Thesis: End-to-end Multi-channel Target Speech Separation and Recognition

Abstract: Speech separation is an essential task for the far-field multi-talker speech recognition task. Neural network-based separation methods achieve better objective performances. However, they often cause the non-linear speech distortion that harms the performance of ASR systems. Beamforming techniques (e.g., minimum variance distortionless response (MVDR)) aim to reduce the noise and preserve the distortionless speech. There are two challenges in estimating the MVDR beamforming weights by using the neural network. First, the gradients of computing the inverse or the division of the complex-valued matrices are not stable when joint trained with neural networks. Second, the conventional beamforming weights are time-invariant that is not optimal for noise reduction. To overcome the challenges, we propose an RNN-based differentiable MVDR beamformer that directly estimates the frame-wise beamforming weights. To investigate the impact of the joint training with the ASR system, we build an end-to-end framework that comprises the RNN-MVDR beamformer and a pre-trained RNN Transducer (RNNT) based ASR. We optimize the MVDR beamformer by utilizing the Connectionist Temporal Classification (CTC) loss for ASR and the scale-invariant source-to-noise ratio (Si-SNR) loss for speech separation.

Committee:

- Professor Michael Mandel, Mentor, Brooklyn College
- Professor Rivka Levitan, Brooklyn College
- Professor Lei Xie, Hunter College

Outside Member:

- Keelan Evanini, Educational Testing Service