THE XMUSPEECH SYSTEM FOR AUDIO-VISUAL TARGET SPEAKER EXTRACTION IN MISP 2023 CHALLENGE

Longjie Luo[1†], *Tao Li*[2†], *Lin Li*[1*], *Qingyang Hong*[2*]

[1]School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China;
[2]School of Informatics, Xiamen University, Xiamen 361005, China;
{lilin,qyhong}@xmu.edu.cn

ABSTRACT

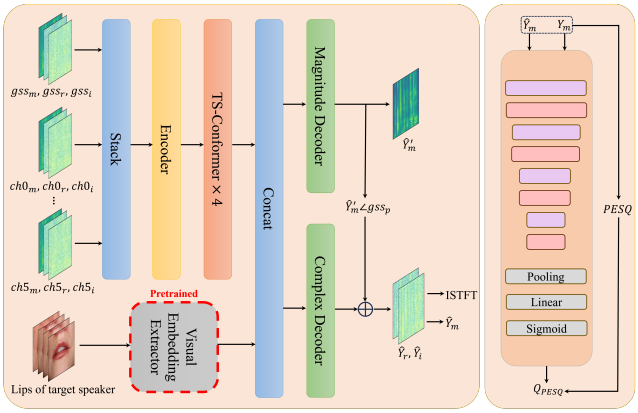
This paper presents the XMUSPEECH system for the Multimodal Information based Speech Processing (MISP) 2023 Challenge. The primary difficulty in this competition lies in effectively suppressing noise in mixtures while maintaining the integrity of the target speaker's speech signal. To achieve this, we introduce the AV-MCCMGAN, a multi-modal multi-channel speech enhancement model. This model employs the raw 6-channels audio as a supplement to the guided source separation (GSS) signal and utilizes the visual clues of the target speaker to further eliminate noise. Ultimately, our system achieves the character error rate (CER) values of 23.3% and 33.41% on the Dev and Eval set, ranking the third.

***Index Terms*—** MISP Challenge, Multi-channel, Multi-modal, Target Speaker Extraction

1. INTRODUCTION

Recently, a growing number of researchers have focused on audio-visual target speaker extraction (AVTSE). Regrettably, there is currently no publicly benchmark for AVTSE. To fill this gap, the MISP 2023 challenge concentrates on the AVTSE task [1]. To reduce difficulties, the oracle diarization results are provided. Therefore, the focus of this paper is to enhance the separated audio, that is, to suppress the noise without damaging the target signal. We first employ GSS [2] to extract the speech signals of each target speaker from 6-channels mixture audio. Based on the conformer-based metric generative adversarial network (CMGAN) [3], we then propose the audio-visual multi-channel CMGAN (AV-MCCMGAN) to further improve the quality of the target speaker's speech. Simultaneously, we modify the loss function to address the issue of domain mismatch between the far-field input and the near-field label during the model training. Our system finally achieves the CERs of 23.3% and 33.41% on the Dev and Eval set.

†Equal contribution
*Corresponding author
This work was supported in part by the National Natural Science Foundation of China under Grants 62371407, 62276220, and 62001405.



(a) Generator (b) Discriminator

Fig. 1. Framework of the proposed AV-MCCMGAN.

2. SYSTEM OVERVIEW

2.1. Data Preparation

The training data comprises two components: 106.09 hours of the official training set and an additional 103.68 hours of simulation data obtained through the application of the official methodology and configuration.

2.2. System Architecture

In our system, GSS [2] utilizes the oracle diarization results to isolate the speech signals of the target speakers from the raw 6-channels audio. Nevertheless, it demonstrates several limitations: inadequate noise suppression; potential enhancement of interference signal rather than target signal in the presence of strong interference; and the risk of compromising the integrity of the target signal. To address these issues and harness the complete potential of the raw audio and visual information, we introduce AV-MCCMGAN. As shown in Fig.1, AV-MCCMGAN comprises both a generator and a discriminator. Let gss_m, gss_r, and gss_i denote the magnitude, real, and imaginary spectrum of the GSS-processed audio. The magnitude, real, and imaginary spectrum of the raw 6-channels audio are similarly named, e.g., $ch0_m$, $ch0_r$, and $ch0_i$. Upon inputting the aforementioned 21-channel data and the visual clues of the target speaker into the generator, the magnitude spectrum $\widehat{Y}'_m$ is directly mapped by magnitude decoder. After

Table 1. CER(%) and Deep Noise Suppression Mean Opinion Score (DNSMOS) of different systems on Dev set.

Method	System	Trainable Params(M)	Training Target	Data Augmentation	CER(%) ↓	DNSMOS		
						SIG ↑	BAK ↑	OVR ↑
M0	GSS	-	-	-	26.4	2.03	1.88	1.57
M1	Baseline	76.5	masking	Yes	26.3	3.18	3.32	2.63
M2	CMGAN	1.83	masking	No	28.2	2.90	2.96	2.36
M3	CMGAN		mapping	No	31.7	2.63	2.79	2.06
M4	MCCMGAN		masking	No	24.9	3.15	3.25	2.60
M5	MCCMGAN		mapping	No	24.6	3.17	3.26	2.62
M6	MCCMGAN		mapping	Yes	23.7	3.23	3.34	2.66
M7	AV-MCCMGAN		mapping	Yes	**23.3**	**3.24**	**3.34**	**2.67**

incorporating the GSS phase gss_p to acquire the magnitude-enhanced complex spectrum, the element-wise sum with the output of the complex decoder is employed to derive the final complex spectrum ($\widehat{Y}_r$,$\widehat{Y}_i$). The discriminator is utilized to emulate the PESQ score and forms a component of the generator loss function, denoted as $\mathcal{L}_{\text{GAN}}$, to optimize the PESQ score of the enhanced speech.

2.3. Loss Function

While training model with the MISP dataset, an issue of audio distortion arises, potentially stemming from the domain mismatch between far-field input and near-field label. For this reason, the generator loss function is designed as follows:

$$\mathcal{L}_{\text{genr}} = \alpha_1 \mathcal{L}_{\text{m}} + \alpha_2(\mathcal{L}_{\text{r}} + \mathcal{L}_{\text{i}}) + \alpha_3 \mathcal{L}_{\text{GAN}} \tag{1}$$

where α_1, α_2 and α_3 represent the weights assigned to the losses, $\mathcal{L}_{\text{m}}$, $\mathcal{L}_{\text{r}}$ and $\mathcal{L}_{\text{i}}$ denote the prediction loss of magnitude, real and imaginary spectrum, respectively.

3. RESULTS AND ANALYSIS

3.1. Ablation and Comparison

Table. 1 illustrates our system and its improvements. Compared with the baseline (M1), the CMGAN (M2) has significantly fewer trainable parameters, and it achieves competitive results even without visual clues and data augmentation. Subsequently, we changed the training target from masking to mapping. However, the results of M3 showed that the refined approach didn't exhibit any advantages. We further carried out experiments on multi-channel CMGAN (MCCMGAN) and different configurations. In comparison among M1, M2, and M4, the MCCMGAN showcased enhancements of 5.3% and 11.7% compared to the baseline and the system M2, respectively. Comparing M5 and M4, we found that mapping-based menthod is preferable to masking-based method in multi-channel condition. Furthermore, by integrating simulation data into our training process, M6 yielded improvements in both CER and DNSMOS [1]. Finally, we used the parameters of M6 to initialize M7 to train the multi-modal multi-channel model and achieved a CER of 23.3% on the Dev set, which is an 11% improvement over the baseline. The multi-system fusion of ours and baseline resulted in a CER of 33.41% on the Eval set, which obtained a 7.4% relative reduction from the baseline, ranking the third in the competition.

Table 2. CER(%) in TV background noise setting on Dev set.

Method	System	w TV	w/o TV
M0	GSS	28.8	24.4
M1	Baseline	28.9	24.3
M7	AV-MCCMGAN	**25.1**	**22.5**

3.2. Results in TV background noise setting

To further describe the strengths of our system, we conducted experiments in the noisy TV background setting, as detailed in Table. 2. Upon comparing M0, M1, and M7, it's obvious that our model exhibits a considerable advantage over both GSS and baseline in the presence of TV background noise, showcasing a 13.2% relative improvement over baseline. Furthermore, even in scenarios without TV background noise, our model still obtains a 7.4% relative improvement over the baseline. The demo of our system is available online[1].

4. CONCLUSION

In this paper, we explored speech enhancement schemes in complex scenarios. Our best result archived a CER of 33.41% on the Eval set, ranking the third out of all teams.

5. REFERENCES

[1] Shilong Wu et al., "The multimodal information based speech processing (misp) 2023 challenge: Audio-visual target speaker extraction," *arXiv preprint arXiv:2309.08348*, 2023.

[2] Christoph Boeddeker et al., "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018, vol. 1.

[3] Ruizhe Cao et al., "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 936–940.

[1] https://leroisoleil2023.github.io/MISP2023-XMU-System-Demo/