Study on monaural speech enhancement by restoring instantaneous amplitude and instantaneous phase

1910442 Vo Duc Duy

Speech is one of the most essential and important means for humans to communicate with each other. With the advancement of science and technology, speech now can be transmitted from a far distance to help connect people around the world through phones or web meetings,... Furthermore, it can also be used as an interface for humans to communicate with machines via automatic speech recognition (ASR) systems. However, in real environments, speech is contaminated by noise, reducing its quality and intelligibility such that it heavily affects the performance of these systems. In order to address this issue, efficient speech enhancement algorithms are needed for both human hearing and ASR systems.

Many techniques have been proposed to separate clean speech from the noisy mixture. Current state-of-the-art methods usually use short-time Fourier transform (STFT) as the means for feature extraction. The word recognition rate of ASR systems utilizing these techniques as the front-end still falls short of expectations, despite the fact that these methods can enhance the quality and intelligibility of noisy speech. Recent studies have showed that temporal envelope and temporal fine structure are crucial cues for speech perception and they also play a significant role in improving the speech intelligibility in noisy conditions. Therefore, speech enhancement by modifying instantaneous amplitude (IA) and instantaneous phase (IPh) extracted from an auditory filterbank is expected to have better improvement in quality, intelligibility as well as word recognition rate of ASR systems than STFT. On this basis, a speech enhancement method based on IA and IPh from the auditory filterbank was proposed. However, this method processed each channel independently, which could neglect important cross-channel information for ASR systems.

The purpose of this research is to investigate a model that can utilize cross-channel information of IA and IPh to explore the ability of this information in elevating the word recognition rate of ASR systems. This model revolves around vector-quantized variational autoencoder to estimate IA, and a complex convolution network to estimate IPh.

The efficacy of the proposed model will be estimated using three metrics: perceptual evaluation of speech quality, short-time objective intelligibility, and word error rate. The outcomes demonstrate that the proposed method can enhance quality, intelligibility of noisy speech and is competitive with some state-of-the-art methods. However, this method still cannot resolve the issue of high word error rate in ASR systems.