Comparison of Two Case Studies in Sinhala Speech Recognition

Thilini Nadungodage
Language Technology Research Laboratory
University of Colombo School of Computing
No 35, Reid Avenue, Colombo 07
hnd@ucsc.lk

Introduction

Speech is the most natural way of communication among humans. Speech recognition is the process of transforming a speech signal into its corresponding word sequence. When the recognition is carried out by a computer program, it is known as Automatic Speech Recognition (ASR). In this poster we present a comparison of two experiments in Sinhala continuous speech recognition.

Two Case Studies

We have tried out two different experiments in building a Sinhala continuous speech recognition system. Main difference of the two experiments was the no of speakers in the speech corpuses.

- **1st corpus**
  - Small (2949 utterances)
  - Consists of a single female voice
  - Minimum noise
  - Recorded with a microphone
- **2nd corpus**
  - Large (62559 utterances)
  - Consists of 2000 various male & female voices
  - With background noise
  - Recorded with mobile phones

Methodology

Building of an ASR system mainly consists of designing two models, namely the Acoustic Model and the Language Model. We have used HTK toolkit [1] to build those two models and evaluate the built systems.

Acoustic Model

An acoustic model is created using audio recordings of speech and their text scripts and compiling them into a statistical representation of sounds which make up words. This is done through modeling the HMMs. The process of acoustic modeling is shown in Figure 1.

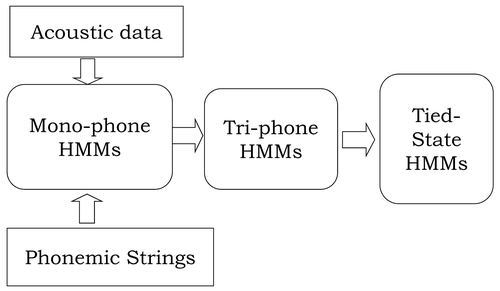

Figure 1: Block diagram of the acoustic modeling process .

Language Model

The way the words are connected to form sentences is modeled by the language model with the use of a pronunciation dictionary. The Language model of the we used is a statistical based bi-gram language model. The process of language modeling is shown in Figure 2.

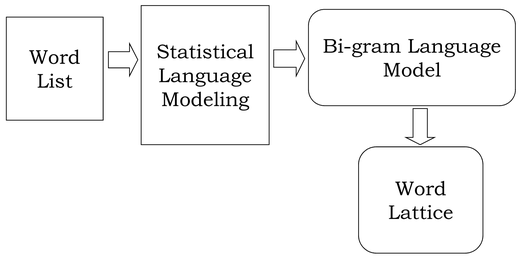

Figure 2: Block diagram of the language modeling process .

Data Set

We collected two data sets for the two experiments separately.

•1st corpus

- Created with sentences which consists most frequently used words in Sinhala language. These sentences were extracted from the UCSC/LTRL Sinhala Corpus.

•2nd corpus

- Created with various types of sentences and keywords. Some of the categories include:
  - Boolean values
  - Date and Time
  - Currency
  - Numbers
  - Phonetically rich sentences

Training

Using above two data sets two acoustic models were trained to recognize Sinhala speech. The basic procedure of building an ASR can be shown as follows:

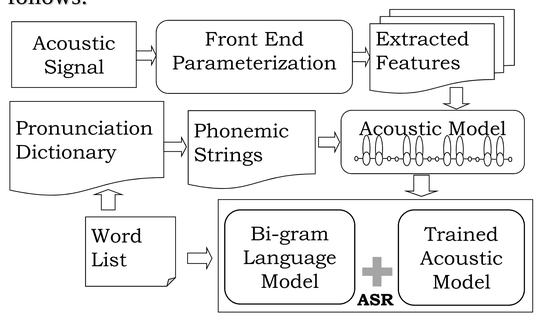

Figure 3: Block diagram of the process of training the ASR.

Evaluation

To test the built systems, some new utterances were sent through those models and obtained the recognized sentences. Those generated sentences were compared with the actual sentences to calculate the performances if the two systems.

- •1st acoustic model – tested with 106 utterances
- •2nd acoustic model – tested with 4865 utterances

Results

Acoustic Model	100% correctly identified sentences (%)	Correctly identified words (%)
1st acoustic model	75.74	96.14
2nd acoustic model	19.55	38.98

Discussion & Conclusion

- •The first acoustic model shows a considerably good recognition rate while the second model shows a poor recognition rate.
- •One reason for this behavior can be because of the large amount of noise that included in the 2nd data set.
- •Another reason can be because the larger corpus was created with various human voices and the training process was very complicated.
- •We are currently trying to improve the performance of the 2nd acoustic model by adjusting the parameter values in the training process.

Reference

[1]. Young, S. Evermann, G. Gales, M. Hain, T. Kershaw, D. Liu, X. Moore, G. Odell, J. Ollason, D. Povey, D. Valtchev, V. and Woodland, P. 2006. *The HTK Book.* Cambridge University Engineering Department, pp. 1-14.