THE MULTIMODAL INFORMATION BASED SPEECH PROCESSING (MISP) 2023 CHALLENGE: AUDIO-VISUAL TARGET SPEAKER EXTRACTION

Shilong Wu[1], *Chenxi Wang*[1], *Hang Chen*[1], *Yusheng Dai*[1], *Chenyue Zhang*[1],
Ruoyu Wang[1], *Hongbo Lan*[1], *Jun Du*[1], *Chin-Hui Lee*[2], *Jingdong Chen*[3],
Sabato Marco Siniscalchi[2,5], *Odette Scharenborg*[6], *Zhong-Qiu Wang*[4], *Jia Pan*[7], *Jianqing Gao*[7]

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

ABSTRACT

Previous Multimodal Information based Speech Processing (MISP) challenges mainly focused on audio-visual speech recognition (AVSR) with commendable success. However, the most advanced back-end recognition systems often hit performance limits due to the complex acoustic environments. This has prompted a shift in focus towards the Audio-Visual Target Speaker Extraction (AVTSE) task for the MISP 2023 challenge in ICASSP 2024 Signal Processing Grand Challenges. Unlike existing audio-visual speech enhancement challenges primarily focused on simulation data, the MISP 2023 challenge uniquely explores how front-end speech processing, combined with visual clues, impacts back-end tasks in real-world scenarios. This pioneering effort aims to set the first benchmark for the AVTSE task, offering fresh insights into enhancing the accuracy of back-end speech recognition systems through AVTSE in challenging and real acoustic environments. This paper delivers a thorough overview of the task setting, dataset, and baseline system of the MISP 2023 challenge. It also includes an in-depth analysis of the challenges participants may encounter. The experimental results highlight the demanding nature of this task, and we look forward to the innovative solutions participants will bring forward.

Index Terms— MISP challenge, target speaker extraction, multimodality, real-world scenarios

1. INTRODUCTION

In real-world scenarios, the complex and adverse acoustic environment is one of the main challenges of automatic speech recognition (ASR) and other back-end tasks. The strong noise, reverberation, and multi-speaker interference result in a serious impact on the system performance. Effective front-end speech processing technologies, like speech enhancement and speech separation [1, 2], have been proven to play a significant role in improving speech quality, thereby enhancing the performance of back-end systems. Recently, research on the cocktail-party problem [3] has shown that people can naturally track the speech of the target speaker from the interference of multiple speakers' conversations and background noise. Inspired by this, researchers have begun to focus on target speaker extraction (TSE), which estimates the speech of the target speaker within a mixed audio stream, leveraging a diverse array of clues including auditory, spatial, visual, and other [4].

Research on TSE systems has mainly focused on audio-only TSE (AOTSE) [5, 6], which utilizes the pre-registered speech from speakers as clues. This method is marked by its inherent simplicity, as it circumvents the need for supplementary equipment. Nonetheless, the development of AOTSE is constrained by several factors, including the challenges associated with acquiring pre-registered audio in real-world scenarios, the potential similarities of acoustic features among multiple speakers, and the presence of significant noise interference [7]. Recently, research [8] in neuroscience suggests that the visual modality, including facial and lip movements, can significantly influence humans' auditory attention, enhancing speech perception by providing additional information about the speaker, especially in noisy environments [9]. In real-world scenarios, acquiring visual clues is common, which can effectively overcome the challenges encountered by AOTSE. As a result, an increasing number of researches focus on audio-visual TSE (AVTSE) [10, 11]. Unfortunately, no publicly available benchmark currently exists for AVTSE. To fill this gap, the Multi-modal Information based Speech Processing (MISP) 2023 challenge focuses on the AVTSE task.

Expanding the scope to front-end speech processing technology, there are already several relevant challenges within the field of speech enhancement. These include the audio-only speech enhancement (AOSE) challenges like the Deep Noise Suppression (DNS) Challenge [12], the Clarity Challenge [13], and multimodal challenges such as the COG-MHEAR Audio-Visual Speech Enhancement Challenge [14]. Nevertheless, there are two main problems with current challenges. Firstly, the evaluation data is either the simulation data obtained by adding a single type of noise or interference speech to clean speech, or it is recorded in a real scene but speakers just read specific sentences or word arrangements. However, in real-life scenarios, people's conversations typically do not have a specific topic, and they encounter complex acoustic environments with multiple types of noise, reverberation, and interference from other speakers, which can lead to a mismatch between simulation and reality. Secondly, these challenges often utilize metrics such as the Deep Noise Suppression Mean Opinion Score (DNSMOS) [15], the Short-Time Objective Intelligibility (STOI) [16], and the Perceptual Evaluation of Speech Quality (PESQ) [17] to evaluate speech quality, or invite staff to score based on their actual listening experience. Research has shown that enhanced speech may experience distortion in terms of comprehensibility, which will worsen the performance of back-end ASR systems [18]. Therefore, evaluating its impact on the performance of back-end systems is also extremely important.

In the previous MISP challenges [19, 20], we released a large distant multi-microphone conversational Chinese audio-visual corpus that focuses on real home-TV scenarios, where 2-6 speakers

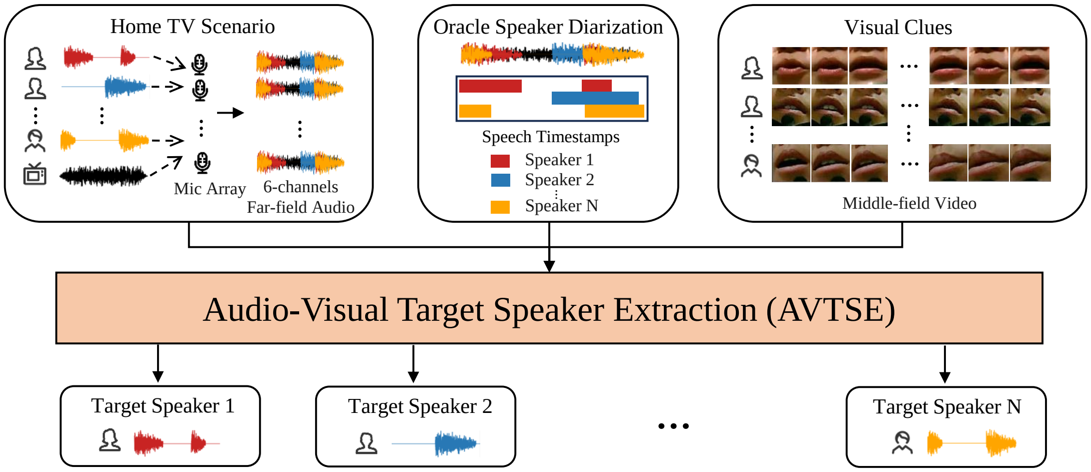

Fig. 1. Overview of the AVTSE task in MISP 2023 challenge.

freely converse without specific topics. Building upon the above analysis, the MISP 2023 challenge aims to improve the accuracy of the back-end ASR system through the AVTSE system in real-world scenarios using the MISP corpus. Specifically, we will use a pre-trained ASR model to decode the speech output from the AVTSE system, with character error rate (CER) as the evaluation metric. Through this challenge, we hope to promote researchers' attention to the AVTSE system, and provide new ideas for the front-end technology application in real scenarios and joint optimization with back-end systems. In this paper, we will introduce the task setting, dataset, and baseline system of the MISP 2023 challenge and conduct an in-depth analysis of the potential difficulties faced by participants. More details can be found on the website[1].

2. DATASET AND TASK SETTING

2.1. Dataset Scenarios and Composition

The MISP corpus [19] focuses on real home-TV scenarios: 2-6 people communicating with each other with TV noise and reverberation in the background. In this scenario, speakers engage in spontaneous conversations without specific topics, posing a challenge due to the significant speech overlap and diversity. Furthermore, in some sessions, strong background noise from television is present, where television programs such as dramas, news, music, and interviews may be playing, further exacerbating the complexity, especially for front-end systems.

We use the training set of AVSR corpus of MISP 2021 challenge [21], with a duration of 106.09 hours, including 21 rooms and 200 speakers. In the training set, the data includes near/middle/far-field audio and middle/far-field videos, allowing participants to choose freely. Additionally, we use the development set released in the MISP 2022 challenge [20], with a duration of 2.51 hours. For the evaluation set, in addition to the MISP 2022 evaluation set, we will also add some new sessions focusing on female dialogue scenarios. Our research has revealed that distinguishing voices in female dialogue scenarios is more challenging, thus increasing the difficulty level for participants. For development and evaluation sets, participants can only use far-field audio and middle-field video.

2.2. Task Overview and Evaluation

The MISP 2023 challenge focuses on the AVTSE task, as shown in Figure 1. Participants must utilize multi-channel far-field audio along with the target speaker's middle-field video to extract the target speaker's speech from audio recordings containing overlapping voices of multiple speakers and background noise. In one session, each speaker is sequentially treated as the target speaker. In addition, we will also provide the oracle diarization results because this is the first edition of the AVTSE challenge, and not providing them would bring greater difficulties to participants. This year, we will open two rankings based on whether external data, apart from the MISP dataset, has been utilized. However, it is worth noting that we will only submit the ranking which is without using them to ICASSP SPGC. This is because we encourage technological innovation rather than relying on large amounts of data.

The objective of the MISP 2023 challenge is to enhance the front-end AVTSE system to extract clearer speech of the target speaker, thereby improving the accuracy of speech recognition while keeping the back-end ASR system unaltered. We will provide a pre-trained ASR model [22], including both model parameters and code. To better investigate the role of the AVTSE system, we only use the audio-only ASR part of that paper. Participants can test the extracted speech in the development set to make adjustments to the AVTSE model. For ranking, we use the character error rate (CER) as the evaluation metric:

$$\text{CER} = (S + D + I)/N \times 100\% \tag{1}$$

where, S, D, and I represent the number of substitutions, deletions, and insertions. N is the number of characters in ground truth. The lower the CER, the higher the ranking.

During the evaluation stage, participants must submit the extracted speech, which we will decode and use to calculate CER. Therefore, participants do not need to modify the ASR model. However, they can still conduct joint training on the front-end and back-end systems with fixed back-end parameters. This is also an avenue we encourage participants to explore. Furthermore, we will calculate the DNSMOS P.835 [15] as a reference to explore the relationship between speech auditory quality and back-end tasks.

3. BASELINE SYSTEM

3.1. Data Preparation

To achieve complete alignment between speech signals, our initial step involves utilizing near-field data to simulate far-field conditions as the training data. Owing to potential noise interference or unclear speech in certain sections of the raw near-field audio data, it is imperative to conduct data cleaning on the raw training set. Firstly, we segment the near-field speech based on timestamps, ensuring that each

[1] https://mispchallenge.github.io/mispchallenge2023

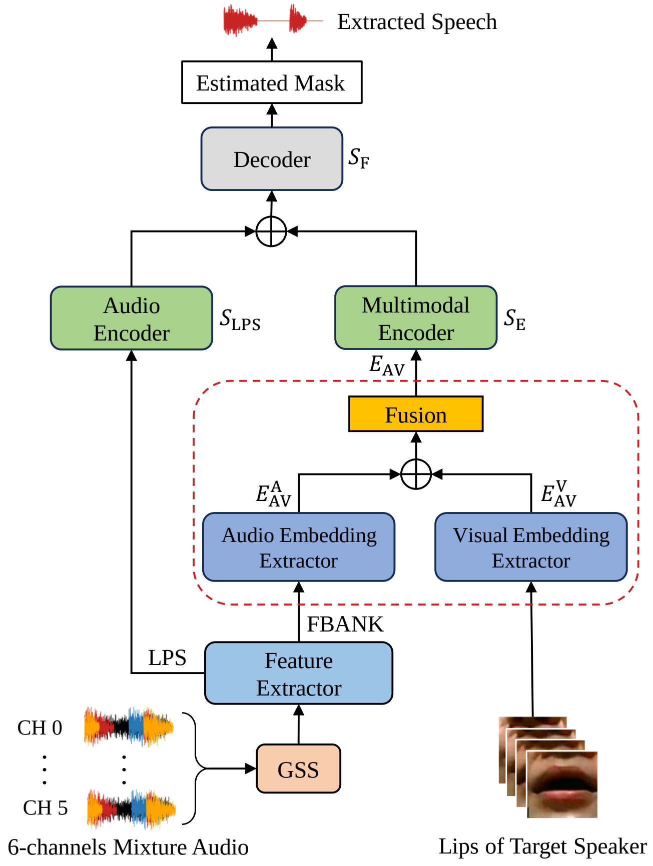

Fig. 2. Diagram of the baseline system.

segment contains the speech of a single speaker. Then we employ the DNSMOS P.835 to identify segments with high speech quality. We meticulously screen the training set, retaining only those near-field speech segments with high overall quality (OVRL) scores. This rigorous selection process results in the inclusion of a total of 22.13 hours of high-quality near-field speech segments. For model training, we randomly combine near-field speech and add noise with different signal-to-noise ratios (SNR) from -10dB to +20dB and reverberation related to room parameters to simulate 6-channels far-field audio. To minimize the mismatch with real data as comprehensively as possible, the noise we utilize is extracted from the far-field audio within the raw training set, containing a large amount of interference from the speaker's voice, as well as environmental and TV noise.

3.2. System Architecture

As shown in Figure 2, the baseline system is mainly based on our recent work - the multimodal embedding aware speech enhancement (MEASE) [9] model, which has achieved "SOTA" in the field of audio-visual speech enhancement (AVSE). Building upon this foundation, we leveraged the oracle diarization results to conduct guided source separation (GSS) [23] on 6-channels mixture audio to initially mitigate the impact of the overlapping speech. Then we use the MEASE model to further extract the speech of the target speaker. The MEASE model comprises a multimodal embedding extractor (in red dashed box) and an embedding-aware enhancement network. We first extract FBANK features and noisy log-power spectra (LPS) features from the audio output of GSS. Subsequently, we use the pre-trained embedding extractor to obtain the deep embeddings from both the FBANK ($A_{\rm FBANK}$) and lip frames (V) of the target speaker. Both audio and visual embedding extractors consist of a spatiotemporal convolution followed by an 18-layer ResNet, but the structure is slightly different and can be represented as follows:

$$\begin{aligned} E_{\rm AV}^{\rm A} &= \text{ResNet18}_{\rm 1D}(\text{BN}(\text{ReLU}(\text{Conv}_{\rm 1D}(A_{\rm FBANK})))) \\ E_{\rm AV}^{\rm V} &= \text{ResNet18}_{\rm 2D}(\text{MP}_{\rm 3D}(\text{BN}(\text{ReLU}(\text{Conv}_{\rm 3D}(V))))) \end{aligned} \tag{2}$$

where ReLu$(\cdot)$, BN$(\cdot)$ and MP$_{\rm 3D}(\cdot)$ represents the ReLU activation, the batch normalization, and the spatiotemporal max-pooling layer, respectively. Then, we fused the audio and visual embeddings using 2-layers of BiGUR. For encoders and decoder, we use the different numbers of 1D-ConvBlocks, where, $N_{\rm LPS}$=5, $N_{\rm E}$=10, and $N_{\rm F}$=15. At the end of the decoder, the hidden representation is activated by a sigmoid activation to generate a magnitude mask. We utilize the ideal ratio mask (IRM) [24] as the learning target and the mean square error (MSE) $\mathcal{L}_{\rm MSE}$ between target IRM and estimated mask as the loss function.

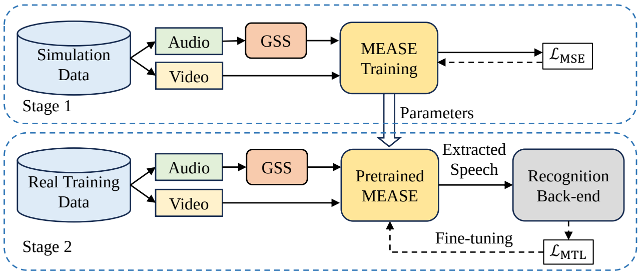

Fig. 3. The two-stage training process of the baseline system.

3.3. Training Process

As shown in Figure 3, the training process of the baseline system encompasses two stages. Firstly, we trained the MEASE model using the simulated data with $\mathcal{L}_{\rm MSE}$ as the loss function. However, this training approach inevitably leads to some degree of distortion in the extracted speech since it does not consider the back-end recognition task, thereby impacting the accuracy of the recognition system. Consequently, in the second stage, we fine-tuned the pre-trained MEASE model using the recognition back-end. Furthermore, to mitigate the issue of mismatch between simulated data and real-world scenarios, we utilized the real far-field data from the training set in the second stage. Because we require that the parameters of the back-end system cannot be changed, we freeze the recognition model and only use the loss returned by it. In our provided ASR back-end [22], the loss function employed is denoted as $\mathcal{L}_{\rm MTL}$:

$$\mathcal{L}_{\rm MTL} = \lambda \log P_{\rm ctc}(Y|X) + (1-\lambda)\log P_{\rm att}(Y|X) \tag{3}$$

where X and Y denote the encoder output and the target sequences, respectively. λ is the weight factor between the CTC loss and the attention cross entropy (CE) loss [25]. Here we set $\lambda = 0.3$.

4. RESULTS AND ANALYSIS

4.1. Baseline Results

Table 1 shows the results of different front-end systems in the speech recognition evaluation metric and DNSMOS P.835. Among these systems, AEASE is a simplified version of MEASE, as it does not utilize the visual modality. The results of "GSS+MEASE+Finetune" serve as our final baseline results. First, the comparison of the results of Beamforming [26] and GSS proves the importance of separating the speaker's speech from the mixed audio in challenging acoustic scenes. Incorporating the AEASE model on top of GSS, we observed an improvement in the DNSMOS, which reflects better realistic auditory quality. However, in terms of ASR metrics, it performed

Table 1. Detailed CER (%) and DNSMOS results of different front-end systems on the development set.

System	Recognition Back-end				DNSMOS		
	S	D	I	CER	SIG	BAK	OVR
Beamforming	31.0	7.2	4.8	43.0	1.45	1.35	1.20
GSS	20.0	4.2	2.2	26.4	1.78	1.80	1.35
GSS+AEASE	21.5	4.9	2.2	28.6	1.97	2.08	1.43
GSS+MEASE	20.4	4.4	2.2	27.0	2.01	2.14	1.46
GSS+MEASE+Finetune	19.8	4.7	1.8	26.3	2.03	2.27	1.50

worse than GSS. This observation underscores that while AEASE can filter out more noise, it also introduces speech distortion simultaneously. Replacing the AEASE model with the MEASE model, which incorporates the visual modality of the target speaker, leads to a notable enhancement in the DNSMOS. Additionally, the introduction of the visual modality can enable the model to extract more of the target speaker's speech, consequently yielding advancements in speech recognition results. Nevertheless, there remains a gap when compared to GSS. In order to balance back-end performance and noise suppression, we use the back-end recognition system to finetune the MEASE model. This enables the model to process signals in a targeted manner to improve the performance of the back-end system. The results show that the model can better preserve the target speaker's speech and remove more irrelevant noise, while the speech recognition results are also improved.

However, although our baseline system surpasses GSS in terms of DNSMOS, its speech recognition performance is only comparable to that of GSS. Due to noise suppression, it is inevitable to introduce more deletion errors. This is sufficient to demonstrate that in real scenarios, our task is very challenging. Therefore, how to improve the AVTSE system to enhance the accuracy of the backend recognition system requires more in-depth research by participants.

4.2. Difficulty Analysis

Figure 4 is an example of a mixture with multi-speaker interference and strong TV background noise. According to the speech timestamps, it can be seen that this is a complex acoustic scene with 2 to 3 speakers speaking simultaneously. We compared the spectrograms and the recognition results obtained from different systems. Firstly, we directly perform ASR decoding on near-field audio and obtain the result that is likely to be at the theoretical limit. Then, we compared beamforming, GSS, and baseline AVTSE system for far-field audio. According to beamforming's spectrogram, it is evident that this is a highly noisy environment so the backend system cannot distinguish the target speaker's speech, resulting in a very poor recognition result. In contrast, GSS can separate the speech of the target speaker but still contains a significant amount of noise. Our baseline AVTSE system further reduced noise, resulting in the best recognition results on far-field audio. However, there is still a significant quality gap between the extracted speech and near-field audio.

As shown in the green boxes in Figure 4, there is a notable presence of the interference speakers' speech in the far-field audio, affecting the clarity of the speaker's speech. However, due to the influence of the visual modality, AVTSE has the capability to filter out this interference. As demonstrated in the blue boxes, the certain background noise that remains in GSS output will largely be eliminated by AVTSE, consequently correcting the back-end recognition results. Nevertheless, as shown in the yellow box, our baseline system inevitably filters out some vocal components while suppressing noise, resulting in deletion errors. Therefore, there is still significant potential for improvement within the AVTSE system. Extracting the target speaker's speech while suppressing noise is challenging.

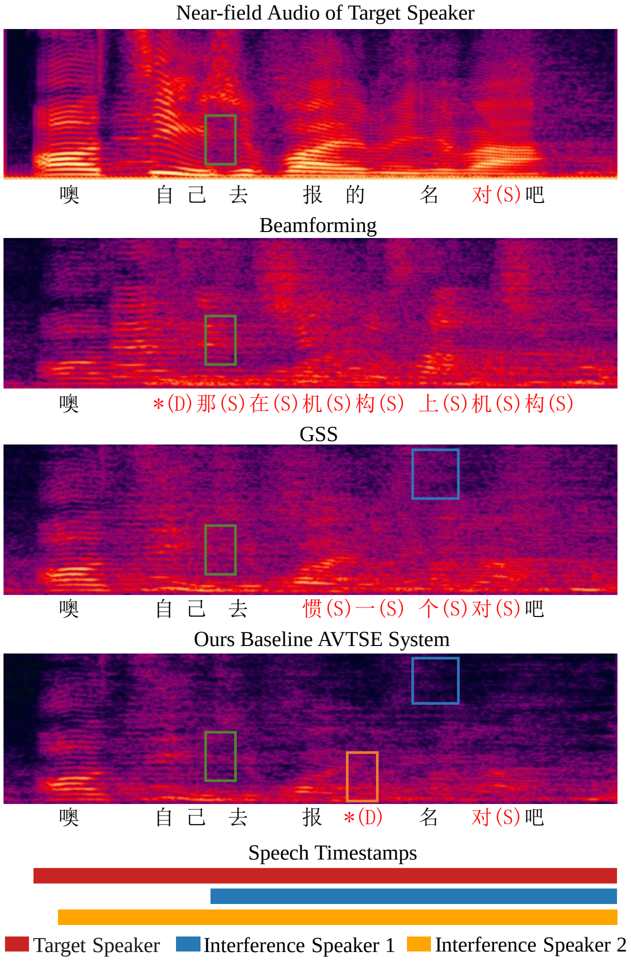

Fig. 4. Spectrograms and recognition results of an example.

5. CONCLUSIONS

In this paper, we provide a detailed description of the dataset, task setting, and baseline system for the MISP 2023 challenge, which is the first benchmark of the AVTSE task. We also conducted a deep analysis of the baseline experimental results, highlighting that the AVTSE task continues to hold significant research potential in real-world scenarios. In the future, we plan to explore the solutions for AVTSE systems on long recordings and incorporate the subjective listening test to further research the relationship between the real speech auditory quality and the performance of back-end tasks.

Acknowledgement: We would like to thank Shinji Watanabe for his valuable comments.

6. REFERENCES

[1] Felix Weninger, Hakan Erdogan, Shinji Watanabe, et al., “Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr,” in *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*. Springer, 2015, pp. 91–99.

[2] Tom O'Malley, Arun Narayanan, Quan Wang, et al., “A conformer-based asr frontend for joint acoustic echo cancellation, speech enhancement and speech separation,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 304–311.

[3] Adelbert W Bronkhorst, “The cocktail-party problem revisited: early processing and selection of multi-talker speech,” *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[4] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, et al., “Neural target speech extraction: An overview,” *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[5] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, et al., “Single channel target speaker extraction and recognition with speaker beam,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[6] Jakub Janský, Jiří Málek, Jaroslav Čmejla, et al., “Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 676–680.

[7] Marc Delcroix, Katerina Zmolikova, Tsubasa Ochiai, et al., “Compact network for speakerbeam target speaker extraction,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6965–6969.

[8] Elana Zion Golumbic, Gregory B Cogan, Charles E Schroeder, et al., “Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”,” *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[9] Hang Chen, Jun Du, Yu Hu, et al., “Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement,” *Neural Networks*, vol. 143, pp. 171–182, 2021.

[10] Andrew Owens and Alexei A Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 631–648.

[11] Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, et al., “Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues.,” in *INTERSPEECH*, 2019, pp. 2718–2722.

[12] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, et al., “Icassp 2023 deep noise suppression challenge,” 2023.

[13] SN Graetzer, Jon Barker, Trevor J Cox, et al., “Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing,” in *Proc. INTERSPEECH*, 2021, pp. 686–690.

[14] Andrea Lorena Aldana Blanco, Cassia Valentini-Botinhao, Ondrej Klejch, et al., “Avse challenge: Audio-visual speech enhancement challenge,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[15] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, “Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[16] Cees H Taal, Richard C Hendriks, Richard Heusdens, et al., “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[17] Antony W Rix, John G Beerends, Michael P Hollier, et al., “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[18] Kazuma Iwamoto, Tsubasa Ochiai, Marc Delcroix, et al., “How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR,” in *Proc. INTERSPEECH*, 2022, pp. 5418–5422.

[19] Hang Chen, Hengshun Zhou, Jun Du, et al., “The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9266–9270.

[20] Zhe Wang, Shilong Wu, Hang Chen, et al., “The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] Hang Chen, Jun Du, Yusheng Dai, et al., “Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis,” in *Proc. INTERSPEECH*, 2022, pp. 1766–1770.

[22] Yusheng Dai, Hang Chen, Jun Du, et al., “Improving audio-visual speech recognition by lip-subword correlation based visual pre-training and cross-modal fusion encoder,” in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 2627–2632.

[23] Desh Raj, Daniel Povey, and Sanjeev Khudanpur, “GPU-accelerated Guided Source Separation for Meeting Transcription,” in *Proc. INTERSPEECH*, 2023, pp. 3507–3511.

[24] Christopher Hummersone, Toby Stokes, and Tim Brookes, “On the ideal ratio mask as the goal of computational auditory scene analysis,” in *Blind source separation: advances in theory, algorithms and applications*, pp. 349–368. Springer, 2014.

[25] Shinji Watanabe, Takaaki Hori, Suyoun Kim, et al., “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[26] Xavier Anguera, Chuck Wooters, and Javier Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.