

Abstract

The automatic speaker recognition technologies have developed into more and more important modern technologies required by many speech-aided applications. The main challenge for automatic speaker recognition is to deal with the variability of the environments and channels from where the speech was obtained. In previous work, good results have been achieved for clean high-quality speech with matched training and test acoustic conditions, such as high accuracy of speaker identification and verification using clean wideband speech and Gaussian Mixture Models (GMM). However, under mismatched conditions and noisy environments, the performance of GMM-based systems degrades significantly, far away from the satisfactory level. While in real-world applications, the matched acoustic conditions cannot always hold. Therefore, robustness becomes a crucial research issue in speaker recognition field.

In this thesis, our main focus is to improve robustness of speaker recognition. We investigate approaches to improve robustness from two directions. First of all, we investigate approaches to use high-level speaker information to improve robustness. We propose new techniques to model speaker pronunciation idiosyncrasy from two dimensions: the cross-stream dimension and the time dimension. Such high-level information is expected to be robust under different channels. Secondly, we investigate approaches to improve robustness for traditional speaker recognition system which is based on low-level spectral information. We introduce a new reverberation compensation approach which along with feature warping in the feature processing procedure improves the system performance significantly. We propose four multiple channel combination approaches, which utilize information from multiple far-field microphones, to improve robustness under mismatched training-testing conditions. Thirdly, we study speaker segmentation and clustering aiming at improving the robustness of speaker recognition as well as automatic speech recognition performance in the multiple-speaker scenarios such as telephony conversations and meetings. Finally, we integrate speaker identification modality with face recognition modality to build a robust person identification system.