WHU-DKU Submission to the 4rd Audio-Visual Speech Enhancement Challenge

Jiarong Du[1,2], *Zhan Jin*[3,4], *Peijun Yang*[1,2], *Bang Zeng*[3,4], *Juan Liu*[2,1], *Ming Li*[3,4]

[1]School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China
[3]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center of Duke Kunshan University, Kunshan, China
[4]School of Computer Science, Wuhan University, Wuhan, China

1. Proposed Methods

Unlike last year's challenge, this year's mixed audio data contains reverberation and more diverse noise types.To address this complex acoustic environment, we propose a two-stage target speaker extraction approach. This method decouples speech separation from reverberation suppression, effectively resolving performance degradation issues of traditional end-to-end models in challenging acoustic conditions. The technical details of each stage are described below.

1.1. Target Speech Separation With Reverb

The first-stage separation model builds on the SOTA system AVSE Challenge of last year[1] with key improvements. Inspired by research demonstrating the benefits of incorporating facial expressions for audiovisual separation, we employ a pre-trained expression estimation network to extract facial expression features from speaker videos. Similarly to lip movement features,the refined expression features are concatenated with audio features along the channel dimension.

Additionally, to improve robustness in complex acoustic scenarios, We incorporated an power compression[2] process into the network, and introduce a progressive SI-SDR loss function inspired by prior work[3]: $\mathrm{L}_{\mathrm{total}} = \frac{l}{K}\sum_{k=l}^{K}\mathcal{L}^{(k)}$, where K=6 denotes the number of GridBlocks[4], the loss in the K-th layer is $\mathcal{L}^{(k)}$ and $\mathcal{L}_{\mathrm{total}}$ is the total loss of the separation model. This layered supervision mechanism enables the network to prioritize strong interference suppression in shallow layers while refining the details of speech reconstruction in deeper layers.

1.2. Reverberation Suppression Post-processing

The reverberant target speech from the first stage is fed into a pre-trained dereverberation network to produce clean target speech. We utilize the SGMSE[5] speech enhancement model, based on a conditional diffusion probabilistic framework. This model operates within the complex STFT domain using a stochastic differential equation (SDE) framework, significantly improving speech quality and robustness. The independent two-stage processing architecture effectively avoids optimization conflicts between separation and dereverberation tasks inherent in end-to-end models. While the post-processing module yields marginal gains in objective measurements, it provides perceptible reverberation reduction according to subjective evaluation.

2. Experiments

For Track 1, the provided training set was utilized for model training while the development set served for validation. Instead of directly using the pre-mixed speech, we dynamically mixed target speaker speech and interference speech at signal-to-noise ratios (SNRs) ranging from -18 dB to 6 dB. The data mixing procedure strictly followed the prescribed data preparation steps, including speech filtering and boundary ramping. During training, mixed audio segments and corresponding target speaker videos (sampled at 25 frames per second) were randomly truncated into 3-second chunks.

The Adam optimizer was employed with an initial learning rate of 0.001. This was halved whenever the best validation loss showed no improvement over three consecutive epochs. Training terminated automatically if no improvement was observed for ten consecutive epochs. For intermediate layer losses, we used the same SI-SDR-SE loss as [4]. All models were trained on eight NVIDIA A40 GPUs with 48GB RAM each, with the batch size fixed at 2. In System I, we employed energy compression. In System II, we conducted end-to-end training by combining lip movements. In System III, we integrated the above two methods and performed fine-tuning.

3. References

[1] Z. Jin, B. Zeng, Z. Li, X. Liu, and M. Li, "A target speaker extraction method for the 3rd audio-visual speech enhancement challenge," *System*, vol. 1, no. 2.932, pp. 0–876, 2024.

[2] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA express letters*, vol. 1, no. 1, 2021.

[3] Z. Hou, T. Sun, Y. Hu, C. Zhu, K. Chen, and J. Lu, "Sir-progressive audio-visual tf-gridnet with asr-aware selector for target speaker extraction in misp 2023 challenge," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 11–12.

[4] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[5] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

Table 1: *THE PROPOSED SYSTEMS' PERFORMANCE IN THREE OBJECTIVE METRICS*

System Name	PESQ	STOI	SISDR
Noisy	1.285795	0.508202	-25.943447
System I	1.947982	0.772565	-18.297653
System II	2.069336	0.798168	-17.809357
System III	2.160946	0.821719	-17.472612