1 Automatic Speech Recognition

Information processing machines have become omnipresent. However, the current modes of human machine communication are geared more towards living with the limitations of computer input/output devices rather than the convenience of humans. Speech is the primary mode of communication among human beings. On the other hand, prevalent means of input to computers is through a keyboard or a mouse. It would be nice if computers could listen to human speech and carry out their commands. Automatic Speech Recognition (ASR) is the process of deriving the transcription (word sequence) of an utterance, given the speech waveform. Speech understanding goes one step further, and gleans the meaning of the utterance in order to carry out the speaker's command.

1.1 MFCC: Audio Feature

The most commonly used feature extraction method in automatic speech (ASR) is Mel-Frequency Cepstral Coefficients (MFCC). This feature extraction method was first mentioned by Bridle and Brown in 1974 and further developed by Mermelstein in 1976 and is based on experiments of the human misconception of words.

MFCC is spectral based parameter used in recognition approach. Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database. All extracted MFCC samples are then statistically analyzed for principal components, at least two dimensions minimally required in further recognition performance evaluation.

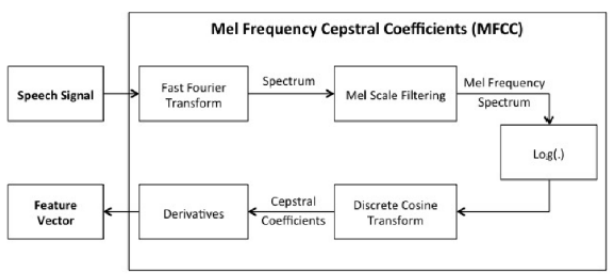

Figure 1

The above figure depicts the flowchart with the various stages of computation for getting the MFCC features.

To extract a feature vector containing all information about the linguistic message, MFCC mimics some parts of the human speech production and speech perception. MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. To represent the dynamic nature of speech the MFCC also includes the change of the feature vector over time as part of the feature vector .