ABSTRACT

Automatic Speaker Verification (ASV) has been the most investigated topic in speech processing since early 1950s. Broadly speaking, it can be formulated as the problem of accepting or rejecting a claim of identity based on a voice sample. The complexity of ASV system and its verification time depends on range of feature vectors, their spatiality, and the complexity of speaker models and therefore the range of speakers. Often ASV systems make errors in conditions where a human speaker could produce only sparse data, i.e., speech that is finally heard may be distorted by a variety of external influences, not related to what was spoken, which affect its characteristics. While humans are not affected by them, ASV systems can be highly sensitive to these distortions.

In other words, ASV systems are not immune to distortions in the speech signal in the manner that humans are. The speech signal carries both useful and unwanted information and therefore extracting or enhancing the useful features from a mixture of conflicting information is highly needed for better verification. Feature selection is one of the most important topics in speech signal processing to improve the performance of the systems in noisy conditions. Unwanted features are the main limiting factors in communication systems. Therefore the modeling and removal of the unwanted features which contribute to improve the performance of ASV system have been at the core of the theory and practice of communications and speech signal processing.

In this regard, the modern methods, which are termed Meta heuristic optimization methods or manifold learning, aim at discovering the underlying structure in data sets as a predecessor to other types of processing. These feature analysis approaches have recently become very popular in machine learning and data mining applications. Kernel Meta heuristic optimization methods have been an area of active research, playing important roles in an ever increasing number of
applications. All the speech enhancement techniques aimed at suppressing the unnecessary features are (naturally) based in one way or the other on the estimation of these features. However, the focus in using these techniques was shifted to extract features with high variability among people. Most commonly used features extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs) have been particularly popular for ASV systems in recent years. This transforms give a highly compact representation of the spectral envelope of a sound. To increase the verification rate, one needs to use the optimization technique for feature selection after the feature extraction technique. After applying the optimization technique, the improvement in optimization of features was about 87% and Hybrid Ant Bee Colony Optimization (HABC) algorithm was tested in different noise conditions using the Berlin Speech Data base and Telephone Conversation. For the second contribution, Multi Objective Hybrid Ant Bee Colony Optimization (MOHABC) algorithm developed which showed much better results than the Hybrid Ant Bee Colony Optimization algorithm in optimizing the feature subset present of a speech signal.

As a third contribution, when sparse data of 2~5 sec is provided for speaker verification, we applied the MOHABC algorithm and GMM mixture indexing in combination to improve the verification accuracy. GMM mixture indexing is a technique where feature frame of the speaker training data is labeled with the highest probability in a table known as Gaussian Mixture Index (GMI). The speaker feature frame is classified into one Gaussian mixture index in the GMI scheme and this indexing process is applied to both the trained data set and test data. The indexed information from the train data is used for the speaker adaption procedure, which used different adaption data depending on the ranking of the indexed information. To fill acoustic holes for sparse training data, the test speaker borrows data from acoustically close speakers. When acoustically
similar data are selected from cohort speaker data, caution should be exercised to ensure a minimum number of cohorts for filling acoustic holes. The method adopted to select similar characteristic speakers is Probability-Score-Method (PS-M). It uses the probability of the feature frame against the speaker model. The average phoneme occurrence is counted as a unit in the indexed mixture of each speech feature frame using the GMI. A balanced acoustic token histogram is formed from this information, and is called the FIND table.

As a result of the difficulties found in the application of traditional models for speaker verification, the proposed hybrid approach performed better when the number of parameters was comparable, and about the same in some cases in which the traditional system used many more parameters. Also, the hybrid system was quite efficient in terms of CPU and memory run-time requirements. In this thesis work we used Multi Objective Optimization and Hybrid Ant Bee Colony Optimization combination for the speaker verification system along with the case of sparse train and test data.