2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-7451-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSPW62465.2024.10627330

SUMMARY ON THE MULTIMODAL INFORMATION-BASED SPEECH PROCESSING (MISP) 2023 CHALLENGE

Hang Chen[1], Shilong Wu[1], Chenxi Wang[1], Jun Du[1,], Chin-Hui Lee[2], Sabato Marco Siniscalchi[2,4], Shinji Watanabe[3], Jingdong Chen[6], Odette Scharenborg[7], Zhong-Qiu Wang[3], Bao-Cai Yin[5], Jia Pan[5]*

[1] University of Science and Technology of China, China [2] Georgia Institute of Technology, USA
[3] Carnegie Mellon University, USA [4] Kore University of Enna, Italy [5] iFlytek, China
[6] Northwestern Polytechnical University, China [7] Delft University of Technology, The Netherlands
✉jundu@ustc.edu.cn

ABSTRACT

Historically, MISP challenges have focused on audio-visual speech recognition (AVSR), where they have been particularly successful in complex acoustic scenarios. However, even the most sophisticated AVSR systems have been found to have performance limitations. Inspired by traditional robust speech recognition systems, where speech enhancement as a front-end can significantly improve accuracy, the MISP2023 challenge focused on audio-visual target speaker extraction (AVTSE). The primary goal of AVTSE is to enhance speech quality by exploiting the lip movements of the target speaker, thereby improving the final recognition performance. This paper provides a comprehensive overview of the challenge framework, describes the results, and summarizes the effective strategies employed by the contributions. In addition, we analyze the prevailing technical hurdles and provide recommendations for future directions to spur further progress in the AVTSE field.

Index Terms— MISP challenge, audio-visual, target speaker extraction, robust speech recognition

1. INTRODUCTION

Automatic speech recognition (ASR) systems are increasingly used in real-world applications. However, they face significant challenges due to the complex and adverse nature of real-world acoustic environments. Strong noise, pronounced reverberation, and multiple simultaneous speakers significantly degrade ASR performance. These problems were addressed in the previous MISP2021 and MISP2022 challenges by allowing visual cues, including lip movements and facial expressions, to improve recognition accuracy, i.e., these challenges focused on audio-visual speech recognition (AVSR). Several international teams participated in the challenges. The lowest character error rates (CERs) achieved with and without oracle speaker diarization were 25.07% and 29.58%, respectively. Compared to the corresponding ASR models, these results represent a relative reduction of 21.45% and 5.04%, respectively, confirming the potential of fusing audio and visual modalities to improve recognition performance, especially in acoustically challenging environments.

Recent empirical observations indicate that, despite considerable progress in challenging acoustic environments, the margin of improvement in the most advanced AVSR systems is exhibiting diminution. This trend does not merely indicate the maturation of existing technologies but instead suggests an imminent necessity for a paradigmatic shift in using visual input to improve recognition performance. In considering the evolution of robust ASR systems, it is widely recognized that implementing speech enhancement as a front-end can ameliorate speech quality and elevate overall recognition performance. In alignment with this perspective, MISP2023 is dedicated to harnessing the visual cues derived from the lip movements of the target speaker to enhance the clarity of target speech, i.e., Audio-Visual Target Speaker Extraction (AVTSE).

This paper summarises the outcomes of the challenge. Specifically, Section 2 provides a brief overview of the challenge and presents the dataset, evaluation metric, and baseline. Section 3 describes the major techniques used by the submitted systems, offering a detailed analysis of their methodologies. Finally, Section 4 summarizes the key findings and proposes prospective directions for future research.

2. CHALLENGE OVERVIEW

The main goal of MISP2023 is to explore various ways of using visual input to improve ASR accuracy in challenging acoustic environments. Specifically, we propose the AVTSE task, which utilizes the visual cues from the target speaker's lip movements to extract more distinct target speech from the recorded mixed speech signal. Subsequently, the recognition performance of the extracted speech is evaluated using a pre-trained ASR model. In this context, CER is employed as the primary metric for evaluation.

The baseline model [1] is a sophisticated cascaded system that combines guided source separation (GSS) with the multimodal embedding aware speech enhancement (MEASE) model. The core of the MEASE model is a pre-trained multimodal embedding extractor (MEE) coupled to an embedding-aware enhancement network. The MEE module integrates noisy Mel Filter Bank (FBANK) features with lip frame data to generate a detailed multimodal embedding. This embedding and the noisy log power spectra (LPS) feature are then processed by the enhancement network, culminating in estimating a magnitude mask.

The training of the MEASE model can be divided into two distinct phases. First, the model is trained on a simulated data set using the Mean Squared Error (MSE) as the loss function. This simulated training set is obtained by corrupting near-field speech from the MISP2021-AVSR corpus [2] with corresponding far-field noise. The second training phase involves fine-tuning the MEASE model with the back-end ASR model. This phase is critical because it uses

*corresponding author

Table 1. Top 3 ranking teams in terms of CER (in %) and their major techniques.

Team	Model	Audio Input			Lip Embedding		Fusion	Training with ASR	Training Strategy	Post-processing	CER (in %)
		Raw 6 Channel	GSS	Average	Pre-training	Jointly Training					
NJU-AALab	SPAV-TFGridNet	×	✓	×	✓	×	FAVA	×	SIR-progressive	ASR-aware selector	33.18
NWPU-ASLP	MSXF	✓	✓	✓	×	✓	Concat	✓	AQMS	\	33.21
XMU-SpeechLab	AV-MCCMGAN	✓	✓	×	✓	×	Concat	×	Adversarial training	Multi-system fusion	33.41

far-field data, effectively bridging the gap between simulated and real-world scenarios. During this fine-tuning phase, the parameters of the ASR model are kept constant. Training focuses exclusively on connectionist temporal classification (CTC) and attention loss.

During the evaluation phase, the available data was limited to far-field audio and mid-field video. The newly released evaluation set was derived by augmenting the existing MISP2022 evaluation set with additional sessions that present more challenging overlapping speech between female speakers.

3. SUBMITTED SYSTEMS

Six (6) teams submitted their results, and the CER for the top 3 teams is summarized in Table 1. The NJU-AALab team [3] won the challenge with the lowest CER of 33.18%. They extended a state-of-the-art (SOTA) audio-only SE model, TF-GridNet, into the SPAV-TFGridNet by incorporating a feature-wise audio-visual attention (FAVA) module and a progressive training strategy based on the signal-to-interference ratio (SIR). In addition, a subsequent ASR-aware selector was used to judiciously choose between the SPAV-TFGridNet and GSS outputs for the lowest CER. Second place went to the NWPU-ASLP team [4], who presented a novel audio-quality-based multi-strategy (AQMS) approach. They employed different extraction strategies depending on the DNSMOS OVRL scores of the audio, skillfully balancing noise removal with speech preservation and improving the subsequent performance of ASR systems. Subsequently, the XMU-SpeechLab team [5] proposed an innovative audio-visual multichannel conformer-based metric generative adversarial network (AV-MCCMGAN). Together with multi-system fusion strategies, they achieved a commendable third place.

We summarize the top three submissions from five critical perspectives: audio and visual inputs, audio-visual fusion, training strategy, and post-processing. This analysis aims to provide a comprehensive overview and critical assessment of recent advances in AVTSE and suggest potential avenues for future investigation.

Audio input: The output of the prior GSS is advantageous for minimizing speech distortion during multichannel audio processing. However, in the face of overwhelming noise interference, GSS may inadvertently eliminate the target speaker's speech signals. With this in mind, both [4] and [5] recommend using raw 6-channel speech to compensate for the loss of target speaker information.

Visual input: Lip embeddings extracted from pre-trained lip reading models have become a mainstream choice [3, 5]. Since SE often relies on the audio modality, we also argue that pre-trained lip embeddings can more effectively circumvent the detrimental effects of visual redundancies on enhanced speech quality.

Fusion: Although frame-level concatenation remains the dominant method for audio-visual fusion [4, 5], we argue that attention-based fusion may be a superior approach to address the temporal resolution discrepancies and potential misalignments between audio and visual streams.

Training Strategy: This is an important innovation for several teams. [3] and [4] introduced the SIR-progressive learning and the AQMS approach, respectively, to circumvent the problem of over-enhancement that degrades ASR performance. Meanwhile, [5] proposed an adversarial training approach to optimize the model to improve perceptual evaluation of speech quality (PESQ). Maintaining a perception-distortion trade-off is beneficial for achieving a task-generic SE model. Furthermore, direct joint training with the ASR back-end did not show a clear advantage but may limit the generalizability.

Post-processing: The essence of post-processing lies in pre-selection the most promising outcomes. The ASR-aware selector, proposed by [3], represents a commendable endeavor in this direction and has demonstrated notable efficacy.

However, all submissions predominantly adhere to conventional data simulation and multi-channel speech processing approaches. This reliance on traditional methodologies suggests considerable room for enhancement and innovation.

4. CONCLUSION

This paper offers an overview of the MISP2023 challenge and encapsulates the achievements of the leading teams. Our analysis indicates that more realistic data simulation techniques and DNN-based multi-channel speech processing methodologies emerge as two promising avenues for AVTSE systems. We anticipate that the insights and methods delineated herein will help future research endeavors and significantly propel the frontiers of knowledge in this domain.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

6. REFERENCES

[1] Shilong Wu, Chenxi Wang, Hang Chen, et al., "The multimodal information based speech processing (MISP) 2023 challenge: Audio-visual target speaker extraction," in *Proc. ICASSP 2024*, 2024.

[2] Hang Chen, Hengshun Zhou, Jun Du, et al., "The first multimodal information based speech processing (MISP) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022, pp. 9266–9270.

[3] Zhongshu Hou, Tianchi Sun, Yuxiang Hu, et al., "Sir-progressive audio-visual tf-gridnet with ASR-aware selector for target speaker extraction in MISP 2023 challenge," in *Proc. ICASSP 2024*, 2024.

[4] Runduo Han, Xiaopeng Yan, Weiming Xu, et al., "An audio-quality-based multi-strategy approach for target speaker extraction in the MISP 2023 challenge," in *Proc. ICASSP 2024*, 2024.

[5] Longjie Luo, Tao Li, Lin Li, et al., "The XMUSpeech system for audio-visual target speaker extraction in MISP 2023 challenge," in *Proc. ICASSP 2024*, 2024.