NLP Applications of Sinhala: TTS & OCR

Ruvan Weerasinghe, Asanka Wasala, Dulip Herath and Viraj Welgama
Language Technology Research Laboratory,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 00700, Sri Lanka
{arw,raw,dlh,wvw}@ucsc.cmb.ac.lk

Abstract

This paper brings together the practical applications and the evaluation of the first Text-to-Speech (TTS) system for Sinhala using the Festival framework and an Optical Character Recognition system for Sinhala.

1 Introduction

Language Technology Research Laboratory [†] (LTRL) of the University of Colombo School of Computing (UCSC), was established in 2004 evolving from work engaged in by academics of the university since the early 1990's in local language computing in Sri Lanka.

Under the scope of the laboratory, numerous Natural Language Processing projects are being carried out with the relevant national bodies, international technology partners, local industry and the wider regional collaboration particularly within the PAN Localization Initiative[*]. The Sri Lankan component of the PAN Localization Project concentrated on developing some of the fundamental resources needed for language processing and some software tools for immediate deployment at the end of the project. Among the resources produced is a Sinhala Language Corpus of 10m words, and a tri-lingual Sinhala-English-Tamil lexicon. The two main software tools developed include a Sinhala Text-to-Speech (TTS) system and an Optical Character Recognition (OCR) system for recognizing commonly used Sinhala publications.

This paper focuses primarily on the end-user applications developed under the above project; Sinhala TTS system and OCR system. The paper describes the practical applications of these tools and evaluates it in the light of experience gained so far.

The rest of this paper is organized as follows: Section 2 gives an overview of the Sinhala TTS system; Section 3 describes the Sinhala OCR system. A summary along with future research directions and improvements are discussed in the last section.

2 Sinhala Text-to-Speech System

Sighted computer users spend a lot of time reading items on-screen to do their regular tasks such as checking email, fill out spreadsheets, gather information from internet, prepare and edit documents, and much more. However visually impaired people cannot perform these tasks without an assistance from other, or without using assistive technologies.

A TTS (text-to-speech) system takes computer text and converts the words into audible speech (Dutoit, 1997). With a TTS engine, application, and basic computer hardware, one can listen to computer text instead of reading it. A Screen Reader (2007) is a piece of software that attempts to identify and read-aloud what is being displayed on the screen. The screen reader reads aloud text within a document, and it also reads aloud information within dialog boxes and error messages. In other words, the primary function of any-screen reading system is to become the "eye" of the visually impaired computer user. These technologies enable blind or visually impaired people to do things that they could not perform before by them-

[†] See website: http://www.ucsc.cmb.ac.lk/ltrl
[*] See project website: http://www.panl10n.net

selves. As such, text-to-speech synthesizers make information accessible to the print disabled.

Within Sri Lanka, there is a great demand for a TTS system in local languages, particularly a screen reader or web browser for visually impaired people. In the case of the Tamil language, work done in India could be used directly. Until the LTRL of UCSC initiatives were launched in 2004, there was no viable TTS system found developed for Sinhala, the mother tongue of 74 % Sri Lankans (Karunatillake, 2004).

A project was launched to develop a 'commercial grade' Sinhala text-to-speech system in UCSC in year 2004. Later, it was extended to develop a Screen Reader which can be used by visually impaired persons for reading Sinhala texts.

The Sinhala TTS system was implemented based on the Festival speech synthesizer (Taylor et al., 1998). The Festival speech synthesis system is an open-source, stable and portable multilingual speech synthesis system developed at Center for Speech Technology Research (CSTR), University of Edinburgh (Taylor et al., 1998, Black and Lenzo, 2003). TTS systems have been developed using the Festival framework for different languages, including English, Japanese, Welsh, Turkish, Hindi, and Telugu (Black and Lenzo, 2003). However, efforts are still continuing to develop a standard Sinhala speech synthesizer in Sri Lanka.

The Sinhala text-to-speech system is developed based on the diphone concatenation approach. Construction of a diphone database and implementation of the natural language processing modules were key research areas explored in this project. In this exercise, 1413 diphones were determined. The diphones were extracted from nonsensical words, and recordings were carried out in a professional studio. Moreover, language specific scripts (phone, lexicon, tokenization) and speaker specific scripts (duration and intonation) were defined for Sinhala. It is worthy to mention the development of context-sensitive letter-to-sound conversion rule set for Sinhala. Incorporation of a high accuracy native syllabification routine (Weerasinghe et al., 2005) and implementation of comprehensive text analysis facilities (capable of producing the accurate pronunciation of the elements such as numbers, currency symbols, ratios, percentages, abbreviations, Roman numerals, time expressions, number ranges, telephone numbers, email addresses, English letters and various other symbols) have been found unique for the language (Weerasinghe et al., 2007). Despite the Festival's incomplete support for UTF-8, the above rules were re-writtenin UTF-8 multi-byte format following the work done for Telugu language (Kamisetty, 2006).

The current Sinhala TTS engine accepts Sinhala Unicode text and converts it into Speech. A male voice has been incorporated. Moreover, the system has been engineered to be used in deferent platforms, operating systems (i.e. Linux and Windows) and by different software applications (Weerasinghe et al., 2007).

2.1 Applications of TTS Synthesis Engine

Sinhala text is made accessible via two interfaces, by the TTS engine. A standalone software named *"Katha Baha"* primarily reads documents in Sinhala Unicode text format aloud. The same application can also be used to record the synthesized speech.

In this way, local language news papers and text books can be easily transformed into audio materials such as CDs. This software provides a convenient way to disseminate up-to-date news and information for the print disabled. e.g. Newspaper company may podcast their news paper, enabling access for print disabled and everyone else. Furthermore, the same application can be utilized to produce Sinhala digital talking books. To ensure the easy access by print disabled, keyboard short cuts are provided.

Owing to the prevalent use of Windows among the visually impaired community in Sri Lanka, it becomes essential that a system is developed within the Windows environment which offers Sinhala speech synthesis to existing applications. The standard speech synthesis and recognition interface in Microsoft Windows is the Microsoft Speech Application Programming Interface (MS-SAPI) (Microsoft Corporation, n.d.). MS-SAPI enabled applications can make use of any MS-SAPI enabled voice that has been installed in Windows. Therefore, steps were taken to integrate Sinhala voice into MS-SAPI. As a result, the MS-SAPI compliant Sinhala voice is accessible via any speech enabled Windows application. The Sinhala voice is proved to work well with "Thunder"[‡] a freely available screen reader for Windows. Additionally, steps were taken to translate and integrate

[‡] Available from: http://www.screenreader.net/

common words found related to Thunder screen reader (e.g. link="සබැඳිය", list item= "ලැයිස්තු අයිතම") (Weerasinghe et al., 2007).

Since most Linux distributions now come with Festival pre-installed, the integration of Sinhala voice in such platforms is very convenient. Furthermore, the Sinhala voice developed here was made accessible to GNOME-Orca and Gnopernicus - powerful assistive screen reader software for people with visual impairments.

It is noteworthy to mention that for the first time in Sri Lankan history, the print disabled community will be able to use computers in their local languages by using the current Sinhala text-to-speech system.

2.2 Evaluation of the Text-to-Speech Synthesis Engine

Text-to-speech systems have been compared and evaluated with respect to intelligibility (understandability of speech), naturalness, and suitability for used application (Lemmetty, 1999). As the Sinhala TTS system is a general-purpose synthesizer, a decision was made to evaluate it under the intelligibility criterion. Specially, the TTS system is intended to be used with screen reader software by visually impaired people. Therefore, intelligibility is a more important feature than the naturalness.

A Modified Rhyme Test (MRT) (Lemmetty, 1999), was designed to test the Sinhala TTS system. The test consists of 50 sets of 6 one or two syllable words which makes a total set of 300 words. The words are chosen to evaluate phonetic characteristics such as voicing, nasality, sibilation, and consonant germination. Out of 50 sets, 20 sets were selected for each listener. The set of 6 words is played one at the time and the listener marks the synthesized word. The overall intelligibility of the system measured from 20 listeners was found to be 71.5% (Weerasinghe et al., 2007).

3 Optical Character Recognition System

Optical Character Recognition (OCR) technology is used to convert information available in the printed form into machine editable electronic text form through a process of image capture, processing and recognition (Optical Character Recognition, 2007).

There are three essential elements to OCR technology. Scanning – acquisition of printed documents as optical images using a device such as flatbed scanner. Recognition- involves converting these images to character streams representing letters of recognized words and the final element involves accessing or storing the converted text.

Many OCR systems have been developed for recognizing Latin characters (Weerasinghe et al., 2006). Some OCR systems have been reported to have a very high accuracy and most of such systems are commercial products. Leaving a land mark, a Sinhala OCR system has been developed at UCSC (Weerasinghe et al., 2006).

Artificial Neural Network (ANN) and Template Matching are two popular and widely used algorithms for optical character recognition. However, the application of above algorithms to a highly inflected languages such as Sinhala is arduous due to the high number of input classes. Empirical estimation of least number of input classes needed for training a neural net for Sinhala character recognition suggested about 400 classes (Weerasinghe et al., 2006). Therefore, less-complicated K-nearest neighbor algorithm (KNN) was employed for the purpose of Sinhala character recognition.

The current OCR system is the first ever reported OCR system for Sinhala and is capable of recognizing printed Sinhala letters typed using widely used fonts in the publishing industry. The recognized content is presented as editable Sinhala Unicode text file (Weerasinghe et al., 2006).

A large volume of information is available in the printed form. The current OCR system will expedite the process of digitizing this information. Moreover, the information available via printed medium is inaccessible to the print disabled, and the OCR system, especially when coupled with Sinhala TTS, will provide access to these information for the print disabled.

3.1 Evaluation of the Optical Character Recognition System

The performance of the Sinhala OCR system has been evaluated using 18000 sample characters for Sinhala. These characters have been extracted from various books and newspapers (Weerasinghe et al., 2006). Performance of the system has been evaluated with respect to different best supportive fonts. The results have been summarized in the Table 1 (Weerasinghe et al., 2006).

Font	*FM*	*DL*	*Lakbima*	*Letter*
% Recog.	97.17	96.26	89.89	95.81

Table 1. Experimental Results of Classification*

From this evaluation it can be concluded that the current Sinhala OCR has average accuracy of 95% (Weerasinghe et al., 2006).

4 Conclusion and Future Work

This paper brings together the development of a diphone voice for Sinhala based on the Festival speech synthesis system and an Optical Character Recognizer for Sinhala.

Future work on the Sinhala TTS engine will mainly focus on improving the prosody modules. A speech corpus containing 2 hours of speech has been already recorded. The material is currently being segmented, and labeled. We are also planning to improve the duration model using the data obtained from the annotated speech corpus. It is also expected to develop a female voice in near future. The current Sinhala OCR system is font dependent. Work is in progress to make the OCR system font independent and to improve the accuracy. Sinhala OCR and the TTS systems, which are currently two separate applications, will be integrated enabling the user friendliness to the print disabled.

A number of other ongoing projects are aimed at developing resources and tools such as a POS tag set, a POS tagger and a tagged corpus for Sinhala, an on-the-fly web page translator, a translation memory application and several language teaching -learning resources for Sinhala, Tamil and English.

All resources developed under this project are made available (under GNU General Public License) through the LTRL website.

Acknowledgement

This work was made possible through the PAN Localization Project, (http://www.PANL10n.net) a grant from the International Development Research Center (IDRC), Ottawa, Canada, administered through the Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan.

* FM – "FM Abhaya", DL – "DL Manel Bold", Letter – "Letter Press"

References

Alan W. Black and Kevin A. Lenzo. 2003. *Building Synthetic Voices*, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from http://festvox.org/bsv/.

Microsoft Corporation. (n.d.). *Microsoft Speech SDK Version 5.1*. Retrieved from:

http://msdn2.microsoft.com/en-/library/ms990097.aspx

T. Dutoit. 1997. *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, Netherlands.

C. Kamisetty, S.M. Adapa. 2006. *Telugu Festival Text-to-Speech System*. Retrieved from:

http://festival-te.sourceforge.net/wiki/Main_Page

W.S. Karunatillake. 2004. *An Introduction to Spoken Sinhala, 3rd edn.*, M.D. Gunasena & Co. ltd., 217, Olcott Mawatha, Colombo 11.

Sami Lemmetty. 1999. *Review of Speech Synthesis Technology*, MSc. thesis, Helsinki University of Technology.

Screen Reader. 2007. *Screen Reader*. Retrieved from: http://en.wikipedia.org/wiki/Screen_reader.

Optical Character Recognition. 2007. *Optical Character Recognition.* Retrieved from: http://en.wikipedia.org/wiki/Optical_character_recognition

P.A Taylor, A.W. Black, R.J. Caley. 1998. The Architecture of the Festival Speech Synthesis System, *Third ESCA Workshop in Speech Synthesis*, Jenolan Caves, Australia. 147-151.

Ruvan Weerasinghe, Asanka Wasala, Kumudu Gamage. 2005. A Rule Based Syllabification Algorithm for Sinhala, *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*. Jeju Island, Korea. 438-449.

Ruvan Weerasinghe, Dulip Lakmal Herath, N.P.K. Medagoda. 2006. A KNN based Algorithm for Printed Sinhala Character Recognition, *Proceedings of 8th International Information Technology Conference*, Colombo, Sri Lanka

Ruvan Weerasinghe, Asanka Wasala, Viraj Welgama and Kumudu Gamage. 2007. Festival-si: A Sinhala Text-to-Speech System, *Proceedings of 10th International Conference on Text, Speech and Dialogue (TSD 2007)*, Pilseň, Czech Republic, September 3-7, 2007. 472-479