Abstract

Research in the area of Large Vocabulary Continuous Speech Recognition (LVCSR) for Indian languages has not seen the level of advancement as in English since there is a dearth of large scale speech and language corpora even today. Tamil is one among the four major Dravidian languages spoken in southern India. One of the characteristics of Tamil is that it is morphologically very rich. This quality poses a great challenge for developing LVCSR systems. In this paper, we have analyzed a Tamil corpora of 10 million words and have exhibited the results of a type-token analysis which implies the morphological richness of Tamil. We have demonstrated a grapheme-to-phoneme (G2P) mapping system for Tamil which gives an accuracy of 99.56%. We have shown the impact of important parameters such as absolute beam width, language weight, number of gaussians and the number of senones on speech recognition accuracy for limited vocabulary (3k). We have presented the results of large open vocabulary speech recognition task for vocabulary sizes of 30k, 60k and 100k on the speaker independent task. The Out Of Vocabulary (OOV) rates are 20.2%, 15.8%, 12.8% respectively. The accuracies are 43.59%, 47.11% and 43.52% respectively.

References
- J. Novak, D. Yang, N. Minematsu, K. Hirose, "Initial and Evaluations of an Open Source WFST-based Phoneticizer", The University of Tokyo, Tokyo Institute of Technology.

Index Terms

Computer Science Artificial Intelligence

Keywords
Speech Recognition Tamil Sphinx Large Vocabulary