ABSTRACT

The primary intention of speech parameterization is to extract useful and efficient cepstral information about what is being spoken from the speech signal. Speech feature extraction is participated as a critical role in Automatic Speech Recognition (ASR) to determine the recognition performance. The Mel Frequency Cepstral Coefficients (MFCC) is widely used feature vectors for speaker and speech recognition system. This approach can work adequately in quiet environments; however, its performance will drop significantly under the adverse situations.

Generally, compared to an ASR system, human is good at tolerating under adverse situations. MFCC was initially suggested for identifying the monosyllabic words in continuously spoken utterances. The calculation of MFCC is intending to implement the ear's working principle artificially. However, the main drawback of this technique is a low-level of noise immunity in recognition process. It leads to a sharp deterioration in speech applications since all MFCCs are varied by the noisy speech when at least one frequency band is distorted. Thus, we propose the perceptual simultaneous masking effect of psychoacoustic modeling is integrated into conventional MFCC to reduce the vulnerability of noise disturbance. The psychoacoustic model is based on the human perception of sound in environment. Even in the audible frequency range, human ear cannot perceive all frequencies in the same way. The weaker sound is made inaudible or uncleared by simultaneously occurring of one louder sound in the certain critical band. The minimum masking threshold, which indicates the sensitive limit or noticeable distortion of the signal, is figured out and integrated into extracting the features to diminish the noise effect.

The Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and Deep Neural Network-Hidden Markov Model (DNN-HMM) training are employed to design a robust speech recognition system. The experiments are carried out on AURORA-2 English connected digits noisy speech corpus to prove the recognition rate can be improved significantly over the standard MFCC. With the use of proposed perceptual masking effect-based cepstral features (PGFCC), the accuracy reached up to 95.88% in signal to noise ratios (SNR) of 10dB, 88.69% in 5dB,

65.04% in 0dB and 26.93% in −5dB with GMM-HMM training. In terms of cross-entropy based DNN-HMM, the accuracy is gained up to 98.14% in SNR of 10dB, 94.43% in 5dB, 81.67% in 0dB and 51.71% in −5dB as well. Using the sequence discriminative DNN-HMM training, we reached the accuracy up to 98.39% in 10dB, 95.03% in 5dB, 83.11% in 0dB and 53.49% in −5dB, accordingly.

The proposed PGFCC outperforms the conventional MFCC using sequence discriminative DNN-HMM training. The relative improvement is 0.39 % in 10dB, 1.2% in 5dB, 6.3% in 0dB and 32.30% in −5dB respectively. The experimental results show that the proposed method provides the significant improvement in low SNRs especially at 0dB and −5dB. The complexity of speech signal can be reduced and minimized the noise effect without affecting the perceived quality of sound by combining the concept of simultaneous masking effect into extracting the features. Moreover, substituting the Bell shape Gaussian filterbank can also tolerate the loss of spectral information in nearby subbands. Besides, we compared and analysed the performance of proposed PGFCC with other modified MFCC denoising scheme.

The relative progress of PGFCC is 0.17% in 10dB, 0.40% in 5dB, 5.75% in 0dB and 50.3% in −5dB over RASTA-MFCC. When our proposed feature is comparing with DWT-MFCC, the relative growth is 0.39% in 10dB, 0.83% in 5dB, 5.71% in 0dB and 34.2% in −5dB, as well. The superior performance of proposed method over PLPCC is 0.06% in 10dB, 0.99% in 5dB, 10.43% in 0dB and 90.28% in −5dB, accordingly. Then, the relative improvement is 0.06% in 10dB, 0.34% in 5dB, 3.46% in 0dB and 24.71% in −5dB over DDDWT-MFCC. Moreover, the recognition performance of open noise situations such as airport, street, restaurant and train station noises that are not included in the training process is investigated. While the mean accuracy achieved (76~86)% under the noise situations that include in the training phase, the recognition accuracy reached up to (74~77)% under the unseen noise environments.

Keywords: Automatic Speech Recognition, Hidden Markov Model, Deep Neural Network, Simultaneous Masking, Mel Frequency Cepstral Coefficients (MFCC), Psychoacoustic Model