Abstract

Mel-frequency cepstral coefficients (MFCCs) have been the most popular and widely used cepstral feature for decades to perform the speaker verification (SV) as well as other speech processing tasks such as automatic speech recognition, speaker diarization, spoofing countermeasures etc. The recently introduced x-vector based SV system, which has shown state-of-the-art performance in previous NIST speaker recognition evaluation, also uses MFCCs as acoustic features. The MFCCs are very popular for the following reasons. First, the computation process utilizes mel filterbank analysis, which is partially inspired by the processing of the audio signal by the human auditory system. Second, the computation process involves fast Fourier transform (FFT) and matrix multiplication, which makes it more computationally efficient compared to other methods such as linear prediction cepstral coefficients (LPCCs) or linear frequency cepstral coefficients (LFCCs). Third, MFCCs are also suitable with different feature level compensation methods such as relative spectral (RASTA) processing, cepstral mean and variance normalization (CMVN), and feature warping.

Though the MFCCs are relatively more robust compared to other cepstral features such as linear frequency cepstral coefficients (LFCCs) or LPCCs, the SV performance with MFCCs are severely degraded in real-world conditions due to the mismatch of acoustic conditions in enrollment (or speaker registration) and verification (or speaker authentication) phase. To overcome some of the shortcomings of MFCCs, various robust acoustic features, for example, frequency domain linear prediction (FDLP), cochlear frequency cepstral coefficients (CFCCs), power-normalized cepstral coefficients (PNCCs), mean Hilbert envelope coefficients (MHECs), Gammatone frequency cepstral coefficients (GFCCs), constant-Q cepstral coefficients (CQCCs), time-varying linear prediction (TVLP), and locally-normalized cepstral coefficients (LNCCs) were proposed. All the above features even though achieve better performance in noisy condition; they require a large number of user-defined parameters. These parameters further need to be manually tuned for different environmental conditions. The overall process seems to be difficult for end-user in real-world applications. The MFCCs, on the other hand, have mainly two free parameters: the number of filters and the number of coefficients to be retained after DCT. Most of the proposed features are also computationally more expensive than MFCCs.

The data-driven methods show the improvement in robustness where large databases are used in training strong discriminative models, as compared to the new features. In this thesis, the focus is to develop a data-driven feature extraction method which captures more speaker specific information in clean as well as noisy conditions. Unlike the above-discussed feature extraction methods which require “hand-crafted” parameters, we derive the parameters from the speech data itself. In this thesis, we used data-driven frequency warping scale for speaker recognition application. The cepstral feature derived from this scale is called as speech-signal-based frequency cepstral coefficient (SFCC). The said cepstral feature-based SV system gives better performance than baseline MFCC based SV system.

The spectral entropy information is then used to investigate an entropy modulated speech-signal-based scale for efficient computation of cepstral features. The proposed feature extraction method is named as MSFCC (Modified speech-signal-based frequency cepstral coefficient). Our study shows that MSFCC outperforms both MFCC and SFCC features. This feature also gives complementary information due to which score level fusion is also done to improve the overall performance of the SV system.

Next, we investigate another approach where we use pitch estimation technique to select specific speech frames which subsequently, are used for creation of speech-signal-based scale. This approach helps to create a better speech-signal-based scale with lesser complexity

in process. Using this, feature extracted from proposed speech-signal-based scale is termed as pitch based SFCC or PSFCC. It is found that the proposed feature-based SV system gives better results than MFCC and SFCC based SV system in both clean and noisy conditions. The score level fusion is also done between MFCC and proposed feature to get better performance of SV system.

Finally, the shape of the filters of the filterbank is investigated by using Gaussian function and PCA based technique. This leads to an optimized filterbank based feature extraction method. We use both the approaches to all the filterbanks used in our thesis, i.e., filterbanks used for MFCC, SFCC, MSFCC and PSFCC features. We find that the optimized filters help to improve the performance of the corresponding cepstral feature based SV systems. Overall, our proposed methods, after optimization process, outperform the MFCC and SFCC based SV systems.

The proposed methods are evaluated on multiple speaker recognition evaluation (NIST SRE) corpora. To investigate the robustness of SV system, we do the experiments on simulated noisy environment created by NOISEX-92 corpus. We also do a detailed analysis of the noise susceptibility of the proposed methods by investigating different noisy conditions. We find that our proposed methods based SV systems are more robust as compared to MFCC and SFCC based SV systems. In order to test the robustness of SV system in real-world conditions, VoxCeleb1 - the real-world corpus is studied using i-vector classifier. In this case, too, our proposed methods perform better. We can conclude that the study conducted in this thesis and features proposed from that, improve real-world application of SV system.

Keywords: Mel scale, Speech-signal-based scale, NIST, Pitch, Spectral entropy, GMM-UBM, Speaker Recognition, Speaker Verification, NOISEX-92, i-vector.