RESEARCH ON SPEECH SEPARATION UNDER COMPLICATED NOISY ENVIRONMENT

ABSTRACT

Speech separation for multiple speakers has been a tough problem in speech signal processing for a long time. In recent years, thanks to the rapid development of deep learning, there has been great improvements in the speech separation when there exists no noise. However, it is still very difficult to achieve the expected effect for the speech separation in actual noise environment. In response to the problem, we researched the separation method for noisy speech signals in the real complicated environments. Specifically, we adopted a two-stage separation framework, which first separates the noise and the speech (also known as speech enhancement or speech noise reduction), and separates the voices for multiple speakers after that. We realized the efficient separation of noisy mixed speech with speech enhancement as the front-end and the speech separation as the back-end. In speech enhancement, we used the long and short-term memory networks to obtain the time-series context information of speech signals from the estimation of the time-frequency mask, comparing with the classical estimation method based on statistical clustering. In terms of speech separation, we used an end-to-end neural network model to estimate the masks, thereby separating the voices of different speakers. Finally, we evaluated the performance of speech enhancement and speech separation based on two speech signal quality evaluation standards, PESQ and signal distortion ratio. The two-stage separation framework is verified under four typical noise environments and different signal-to-noise ratio conditions.

Key words: Speech Enhancement, Speech Separation, Time-Frequency Analysis, Beamforming, Deep Neural Network