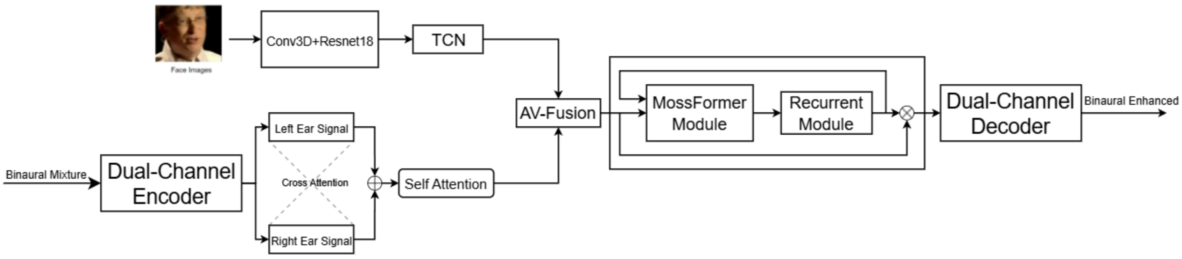
Team TNP’s Submission For AVSE4: Binaural Audio-Visual Speech Enhancement with Mossformer2

Introduction

We participated in Track 1 of AVSE4, and this report provides a systematic description of the model we proposed.

Model Architecture

We propose BAV-mossformer2, whose architecture is shown below. For the visual part, we use Conv3D and Resnet18 to extract features, which are pre-trained, and then refine the visual features through TCN. For the speech branch, we encode the signals from both ear canals through a dual-channel encoder. To simulate more realistic human hearing, we use the signals from the left and right ears as references and apply cross-attention to obtain mixed features, which are then processed through self-attention. Finally, the visual and speech signals are fused and passed through the audio enhancement backbone MossFormer2[1].



Training parameters

We trained BAV-MossFormer2 end-to-end, with the visual pre-trained model being fine-tuned together. The batch size was set to 2, the learning rate was set to 1e-3, the Adam optimiser was used, and a total of 50 epochs were trained.

References

[1]Zhao, Shengkui, et al. "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024