SIR-PROGRESSIVE AUDIO-VISUAL TF-GRIDNET WITH ASR-AWARE SELECTOR FOR TARGET SPEAKER EXTRACTION IN MISP 2023 CHALLENGE

Zhongshu Hou[1,2,*] *, Tianchi Sun*[1,2,*]*, Yuxiang Hu*[2]*, Changbao Zhu*[2]*, Kai Chen*[1,2]*, Jing Lu*[1,2]

[1] Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China
[2] NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China

{zhongshu.hou, tianchi.sun}@smail.nju.edu.cn, {chenkai, lujing}@nju.edu.cn, {yuxiang.hu, changbao.zhu}@horizon.cc

ABSTRACT

TF-GridNet has demonstrated its effectiveness in speech separation and enhancement. In this paper, we extend its capabilities for progressive audio-visual speech enhancement by introducing an attention-based audio-visual fusion module and a progressive learning strategy based on the signal-to-interference ratio (SIR). The model is integrated with a prior guided source separation (GSS) process for robust target speech extraction. A subsequent automatic speech recognition (ASR)-aware selector is employed to choose the enhancement output for better ASR performance. The proposed system achieves a final character error rate (CER) of 33.18% on the evaluation set and ranks first in the ICASSP 2024 Signal Processing Grand Challenge: Multimodal Information based Speech Processing (MISP) 2023 Challenge.

Index Terms— Audio-visual speech enhancement, SIR-progressive learning, ASR-aware selector, neural network

1. INTRODUCTION

Compared with previous audio-visual speech enhancement (AVSE) challenges, the ICASSP 2024 Signal Processing Grand Challenge: Multimodal Information based Speech Processing (MISP) 2023 Challenge† aims to explore innovative front-end technology for real-world application and its joint optimization with back-end ASR systems in more challenging scenarios.

TF-GridNet [1] is the state-of-the-art (SOTA) model of speech enhancement (SE) and separation, which can also be developed into an efficacious AVSE network by combining visual embeddings [2]. However, directly recovering clean speech under low signal-to-noise ratio (SNR) often leads to the degradation of speech components [3]. To tackle this, SNR-progressive learning [3] decomposes the SE task into a series of intermediate targets, each with a small SNR increment. However, it only addresses noise interference while disregarding other disturbances, such as interfering speech and reverberation. Jointly optimizing SE and ASR systems is feasible but faces the challenge of aligning the gradients of SE and ASR tasks, potentially leading to suboptimal ASR performance [4].

In this paper, we incorporate a feature-wise audio-visual attention module into TF-GridNet for effective AVSE, and propose a more comprehensive SIR-progressive learning approach, considering background noise, interfering speech and reverberation as interference. Moreover, we employ an ASR-aware selector to choose the enhanced speech with better ASR results. The proposed model ranks first in the MISP 2023 challenge.

* These authors contributed equally to this work.
† The website of the challenge is Multimodal Information Based Speech Processing (MISP) 2023 Challenge (mispchallenge.github.io).

2. THE PROPOSED SYSTEM

2.1. SIR-progressive AVSE

The diagram of the proposed SIR-progressive audio-visual TF-GridNet (SPAV-TFGridNet) is shown in the training stage 1 of Fig. 1. Let $\mathbf{S}$, $\mathbf{X} \in \mathbb{C}^{T\times F}$ denote clean and noisy complex spectrograms, where T and F denote the time and frequency dimensions, respectively. The visual embedding $\mathbf{V} \in \mathbb{R}^{1\times T\times F_V}$, with F_V the dimension of visual features, is generated from the grayscale images of lip movements through a 3-dimensional convolutional layer (Conv3D), a ResNet-18 layer and temporal linear interpolation, which are pre-trained on lip-reading task [5] and remain frozen throughout training. The audio representation after the k-th TF-GridNet block, denoted as $\mathbf{A}_k \in \mathbb{R}^{C\times T\times F_A}$, with F_A the dimension of audio features and $k = 1,2,\dots,K+1$, is fused with $\mathbf{V}$ through a feature-wise audio-visual attention (FAVA) module, as also depicted in the training stage 1 of Fig. 1.

The output stream of the k-th FAVA module is decoded by a 2-dimensional deconvolution layer (Deconv2D), estimating an intermediate target $\mathbf{S}_k$ with an increased SIR level, where the SIR gain between adjacent targets is denoted as Δ_{SIR} in dB. Besides considering the power of interfering speech, we progressively reduce the reverberation level of the intermediate outputs, which involves gradually decreasing the reverberation time of $\mathbf{S}_k$ relative to $\mathbf{S}$ in practical implementation.

2.2. ASR-aware selector

SE can effectively attenuate interference but may negatively impact ASR [6]. Considering that the output of GSS, $\mathbf{X}_{GSS}$, is guaranteed to hold the desired speech, we propose an ASR-aware selector to choose the preferred output between $\mathbf{X}_{GSS}$ and its subsequent AVSE result $\tilde{\mathbf{S}}_{k+1}$, guided by the provided ASR back-end [7]. As illustrated in the training stage 2 of Fig. 1, lower ASR loss $\mathcal{L}_{\text{MTL}}$ [7] indicates better ASR performance of the given speech, and selection label can be expressed as

$$y = \begin{cases} 1, & \mathcal{L}_{\text{MTL}} \text{ of } \mathbf{X}_{GSS} < \mathcal{L}_{\text{MTL}} \text{ of } \tilde{\mathbf{S}}_{k+1} \\ 0, & \mathcal{L}_{\text{MTL}} \text{ of } \mathbf{X}_{GSS} \geq \mathcal{L}_{\text{MTL}} \text{ of } \tilde{\mathbf{S}}_{k+1} \end{cases}. \quad (1)$$

Correspondingly, a one-hot vector $\boldsymbol{y} = [y, 1-y]$ is set as the learning target, with $y = 1$ signifying the selection of $\mathbf{X}_{GSS}$ and $y = 0$ signifying the selection of $\tilde{\mathbf{S}}_{k+1}$. A conformer-based ASR-aware selector, as depicted in the training stage 2 of Fig. 1, is utilized to predict the selection probability $\tilde{\boldsymbol{y}} = [\tilde{y}, 1-\tilde{y}]$. Note that $\tilde{y}$ is a real number between 0 and 1, and we set a threshold coefficient η for selection in the inference stage,

$$\text{Selection result} = \begin{cases} \mathbf{X}_{GSS}, & \tilde{y} > \eta \\ \tilde{\mathbf{S}}_{k+1}, & \tilde{y} \leq \eta \end{cases}, \quad (2)$$

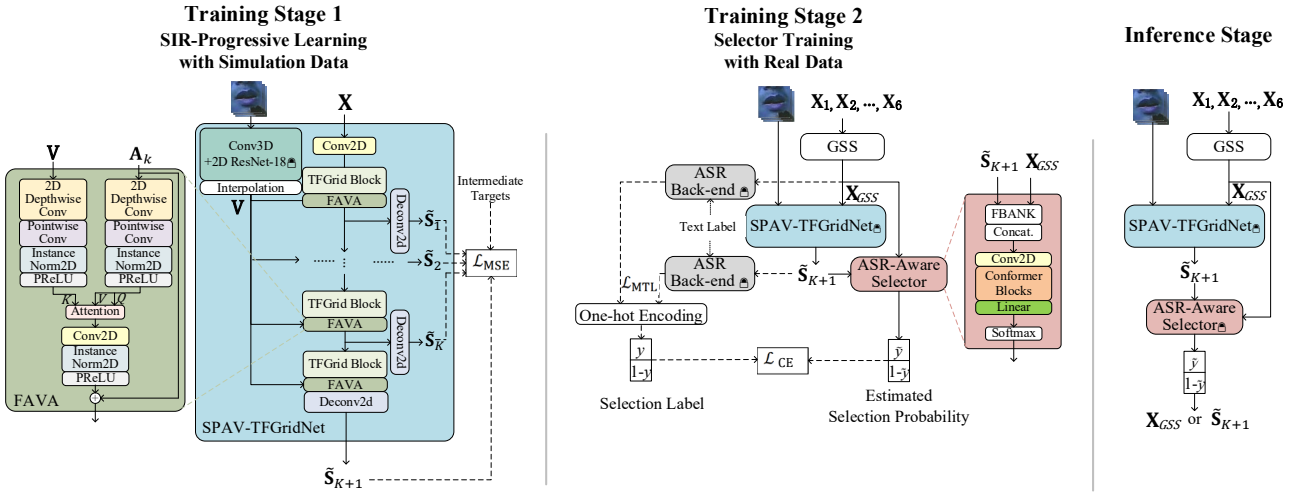

Fig. 1. The schematic diagram of the training and inference procedures of the proposed system.

2.3. Training process and loss functions

As illustrated in Fig. 1, the training procedure is divided into two stages. In the first stage, the SPAV-TFGridNet is trained with the spectral power compression loss function $\mathcal{L}_{\text{SPC}}$ [8] and the overall loss function for progressive learning is

$$\mathcal{L}_{\text{MSE}} = \sum_{k=1}^{K+1} \mathcal{L}_{\text{SPC}}(\tilde{\mathbf{S}}_k, \mathbf{S}_k). \quad (3)$$

In the second stage, ASR-aware selector is trained together with the SPAV-TFGridNet and the ASR back-end model, where only the parameters of selector are learnable. The cross-entropy loss function $\mathcal{L}_{\text{CE}}$ is utilized for training,

$$\mathcal{L}_{\text{CE}} = -\sum [\boldsymbol{y}\log\tilde{\boldsymbol{y}} + (\mathbf{1}-\boldsymbol{y})\log(\mathbf{1}-\tilde{\boldsymbol{y}})]. \quad (4)$$

3. EXPERIMENTS

3.1. Dataset and implementation details

The data we use for training is the MISP corpus [7], where the audios are segmented according to the provided speaker-dependent time stamps, and the middle-field videos are pre-processed to generate the region of interest (ROI) of lip movements based on the baseline approaches [7]. Specifically, we use the segmented monaural near-field speech, noise and room impulse responses (RIRs) to generate the simulation data in the first training stage, where dynamic mixing is applied and one audio clip contains 6 speakers at most. The input SIR ranges from -5 dB to 15 dB for noise and from -5 dB to 10 dB for each interference speaker. In the second training stage, 6-channel far-field data is employed to train the ASR-aware selector. The maximum length of an audio clip is 5 seconds during training.

The window and hop size of short-time Fourier transformation (STFT) are 40ms and 20ms, respectively, and a periodic Hanning widow is used. The ROI images are transformed to grayscale with pixel dimensions of 112×112. The SIR gain Δ_{SIR} is set to 4 dB and 5 TF-GridNet blocks are utilized, i.e. $K = 4$. $\mathbf{S}$ is convolved with the first [150,120,90,70,50]ms of RIRs to generate intermediate targets before noise is added. The threshold coefficient η is set to 0.58.

3.2. Results and discussion

The CER results on the evaluation testset are presented in Table 1. It can be seen that the proposed SPAV-TFGridNet without ASR optimization already outperforms the baseline system. The selector trained under the supervision of given ASR back-end further improves the accuracy of recognition. Employing $\eta = 0.58$ as the selection threshold notably reduces the CER compared with $\eta = 0.50$, indicating that the AVSE result $\tilde{\mathbf{S}}_{k+1}$ tends to better extract the target speaker.

Table 1. Results on the evaluation testset.

Systems	CER (%)
GSS	37.60
GSS+MEASE+Finetune [7] (baseline)	36.10
GSS+SPAV-TFGridNet	34.49
GSS+SPAV-TFGridNet+Selector ($\eta = 0.50$)	33.74
GSS+SPAV-TFGridNet+Selector ($\eta = 0.58$)	**33.18**

4. CONCLUSIONS

In this paper, we propose an audio-visual target speaker extraction model combining an SIR-progressive learning strategy and a conformer-based ASR-aware selector. The proposed network integrated with an initial GSS process achieves the best ASR performance in the MISP 2023 challenge.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 12274221).

6. REFERENCES

[1] Wang, Zhong-Qiu, et al. "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation." *ICASSP 2023*. IEEE, 2023.

[2] Pan, Zexu, et al. "Scenario-Aware Audio-Visual TF-GridNet for Target Speech Extraction." *ASRU*. IEEE, 2023.

[3] Tu, Yan-Hui, et al. "A multi-target SNR-progressive learning approach to regression based speech enhancement." *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020)*: 1608-1619.

[4] Hu, Yuchen, et al. "Gradient remedy for multi-task learning in end-to-end noise-robust speech recognition." *ICASSP 2023*. IEEE, 2023.

[5] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "Deep lip reading: A comparison of models and an online application." in *Proc. Interspeech*, 2018.

[6] Chen, Yu-Wen, Julia Hirschberg, and Yu Tsao. "Noise robust speech emotion recognition with signal-to-noise ratio adapting speech enhancement." *arXiv preprint arXiv:2309.01164* (2023).

[7] Wu, Shilong, et al. "The Multimodal Information Based Speech Processing (MISP) 2023 Challenge: Audio-Visual Target Speaker Extraction." *arXiv preprint arXiv:2309.08348* (2023).

[8] Li, Andong, et al. "On the importance of power compression and phase estimation in monaural speech dereverberation." *JASA express letters 1.1* (2021).