Enhancement System Submission to the 3rd Audio-Visual Speech Enhancement Challenge

Zhan Jin[1], *Bang Zeng*[1], *Ming Li*[1], *Zhuo Li*[2], *Xin Liu*[2]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, China
[2]Hardware Engineering System, OPPO, Beijing, China

1. Proposed Methods

Our system is built upon AV-GridNet[1], the SOTA system from the previous AVSE Challenge. Target speaker's lip movement visual features are extracted from the lip ROIs, and then fused with the time-frequency features of noisy audio to gain synchronous audio-visual tensors, and be sent to a series of GridBlocks and a deconvolution module of TFGridNet[2], finally the audio of target speakers is extracted.

We make some changes to the original AV-GridNet to improve its speech enhancement performance in noisy environments. We fuse the audio and visual features by channel-wise concatenation only once before entering GridBlocks. Apart from the fusion policy, the unfold operation is also omitted. This operation is applied before entering the LSTM module in the GridBlocks, and is aimed at increasing the receptive field of the LSTM hidden units in the respective dimensions, either frame-wise or frequency-wise. Instead of unfold, we expand the channel dimension in the beginning in order to maintain the information flow of the whole system.

Three systems are implemented and evaluated with the test data. System I concatenates audio and visual tensors on the feature dimension followed by a linear layer to restore the audio-visual tensor's feature dimension the same as the original audio tensor. System II shares the same network structure with System I except that it applies channel-wise concatenation of audio and visual tensors before entering the GridBlocks. System III follows the process of SAV-GridNet system substitutes the original AV-GridNet with System II. In SAV-GridNet a frontend scene classifier is used to distinguish between speaker and non-speaker interferer.

2. Experiments

For track 1, we use the provided train set for training and dev set for validation. Dynamic mixing is applied for both training and validation phases. The mixing process exactly follows the data preparation steps including speech filtering and boundary ramping. Mix audio and the synchronous ROIs of 25 frames every second are randomly truncated to 3-second chunks. The Adam optimizer is used with the initial learning rate of 0.001, which will be halved if the best loss on the dev data is not improved within 3 consecutive epochs. The training process will stop if no improvement is seen within 10 consecutive epochs. For System I and II we use the SI-SDR-SE with Mixture-Constraint loss[2]. For System III, we finetune System II with a hybrid loss[3] starting with the learning rate of 0.0005. The scene classifier is also trained with the binary cross-entropy loss. All models are trained on 8 A40 of 48GB RAM, and the batch size is set to 32.

3. References

[1] Z. Pan, G. Wichern, Y. Masuyama, F. G. Germain, S. Khurana, C. Hori, and J. Le Roux, "Scenario-aware audio-visual tf-gridnet for target speech extraction," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[2] 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*, 2023.

[3] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," *arXiv preprint arXiv:2203.16843*, 2022.

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

System Name	PESQ	STOI	SISDR
Noisy	1.467	0.610	-5.494
System I	2.932	0.876	11.999
System II	2.997	0.879	12.434
System III	3.004	0.886	12.701