standardized by the Sri Lanka standards (SLS) institute as Sinhala Character Code for Information Interchange (SLS 1134:1996) and subsequently revised in 2004 and 2011 (SLS 1134:2004, SLS 1134:2011) [5]. The Sinhala alphabet representation in Unicode is first added to the Unicode standard version 3 based on the ISO/IEC 10646 [4].