An SPMamba-Based Audio-Visual Speech Enhancement Model for the 4th COG-MHEAR AVSE Challenge

Chih-Ning Chen[1], Jen-Cheng Hou[2], Jun-Cheng Chen[2], Yu Tsao[2], Shao-Yi Chien[1]
[1] National Taiwan University
[2]Academia Sinica
Email:ning@media.ntu.ee.edu.tw

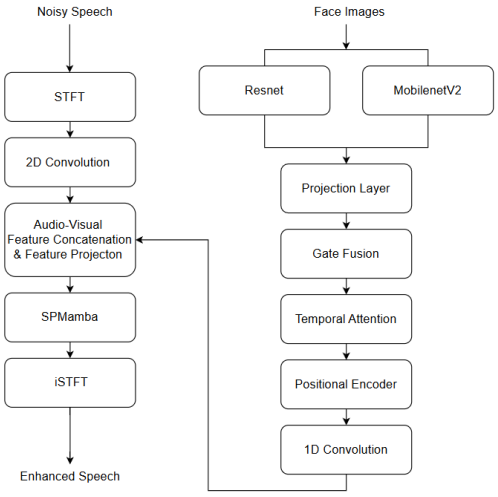

Figure 1: Our proposed AVSE model.

Method

We propose an audio-visual speech enhancement (AVSE) system for the 4th COG-MHEAR AVSE Challenge. As shown in Fig. 1, the model takes SPMamba [1] as the audio backbone model. For the visual part, a dual-branch structure is used to extract visual features from face images. The visual features extracted from the ResNet [3] and MobileNetV2 [6] are aggregated and projected to the same dimension through a linear layer. Gated fusion are then applied to the fused visual features, which then pass through Temporal Attention and Position Encoding to obtain the final visual features.
The final visual features are concatenated with speech features, which are produced by short-time Fourier Transform (STFT) and 2D convolution operations. The combined features are subsequently fed into the SPMamba model. The training objective is based on the Permutation Invariant Training Loss [5].

Experiments

The challenge dataset [4], based on the LRS3 [2] dataset, consists of 34,525 and 3,365 utterances for the training and development set, respectively. The noisy speech contains noise mixutures from different scenes and talkers. We use the provided visual videos, i.e., the cropped face region, as our visual inputs.
The ResNet used for visual feature extraction adopts the same configuration as used in the baseline model, which is a four-layer ResNet. For SPMamba, We follow the model configuration set by the authors' implementation [8]. The full audio-visual model has about 21M trainable parameters. The batch size is 1. Adam is used as our optimizer. The initial learning rate is set to 0.0001, which is halved if the SI-SNR doesn't improve for 10 epochs on the development set. Training was performed on eight NVIDIA V100 GPUs, which has 32 GB VRAM each. Each epoch takes approximately 75 minutes, and total of 80 epochs are trained. The total amount of GPU hours consumed in training is around 800.

Results

Model	MBSTOI
Noisy	0.416
Baseline	0.515
Ours	0.659

Table 1: The mean MBSTOI scores of the enhanced speech on the development set.

We evaluate the enhanced speech with MBSTOI [7], a metric assessing the intelligibility of binaural speech. As shown in Table 1, our approach outperforms the baseline model provided by the organizer by 0.144 in MBSTOI, a 27.9% improvement, suggesting the superior effectiveness of our approach.

References

[1] K. Li, G. Chen, R. Yang, and X. Hu, "Spmamba: State-space model is all you need in speech separation," *arXiv preprint arXiv:2404.02063*, 2024.

[2] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] lorena aldana, "avse_challenge," https://github.com/cogmhear/avse_challenge, 2025, online; accessed 2025-07-05.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[7] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[8] J. Lee, "Spmamba: State-space model is all you need in speech separation," https://github.com/JusperLee/SPMamba/tree/main, 2024, accessed: 2025-07-05.