ABSTRACT

Speaker recognition is the task of establishing identity of an individual based on his/her voice. Due to recent advances in the field of speaker recognition, it has become possible to use these systems for authentication and in forensics. It does not require sophisticated or dedicated hardware and has a significant potential as a convenient biometric method. However, in real life situations such speaker identification systems are required to work in noisy environment. The main emphasis in this thesis has been on designing a robust speaker identification system with enhanced performance in practical conditions. The areas of possible improvement in the field of speaker recognition have been investigated.

The task of Speaker Recognition is achieved by two-stage signal processing i.e. training and testing. The training process calculates speaker-specific feature parameters from the speech. These features are used to generate statistical models of different speakers. In the testing phase, speech samples from unknown speakers are compared with the speaker models and classified.

In order to improve the quality of the speech used, the use of voice activity detection (VAD) has been investigated. Results show that the proposed VAD algorithm gives an improvement of about 7% in reduction of error rate. Mel-Frequency Cepstral Coefficients (MFCC) and its derivatives have been used as a feature extraction technique. The key features are extracted from the speech signal using MATLAB and represented by a matrix of cepstral coefficients. The language/data related and speaker dependent factors such as language, length of sample, accent/style of speaking, disguise and emotional state of speaker that affect the accuracy of a MFCC-based speaker recognition system have been investigated. Also, the effects of microphone/sample quality and noise have been analyzed. It has been observed that the most significant factors, among the tested ones, are the noise, microphone quality and disguise. The results also show that the emotional state of the speaker causes a remarkable variation in fundamental frequency of speaker.

Then using the features extracted from speech signals, a statistical model is created to establish the identity for each person enrolling in the system. In this thesis, Back Propagation neural network and Support Vector Machine (SVM) based classifiers have been used for Speaker Modeling. Due to their stacked, layer-wise structure they
have been proved to model highly nonlinear relations between inputs and outputs of a system with high performance. A comparative analysis between the two techniques has been carried out. In this comparison, 2.3% equal error rate (EER) is achieved for Neural Network classifier and the EER dropped to 1.2% for SVM classifier. Noise compensation technique has been proposed to improve the performance in noisy environment. Tests have been performed using TIMIT and NOIZEUS database. Results obtained have been presented in terms of detection error tradeoff (DET) curve and EER. In the scope of this work, hardware has been assembled through which the software, including the algorithm and developed models could operate. An effort has been made to demonstrate the operation of the speaker recognition technology in real-life scenarios for Hindi words.

The identified drawbacks of the current speaker recognition systems are scalability and slow convergence rates of the modeling techniques. The number of feature vectors, their dimensions and the complexity of the speaker models are the factors which affect the performance and processing time of the system. In this thesis, emphasis has been given on reduction of the dimension of feature space by selecting relevant features. An optimization method based on Genetic Algorithm (GA) has been used. The results obtained using GA has been compared with that of Principal Component Analysis (PCA) for speaker identification task. Simulation results show an improvement of approximately 2.3 % in speaker identification rate for SNR of 30 dB in case of GA. Also, the EER obtained with 61.1% of reduction of features is almost same as that obtained with complete feature set.

Hence, it can be concluded that after GA transformation, the computational overhead of the subsequent processing stages is reduced and at the same time, effect of noise is minimized which results in improved accuracy.

Keywords: Speaker Recognition, Feature Extraction, MFCC, Voice Activity Detector, Feature Matching, Support Vector Machine, Neural Network and Genetic Algorithm.