Abstract

Over the last decade, an *Speaker Identification* (SI) systems, have been employed efficiently as an add-on module with various speech related applications like *Speech Recognition*, *Speaker Verification* (SV), *Personalized Speech Coding*, and many other *Personalized User Interfaces*. At the time of testing, the identification time for an unseen speech sample depends on the number of feature vectors, their dimensionality, the complexity of the speaker models and the number of speakers. Methods like *Pre-quantization* (PQ) and *Speaker Pruning* help the system to achieve considerable speed-up gain. However, the performance of the system degrades somewhat as some of the important parameters have been pruned out during testing in order to make it faster. In this thesis, we focus on decreasing the computational load in the identification phase while an attempt is made simultaneously to keep the recognition accuracy reasonably high through various fusion strategies that take evidence from conventional as well as complementary feature sets and different feature sets with their reduced dimensionality.

First, we have proposed a complementary feature set, which can describe high frequency part of the spectrum more than its counterpart, baseline *Mel-frequency Cepstral Coefficients* (MFCC) method. The SI performances of the proposed complementary feature set has been compared with baseline. Next, a fusion strategy is developed to fuse the score from the models separately developed for baseline and feature sets. PQ method has been introduced in the fusion strategy to compare the performance on equivalent computational load. The results of our analysis indicate that, using the proposed complementary features and PQ based fusion scheme, one can achieve an appreciable enhancement in SI accuracy while utilizing the same amount of resource that a baseline system does. Subsequently, a study has been done to observe the effect of using various filter bank shapes on identification accuracy. *Gaussian filter* (GF) has been introduced to average the speech spectrum for determining the spectral envelope. Six

different shapes of filters that include Rectangular, Triangular, and Gaussian with its four varieties have been explored to extract meaningful cepstral parameters. The study reveals that cepstral features obtained from GF outperform other variants for a closed set SI task. Complementary features are also obtained using the Gaussian shaped filter and in a similar way complementary speaker models have been fused. In sequel, a straightforward and non-exhaustive search based *Feature Selection* (FS) method based on *Singular Value Decomposition* (SVD) followed by *QR Factorization with Column Pivoting* (QRcp) has been proposed in this thesis for achieving higher identification rate with reduced feature set compared to full feature set. The performance of the system has been compared with the system, which uses the features selected by well known *Feature Selection* (FS) method *F-Ratio* (FR) based feature selection. It has been observed that our proposed FS method is superior both in terms of number of features required to be used and error rate. Finally, some candidate strategies have been proposed for fusing the information from multiple sources and applied at various levels of an SI system to enhance the system's performance over single stream baseline system. Using the best fusion technique, we obtain significant relative improvements in terms of SI error rate as compared to conventional (baseline) system for two public databases, namely **YOHO** and **POLYCOST**, respectively comprising more than **130** speakers.

Keywords: Divergence, Fusion, Gaussian Mixture Model (GMM), GIMFCC, GMFCC, IMFCC, MFCC, PQ, QRcp, SI, Speaker Recognition, Speed-up, Subband, SVD.