Developing a Speech Corpus for Sinhala Speech Recognition

Thilini Nadungodage
Language Technology Research Laboratory,
University of Colombo School of Computing,
Colombo 00700
hnd@ucsc.lk

Viraj Welgama
Language Technology Research Laboratory,
University of Colombo School of Computing,
Colombo 00700
wvw@ucsc.lk

Ruvan Weerasinghe
Language Technology Research Laboratory,
University of Colombo School of Computing,
Colombo 00700
arw@ucsc.lk

Abstract

Speech corpora is a main part of statistical model based speech recognition research and highly affect the performance of a speech recognizer. Some effort should be given to build a good speech corpus for low-resourced languages such as Sinhala. Sinhala language suffers from the lack of proper speech corpora for speech recognition research. In this paper we present our effort on designing and developing a speech corpus for the Sinhala language as it would facilitate future research on Sinhala speech recognition.

1 Introduction

Speech recognition is a very active research area, which tries to provide a friendly environment for human-computer interaction. Research on speech recognition has been carried out over last few decades and currently it has matured to a stage where it can be productively used in many practical applications (Furui, 2005). Automatic Speech Recognition (ASR) has successfully been applied to many languages so far, mainly in Latin languages. Most of the modern ASR systems use statistical models trained using corpora of relevant speech (Barnard et al., 2009). Such systems highly rely on the comprehensiveness of the relevant speech corpora.

Experiments have shown that performance of ASR systems mostly depend on the characteristics of their training corpora. It is said that better performance can be gained by using domain specific data for different applications rather than using a general data set for all applications.

Speech corpus is a database of speech audio files and text transcriptions of them, in a format that can be used to create Acoustical Models using speech recognition engines. The first step of building an ASR system is building the speech corpus. For languages such as English, French, German and many others, this step is almost not relevant because they have decades of collected speech data and successfully built speech corpora for experiments (Cucu et al., 2012). For low-resourced languages including most of Non-Latin languages, it has to build speech data sets from the scratch since most of them do not have pre-created speech corpora.

There are examples in literature where attempts have been made to develop speech corpora for low-resourced languages. For example International Institute of Information Technology, India has developed speech corpus for three Indian languages: Tamil, Telugu and Marathi (Anumanchipalli et al., 2005). The National Language Research Institute of Japan presents a spontaneous speech corpus of Japanese in (Maekawa et al., 2000). Development of a Thai speech corpus is presented in (Kasuriya et al., 2003) and building a Romanian speech corpora is presented in (Cucu et al., 2012).

Our goal in this paper is to present the steps we followed and the experience we gained on designing and developing a speech corpus as a first step of building comprehensive speech corpora for the Sinhala language. Sinhala is an Indo-Aryan language spoken by more than 16 million people in Sri Lanka. Sinhala is a highly inflectional language as are many other Indic languages, and like many of them, can be considered as a low-resourced language with respect to the linguistic resources available for NLP. Lack of proper data is the main problem we face when it comes to speech recognition research in the Sinhala Language. Our speech corpus is expected to be used in developing a speech recognition system for mobile phone music requesting application. We hope that this work will also be helpful in future research on Sinhala speech recognition.

2 Speech Corpus Building

Speech corpus should be designed in a way that it contains phonetic events (phones, di-phones, tri-phones) according to their frequency of occurrence in natural speech. This type of data set is defined as *phonetically-balanced* data set (Radova and Vopalka, 1999). Audio data of a speech corpus should represent the users of the applications that built using that corpus. Parameters such as number of different voices, speaker genders and age groups among others are considered as important factors of audio data.

A Speech corpus can be developed mainly in two ways. One way is to collect existing speech data (speech that is already been recorded) and manually transcribe them in to text. This is a very time consuming and tedious task which requires a lot of resources. The second way is to design the text corpus first and record the speech by reading the collected text. We used the second approach to build our speech corpus.

2.1 Text Corpus Collection

Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants (Wasala et al., 2006). To collect a phonetically-balanced data set we use the UCSC[1] 10M Words Sinhala Corpus (Weerasinghe et al., 2007) to extract phonetically-balanced sentences and phrases. Since our main target is to build a music request application, we included the vocabulary of Sinhala music domain to create those sentences/phrases. However, to make the speech corpus more general, we added some other fields such as commonly used words like yes/no, digits and standard built-in items (number, dates and times). Following section gives a brief description of the data fields we used to create the text corpus.

2.1.1 Data Fields

Our text corpus consists of 12 different types of data fields, namely; phonetically balanced sentences, keywords, currency, boolean data, digits, dates, times, music genres, proper names, songs, numbers and spontaneous questions.

- Phonetically Balanced Sentences
  A set of 2000 phonetically-balanced sentences. Initially, a set of simple sentences in the length of 8-12 words were selected and calculated its phoneme distribution. Then, missing phonemes and low frequency phonemes were identified by analyzing the phoneme distribution. We created a threshold in a way that the rarest phoneme should atleast have a frequency of 50 occurrences in the sentences. After identifying the low frequency phonemes we searched the text corpus for sentences which contain these phonemes and added them to the selected sentence set until the phoneme coverage is balanced.

- Keywords
  Keywords used in the music application domain. We got a list of English keywords which are used in the music domain such as 'song', 'play', 'pause' and translated them into Sinhala.

- Currency
  A list of phrases that shows how the currency terms are spoken in the Sinhala language. This list included 400 different currency terms in the range of one Rupee to 900,000 Rupees including phrases with cents.

- Boolean Data
  A list of words that represent various usages of 'yes' and 'no' in Sinhala.

- Digits
  A list of phrases that shows how strings of digits are spoken in Sinhala. We collected digit strings under three types as credit card numbers, telephone numbers and randomly generated 7 digit strings. Our list included 1000 credit card numbers 500 telephone numbers and 500 random generated strings representing various styles of speech.

- Dates
  A list of phrases including 500 randomly generated dates with various years, months and dates.

- Times
  A list of phrases including 500 randomly generated times with hours and minutes.

- Music Genres
  A list of various music genres in Sri Lanka including traditional music and music with western influence.

[1] University of Colombo School of Computing, Sri Lanka

- Names
  A list of various names including artists names, local cities, international cities and proper names.
- Songs
  A list of phrases which each phrase includes few words from a song to identify the song. This included 2000 Sinhala songs and 1500 English songs.
- Numbers
  A list of phrases that shows how numbers are spoken in Sinhala. This list included numbers from 10-99 and 500 random numbers from 100-9999.
- Spontaneous Questions
  A list of questions which we expected to get various answers. Some of the questions were like: 'What is your name?'. 'What is your age?', 'Who is your favorite artist?'.

2.1.2 Caller Sheet Preparation

Using the data collected to design the text corpus, caller sheets were prepared in a predefined format. A caller sheet is the document we supply to the speakers to read. It consists of all the data types in various proportions. First part of each caller sheet was started with 10 spontaneous questions which we expected answers from the speaker. The second part we expected the speaker to read aloud. This part contained prompts for the speaker to read 5 phonetically balanced sentences, 7 keywords, 2 currency items, 4 digit strings, 1 date, 1 time, 1 music genre, 5 names, 8 songs and 2 numbers which added up to 36 items. A caller sheet altogether consisted with 46 items which was the number of utterances we expected to collect from one speaker. Our target was to collect speech from 2000 speakers. Hence 2000 caller sheets were created by randomly selecting items for each caller sheet from the text corpus.

2.2 Audio corpus

After creating the text corpus and caller sheets next task was to collect audio data. We recorded telephone calls from selected speakers who read the given caller sheets and stored them in .wav format.

2.2.1 Speaker Distribution

When selecting speakers main concern was given to the age of the speakers. Since music requesting over mobile phones is most popular among young people, more proportion of speakers were selected in the age group of 16-40 years. Our expected age distribution of speakers is shown in Table 1. Our target was to select equal number of speakers from males and females for the gender distribution.

16-25 (%)		**26-40** (%)		**41-60** (%)		**<16,>60** (%)	
30		40		20		10	
M	F	M	F	M	F	M	F
15	15	20	20	10	10	5	5

Table 1: Expected speaker distribution according to age levels and genders

We selected a group of university students as our primary speakers. Initially they recorded their own voices by reading the given caller sheets and then they were given the instructions to select speakers from their hometowns and among their relatives. By following this procedure we were able to collect speakers from different age levels, education levels and different dialects from all over the country. Although our expectation was to select speakers in the given proportions in Table 1, practical reasons lead us not to strictly depend on these measures. Table 2 shows the approximate values of the actual speaker distribution according to age groups.

16-25 (%)	**26-40** (%)	**41-60** (%)	**<16,>60** (%)
50	24	18	8

Table 2: Actual speaker distribution according to age levels

2.2.2 Recording

We used mobile phones to record voices. We provided a mobile phone to the speaker with a fixed number to call, where it directed to the sever which the recording is setup. Each speaker was given instructions prior to making the call on how to make and proceed with the call. Also they were instructed to go to a location where no too much noise to make the call. When the call was made, pre-recorded instructions were played to guide the speakers throughout the caller sheets.

3 Corpus Cleaning and Noise Tagging

In a speech corpus, audio data and text data should be aligned. So we manually listened to each recorded sentence or phrase and checked them with their corresponding text transcription and made necessary corrections. For the spontaneous questions we wrote down the answers of the speakers.

We used separate tags to identify the noises recorded with human voices. Six different noise tags were used to distinguish between different noises.

- Speech
  A word mispronounced by the speaker which does not has any meaning.
- Filler
  A sound or word that is spoken in between an utterance by the speaker when he/she pauses to think. Sounds like 'um', 'ah', 'er', 'mm'.
- Speaker Noise
  Sounds made by speaker other than words, like clearing throat or exhaling.
- Dtmf
  Dual Tone Multiple Frequency tones. Sounds made by keypad pressing in mobile phones.
- Int
  Background sounds with high intensity like vehicle hone sound and sound of a door slamming.
- Noise
  Other background noises which does not fall into the above categories.

Utterances with lot of noises and utterances that are unclear were removed to improve the quality of the speech corpus.

4 Corpus Statistics

Our final speech corpus was consisted with 78,667 utterances which included altogether 65 hours of speech. To measure the quality of the built speech corpus here we present some of the statistical features of it.

4.1 Basic Statistics

In Table 3 we give the figures of basic statistics such as; # of words, n-grams, phone and di-phone counts of the speech corpus

Attribute	Value
# of Phrases	78,667
# of Words	307,756
Uni-grams	13,204
Bi-grams	43,729
Tri-grams	41,185
# of phonemes	1,709,326
# of di-phones	1,631,517

Table 3: Basic statistical measures of the Sinhala speech corpus

4.2 Phonetic Distribution

To show the phonetic distribution of the built corpus we calculated phone frequencies occurred in it. Most frequent phone has occurred 253,229 times and least frequent phone has occurred 82 times. Figure 1 shows this phonetic distribution as a graph of phone frequencies against Sinhala phones.

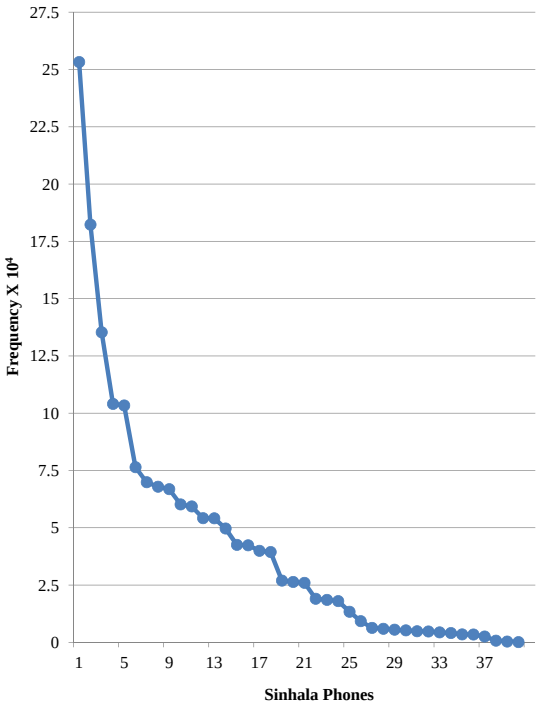

Figure 1: Phonetic distribution of the built corpus as a graph of phone frequencies against Sinhala phones.

4.3 Language Model Perplexity

Perplexity of a language model is a measure which indicates how good is the language model. Using

the text corpus of the developed speech corpus, we built a language model and measured its perplexity using a pre-defined test data set.

We randomly selected 73,802 sentences/phrases from the text corpus as the training set of the language model. Remaining 4,865 sentences/phrases we considered as the test data set. We used the SRILM language modeling toolkit developed by the SRI International Speech Technology and Research Laboratory (Stolcke and others, 2002) to develop the language model and measure its perplexity. The results are shown in Table 4.

Attribute	**Value**
# of sentences	4,865
# of Words	19,239
OOV Words	155
Perplexity	14.9101

Table 4: Perplexity measures of the language model for a given test data set

5 Summary

In this work we have designed and developed a speech corpus for the use of Sinhala Automatic Speech Recognition. Our corpus is consisted with 78,667 utterances recorded from mobile phone calls and lengthed around 65 hours of speech.

Acknowledgment

The authors are very grateful to Mr. Chamila Liyanage, Ms. Jeewanthi Liyanapathirana, Mr. Randil Pushpananda, Ms. Dilhani Samaranayake, Mr. Namal Udalamaththa and former members of Language Technology Research Laboratory, UCSC for their significant contribution on developing the speech corpus. This work is partially supported by OnMobile Global Limited, India.

References

Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, RNV Sitaram, and SP Kishore. 2005. Development of indian language speech databases for large vocabulary speech recognition systems. In *Proc. SPECOM*.

Etienne Barnard, Marelie Davel, and Charl Van Heerden. 2009. Asr corpus design for resource-scarce languages. ISCA.

Horia Cucu, Audi Buzo, Corneliu Burileanu, et al. 2012. Asr for low-resourced languages: Building a phonetically balanced romanian speech corpus. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2060–2064. IEEE.

Sadaoki Furui. 2005. 50 years of progress in speech and speaker recognition. In *10th International Conference on Speech and Computer-SPECOM, Patras, Greece*, pages 1–9.

Sawit Kasuriya, Virach Sornlertlamvanich, Patcharika Cotsomrong, Supphanat Kanokphara, and Nattanun Thatphithakkul. 2003. Thai speech corpus for thai speech recognition. In *Proceedings of Oriental COCOSDA*, pages 54–61.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *LREC*.

Vlasta Radova and Petr Vopalka. 1999. Methods of sentences selection for read-speech corpus design. In *TSD*, volume 99, pages 165–170. Springer.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Asanka Wasala, Ruvan Weerasinghe, and Kumudu Gamage. 2006. Sinhala grapheme-to-phoneme conversion and rules for schwa epenthesis. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 890–897. Association for Computational Linguistics.

Ruvan Weerasinghe, Dulip Herath, Viraj Welgama, Nishantha Medagoda, Asanka Wasala, and Eranga Jayalatharachchi. 2007. Ucsc sinhala corpus - pan localization project-phase i.