ICT-AVSU's Submission for The 3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge

We participated in the AVSE Challenge Track 1 and Track 2. This report presents a system description of our model, AVCMGAN.

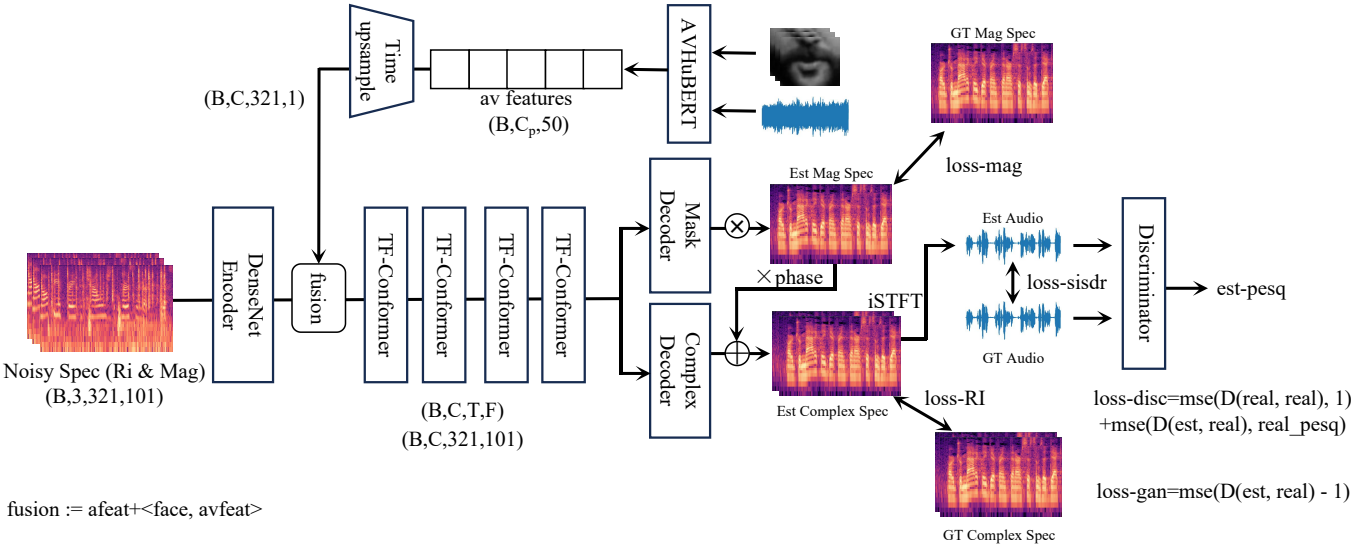

Fig. 1. AVCMGAN model

I. Network

The model architecture, as shown in Fig. 1, consists of an audio branch and a feature extractor. The audio branch is built upon CMGAN [2], which includes a generator composed of a DenseNet encoder, TF-Conformer, and two DenseNet decoders. The discriminator in this branch utilizes a convolutional structure.

For track 1, the feature branch employs a typical lip-reading network consisting of Conv3D, ResNet18, and TCN. The lip-reading network takes speaker's faciaFl videos paired with noisy audio as input and outputs visual features.

As for the track 2, we utilizes a pre-trained Robust AVHubert model [1] to extract audio-visual features. AVHubert takes paired noisy audio and video as input, and we use the features from the 24th layer of the large model for subsequent training.

The visual features or audio-visual features are passed through a deconvolutional network to perform temporal upsampling and align the temporal dimension with the audio features. Afterward, they are added with the audio features and input to the subsequent network.

II. Training

We trained AVCMGAN end-to-end, with the AVHubert model frozen. We extracted the required features using AVHubert in advance and loaded these features directly during the training phase, thereby saving training time. The training parameter settings are as follows:

- Number of training parameters: AVCMGAN-Track 1 is 15.5M, AVCMGAN-Track 2 is 15.1M.
- Init LR: 1e-3.
- Optimizer: AdamW.
- BatchSize: 64.
- Epochs: 26.
- Scheduler: StepLR with 5 steps.

References

[1] Ruizhe Cao, Sherif Abdulatif, and Bin Yang. CMGAN: conformer-based metric GAN for speech enhancement. In *INTERSPEECH*, pages 936–940. ISCA, 2022.

[2] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. In *INTERSPEECH*, pages 2118–2122. ISCA, 2022.