System Introduction

We introduce a two-stage speech enhancement system to tackle the AVSE challenge, including an audio-visual separation network and an audio-visual selection network. In the first stage, the separation network receives the mixed audio input and the visual cue of the target speaker, and then generates two-channel output, literally the predicted target speech and interferer audio. In the second stage, the selection network evaluates on both channels the compatibility of the audio with the corresponding visual cue, and finally picks one of them as the predicted target speech.

The reason we design the separation network with two-channel output is because the ground truth interferer audio is given in the challenge, so we want to take advantage of both target and interferer information. The selection network in the second stage is inspired by Vocalist [1] an audio-visual synchronization model. Instead of STFT, a multi-scale time domain encoder is applied to extract the audio feature in different granularity. The encoded input audio together with the visual embedding passes three cross-attention block as implemented in Vocalist and becomes weighted time-domain signals, which finally passes two feed forward layers and predicts a score representing audio-visual compatibility. The key problem of this structure is that, the final speech selection result is upper bounded by the separation performance. Therefore, after achieving a validation accuracy of 99.5% in the selection block, we started to focus on the separation performance. At first, Sepformer [2] was used in the first stage to separate the audio, and the hearing test result was already quite clear compared to the ground truth. To further improve the separation performance, the visual cue and multi-modal fusion block were incorporated into Sepformer. The final structure of the separation network is similar to audio-visual Sepformer [3].

Training Details

We followed the rule of track 1 to only use AVSE2 data for training and validation. The visual feature extractor is pretrained in a visual-only word recognition task on LRW, and used in both separation and selection. In the first stage, the model is trained by SI-SDR with PIT. The parameter size of the separation network is 29.7M. Similar to the training scheme of audio-visual sepformer [3], it is trained with Adam in a learning rate of 1.5e-4, and applies a halving strategy if the validation result has no improvement in 3 continuous epochs. In practice, we trained the separation network on 16 Sugon DCU-Z100-16G with 64 two-second utterances every batch under random sampling mode. The separation model achieved best performance within 100 epochs. In the second stage, the selection model is trained by BCE. The parameter size of the selection network is 13.7M. It is trained with Adam in a learning rate of 5e-5. We randomly sampled 64 0.6-second utterances in every batch, and trained the selection network on 16 Sugon DCU-Z100-16G. This model achieved best performance within 500 epochs.

Reference

[1] Kadandale, Venkatesh S., Juan F. Montesinos, and Gloria Haro Ortega. "VocaLiST: an audio-visual synchronisation model for lips and voices." Proc. Interspeech 2022; 2022 Sep 18-22; Incheon, South Korea.[Baixas]: International Speech Communication Association; 2022. p. 3128-32. (2022).

[2] Subakan, Cem, et al. "Attention is all you need in speech separation." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[3] Lin, Jiuxin, et al. "Av-Sepformer: Cross-Attention Sepformer for Audio-Visual Target Speaker Extraction." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.