TEAM CITISIN: Leveraging Mamba for Audio-Visual Speech Enhancement

Rong Chao^{1,2}, *Wenze Ren*^{1,2}, *You-Jin Li*^{1,2}, *Kuo-Hsuan Hung*^{1,2}, *Szu-Wei Fu*^{3}, *Wen-Huang Cheng*^{2}, *Yu Tsao*^{1}

^{1}CITI, Academia Sinica, Taiwan
^{2}CSIE, National Taiwan University, Taiwan
^{3}NVIDIA,

`d13922037@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw`

1. System Description

We propose a hybrid audio-visual speech enhancement system for the AVSEC-4 Challenge, extending the recently introduced SEMamba framework[1]. Our key innovation lies in adapting Mamba-based speech enhancement to accept full-face visual inputs. Inspired by AVSEC-3[2, 3], we explore the use of richer visual context beyond the lip region to improve robustness in visually challenging scenarios.

- **Audio Stream:** A waveform-level encoder initialized from scratch and trained end-to-end. It follows the SEMamba[1] design, integrating convolutional frontends with Mamba-based sequence modeling to capture long-range temporal dependencies efficiently.
- **Visual Stream:** A pretrained 3D ResNet-18 model extracts full-face visual features from 25 FPS video frames. The ResNet backbone is kept frozen to mitigate overfitting and accelerate convergence. A lightweight temporal convolutional network (TCN) aligns the visual features temporally with the audio stream.

The audio and visual representations are fused and passed through a shared TF-Mamba blocks and decoders to estimate the enhanced clean speech waveform.

2. Training Configuration

2.1. Training Settings

- **Epochs:** 450
- **Batch Size:** 2
- **Optimizer:** AdamW
- **Learning Rate:** 5e-4
- **Loss Functions:** Time domain L1-loss, magnitude loss, complex loss, phase loss, and consistancy loss.
- **Data:**
  - **Training dataset:** 34,524 scenes (113hours 17mins)
  - **Development dataset:** 3,306 scenes (8hrs 38mins)

2.2. Model Statistics

- **Total Parameters:** 18.0M
- **Trainable Parameters:** 6.9M
- **Frozen Parameters:** 11.2M (primarily 3D ResNet-18)

2.3. Hardware and Runtime Environment

- **Training Hardware:** 2 × NVIDIA RTX 3090
- **Training Time:** ≈ 3 hours per epoch
- **GPU Memory Usage:** ≈ 23 GB per GPU

3. Results

3.1. Development Set (Binaural)

Table 1 shows the MBSTOI and PESQ scores on the development set. Our AVSEMamba model significantly outperforms the AVSEC baseline and the noisy input across both metrics.

Method	MBSTOI	PESQ
Noisy Input	0.4161	1.30
AVSEC Baseline	0.5150	–
Ours (AVSEMamba)	**0.8037**	**2.97**

Table 1: *MBSTOI and PESQ scores on the development set (binaural).*

3.2. Blind Test Set (Leaderboard, Monaural)

We report PESQ, STOI, and UTMOS on the blind test set in Table 2. Our system achieves substantial improvements in intelligibility and perceptual quality compared to the noisy input.

Method	PESQ	STOI	UTMOS
Noisy Input	1.31	0.55	1.37
Ours (AVSEMamba)	**2.31**	**0.77**	**2.24**

Table 2: *PESQ, STOI, and UTMOS scores on the blind test set (monaural).*

4. Conclusion

We present a full-face audio-visual speech enhancement system based on the AVSEMamba framework, demonstrating substantial improvements over the AVSEC-4 baseline. The combination of pretrained spatiotemporal visual features with lightweight temporal modeling provides a favorable balance between performance and training efficiency. Future work will focus on adapting this system to real-time constraints and exploring its effectiveness in dynamic conversational scenarios.

5. References

[1] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," *in Proc. IEEE SLT*, 2024.

[2] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[3] W. Ren, K.-H. Hung, R. Chao, Y. Li, H.-M. Wang, and Y. Tsao, "Robust audio-visual speech enhancement: Correcting misassignments in complex environments with advanced post-processing," in *O-COCOSDA*, 2024.