Towards Solving Cocktail-Party: The First Method to Build a Realistic Dataset with Ground Truths for Speech Separation

Rawad Melhem[1], *Assef Jafar*[1] *and Oumayma Al Dakkak*[1]

[1](*Higher Institute for Applied Sciences and Technology, Damascus, Syria*)

Abstract

Speech separation is very important in real-world applications such as human-machine interaction, hearing aids devices, and automatic meeting transcription. In recent years, a significant improvement occurred towards the solution based on deep learning. In fact, much attention has been drawn to supervised learning methods using synthetic mixtures datasets despite their being not representative of real-world mixtures. The difficulty in building a realistic dataset led researchers to use unsupervised learning methods, because of their ability to handle realistic mixtures directly. The results of unsupervised learning methods are still unconvincing. In this paper, a method is introduced to create a realistic dataset with ground truth sources for speech separation. The main challenge in designing a realistic dataset is the unavailability of ground truths for speakers' signals. To address this, we propose a method for simultaneously recording two speakers and obtaining the ground truth for each. We present a methodology for benchmarking our realistic dataset using a deep learning model based on Bidirectional Gated Recurrent Units (BGRU) and clustering algorithm. The experiments show that our proposed dataset improved SI-SDR (Scale Invariant Signal to Distortion Ratio) by 1.65 dB and PESQ (Perceptual Evaluation of Speech Quality) by approximately 0.5. We also evaluated the effectiveness of our method at different distances between the microphone and the speakers, and found that it improved the stability of the learned model.

Keywords: Single Channel, Speech Separation, Deep Learning, Realistic Datasets, Ground Truths.