Abstract

Spoken Language Identification is a task of recognizing the language from an unknown utterance of speech. The ability of machines to distinguish between different languages becomes an important concern with the emerging trends in global communications which are multilingual nature. This paper describes a text independent language recognition system using a common code book and discrete hidden Markov models (DHMM) to achieve a very good LID recognition performance with less computation time comparing with that of a state of art phone based systems available in literature. This approach includes generation of a common codebook and training of DHMM, one for each language. The experiments are carried out on the database of Indian language consists of six languages namely Telugu, Tamil, Hindi, Marathi, Malayalam and Kannada.

References

- M. A. Zissman "Comparison of Four Approaches to Automatic Language
Identification of Telephone Speech; IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 4, NO. 1, p. p 31-44 JANUARY 1996.

- Language Technology Research Center, IIIT Hyderabad.

**Index Terms**

Computer Science

Pattern Recognition

**Keywords**

Language Identification (lid)  Mel Frequency Cepstral Coefficients (mfcc)  Vector Quantization (vq)

Discrete Hidden Markov Model (dhmm)