Design and Development of Automatic Speech Recognition System for Tamil Language Using CMU Sphinx 4

1M. Kalith

Faculty ofApplied Science, South Eastern University ofSri Lanka,
*Corresponding Author: imkalith@fas.seu.ac.lk

This paper presents a design and development of Speech Recognition System for Tamil language. This system is based on CMU Sphinx 4 open source speech recognition (ASR) engine developed by Carnegie Mellon University. This system should be adapted to speaker specific automatic, continuous speech. One of the main components of this system is a core Tamil speech recognition system that can be trained with field specific data. The target domain is the accent spoken by illiterate Tamil-speaker from Eastern area of Sri Lanka. The phonetically rich and balanced sentence text corpus were developed and recorded in conditional environment to set up speaker specific speech corpus. Using this speech corpus the system was trained and tested with speaker specific (testing with same word uttered by same person) and speaker independent data (testing with different word uttered by different person). The system currently gives a satisfactory peak performance of 39.5% Word Error Rate (WER) for speaker specific and unsatisfactory rate for speaker independent data, which is comparable with the best word error rates of most of the recognition systems for continuous speech available for any language.

Key words: - Speech recognition, CMU Sphinx 4, Tamil language