The Multimodal Information Based Speech Processing (MISP) 2025 Challenge: Audio-Visual Diarization and Recognition

Ming Gao[1], *Shilong Wu*[1], *Hang Chen*[1], *Jun Du*[1,*], *Chin-Hui Lee*[2], *Shinji Watanabe*[3], *Jingdong Chen*[4], *Siniscalchi Sabato Marco*[5], *Odette Scharenborg*[6]

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

✉jundu@ustc.edu.cn

Abstract

Meetings are a valuable yet challenging scenario for speech applications due to complex acoustic conditions. This paper summarizes the outcomes of the MISP 2025 Challenge, hosted at Interspeech 2025, which focuses on multi-modal, multi-device meeting transcription by incorporating video modality alongside audio. The tasks include Audio-Visual Speaker Diarization (AVSD), Audio-Visual Speech Recognition (AVSR), and Audio-Visual Diarization and Recognition (AVDR). We present the challenge's objectives, tasks, dataset, baseline systems, and solutions proposed by participants. The best-performing systems achieved significant improvements over the baseline: the top AVSD model achieved a Diarization Error Rate (DER) of 8.09%, improving by 7.43%; the top AVSR system achieved a Character Error Rate (CER) of 9.48%, improving by 10.62%; and the best AVDR system achieved a concatenated minimum-permutation Character Error Rate (cpCER) of 11.56%, improving by 72.49%.

Index Terms: multimodal speech processing, speaker diarization, speech recognition, audio-visual fusion

1. Introduction

In recent years, the proliferation of speech-enabled applications has led to increasingly complex usage scenarios, such as home environments, and professional meetings. These scenarios present considerable challenges due to adverse acoustic conditions, including far-field audio, background noise, and reverberation. Additionally, conversational interactions often involve multiple speakers with substantial speech overlap. Most publicly available meeting corpora have limited scope and lack summary annotations. For instance, the CHIL dataset[1] contains only 20 English meetings with 80 speakers and 72 hours of content. Audio-only datasets like AliMeeting[2] and Aishell-4[3] include 500 and 60 Mandarin meetings, respectively. The simulated LibriCSS dataset[4] recreates meeting dynamics by playing LibriSpeech[5] utterances through loudspeakers, but the dialogues lack the continuity of real conversations.

Current audio-only speech processing techniques are facing performance plateaus, with systems like those in CHiME-6[6] achieving a word error rate (WER) of 30%, and a character error rate (CER) of 20%, limiting their real-world deployment. The McGurk effect and subsequent studies show that visual cues, such as lip movements, enhance speech perception, particularly in noisy environments. Motivated by this, the MISP challenge aims to advance speech processing by incorporating additional modalities, namely video.

Previous MISP 2021[7], 2022[8] and 2023[9] challenges targeted the home scenario, where several people chatted in Chinese while watching TV in a living room. A large-scale audio-visual Chinese home conversational corpus was released to support tasks like wake word spotting, target speaker extraction, speaker diarization, and speech recognition. It is the first and largest distant multi-microphone Chinese audio-visual corpus. The dataset's availability led to over 150 teams downloading it, more than 60 teams submitting results, and 15 papers presented at ICASSP 2022-2024. Motivated by the success of previous MISP challenges, the MISP 2025 challenge focuses on multi-modal multi-device meeting transcription and aims to push the boundaries of current techniques by introducing additional modality information, namely the video modality. The specific tasks considered in the challenge are Audio-Visual Speaker Diarization (AVSD), Audio-Visual Speech Recognition (AVSR) and Audio-Visual Diarization and Recognition (AVDR).

Table 1: *Details of MISP-Meeting corpus*

Dataset	Train	Dev	Eval	Total
Duration (h)	119	3	3	125
Session	72	9	9	90
Room	15	4	4	23
Participant	233	15	15	263
-Male	115	7	8	130
-Female	118	8	7	133
Overlap Ratio	57.30%	46.54%	53.61%	56.95%

This paper makes the following key contributions to the field of multimodal speech processing:

- **An innovative audio-visual speech recognition challenge (MISP 2025).** We propose a new challenge focused on audio-visual speech recognition, providing a large-scale, multi-device, multi-modal dataset specifically designed for meeting scenarios. The challenge includes synchronized audio and video recordings captured under realistic conditions, addressing the need for high-quality resources in this domain.
- **A comprehensive baseline system.** We provide a complete set of baseline implementations for the tasks of audio-visual speaker diarization, audio-visual speech recognition, and joint diarization and recognition. These baselines are designed to facilitate rapid experimentation and benchmarking by researchers and practitioners.
- **In-depth analysis of participant methods.** We present a thorough summary and analysis of the techniques submitted by participants in the MISP 2025 Challenge. That includes a discussion of effective approaches, their strengths and limitations, and insights into promising directions for future.

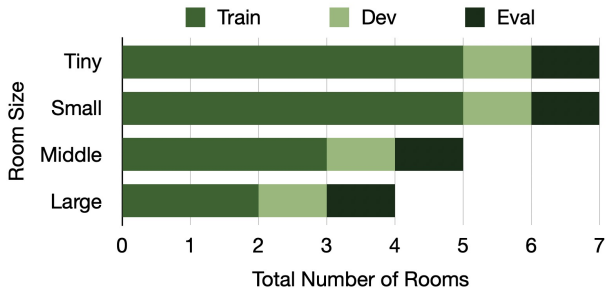

Figure 1: *Statistics of meeting rooms.*

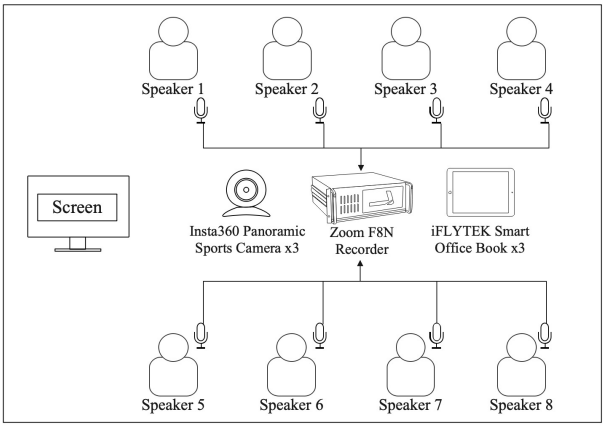

Figure 2: *Example of recording venue, and used devices.*

2. Dataset

2.1. Statics

The MISP-Meeting dataset[10] comprises a total of 125 hours of synchronized audio and video data, meticulously curated to reflect real-world meeting scenarios. The dataset is partitioned into three distinct subsets: 119 hours for training (Train), 3 hours for development (Dev), and 3 hours for evaluation (Eval), the latter serving as the basis for rigorous scoring and ranking. Specifically, the training, development, and evaluation sets consist of 72, 9, and 9 sessions, respectively. To ensure robustness and generalizability, there is no overlap in speakers or recording rooms across these subsets. Each session involves 4-8 participants engaged in natural discussions, with session durations tailored to the subset: training sessions span 2 hours, while development and evaluation sessions are condensed to 20 minutes. This design ensures that training sessions encompass multiple topic transitions, simulating the dynamic nature of real meetings. The dataset includes a total of 233 participants in the training set, and 15 participants each in the development and evaluation sets, with a balanced gender representation. Notably, all participants are professionals, or students whose fields of expertise align closely with the meeting topics, enhancing the authenticity of the discussions while minimizing prolonged silent intervals. The proportion of overlapping speech segments in the training, development, and evaluation sets is 57.30%, reflecting the natural conversational dynamics of real-world meetings.Details of the dataset, including session counts, participant demographics, and speech overlap statistics, is in Table 1.

A distinguishing feature of the MISP-Meeting corpus is the diversity of its meeting room environments. As summarized in Figure 1, the dataset encompasses 23 meeting rooms categorized into four size groups: tiny (8.79–20 m²), small (20–40 m²), medium (40–80 m²), and large (80–117.6 m²). Each subset includes rooms from all size categories, ensuring a wide range of acoustic properties and spatial configurations. The rooms are further characterized by varied wall materials (e.g., cement, glass) and furnishings (e.g., sofas, TVs, blackboards, fans, air conditioners, plants), which collectively contribute to the acoustic diversity of the dataset. Detailed acoustic parameters for each meeting venue will be released alongside the training data, providing researchers with a valuable resource for in-depth acoustic analysis and model development.

2.2. Data Collection

As shown in Figure 2, meeting attendees are seated around an 8-microphone array and a panoramic camera, both placed on the table in a standard meeting room. Participants engage in natural conversations on various topics, including medical treatment, education, business, and industrial production. During the sessions, different indoor noises, such as clicking, keyboard typing, door opening and closing, and fan sounds, naturally occur.

The microphone array is integrated into the iFLYTEK Smart Office Book X3, configured in a rectangular shape with dimensions of 197 mm by 134 mm. It consists of eight omnidirectional microphones symmetrically distributed along the width edges, each capturing audio at a 16 kHz sampling rate and 32-bit resolution. Adjacent to the array is an Insta360 Panoramic Sports Camera X3, measuring 46 mm by 114 mm, with two fisheye lenses on the front and back surfaces and two omnidirectional microphones on the sides. The setup is mounted 30-40 cm above the table surface. The outputs include 360-degree panoramic video in MP4 format, with a resolution of 3840x1920 at 30 fps, and 2-channel audio recorded at 48 kHz and 16-bit resolution.

Each participant wore a headset microphone that captured near-field speech at 44.1 kHz and 16-bit resolution, minimizing interference from off-target sources and ensuring a signal-to-noise ratio (SNR) greater than 15 dB for high-quality manual transcription. All microphones were connected to a Zoom F8N Recorder with a shared clock for synchronized recording. Despite the synchronized microphones, there were still three distinct clocks within the system: the microphone array, the camera, and the recorder. These clocks were manually synchronized by identifying a common reference point, such as knocking a cup. The visual frame capturing the moment of impact, and the corresponding sound waveform were manually aligned using the provided timestamps.

3. Challenge Description

3.1. Task 1: Audio-Visual Speaker Diarization (AVSD)

Audio-visual speaker diarization aims to solve the "who spoke when" problem by labeling speech timestamps with classes corresponding to speaker identity using multi-speaker audio and video data. Training and development sets provide all audio and video recordings along with the corresponding ground-truth segmentation timestamp. In contrast, the evaluation set does not include the near-field speech, and the transcriptions. Participants need to determine the speaker at each time point.

We follow our previous work [11] to deploy the baseline system, with adaptations to accommodate the far-field audio and video data used in this challenge. The baseline system

consists of a visual encoder, audio encoder, and decoder. The visual encoder processes lip region inputs with a lipreading model[12], Conformer blocks[12], and a bidirectional LSTM (BLSTM) layer to generate visual embeddings and an initial diarization result. The audio encoder processes the dereverberated audio, using the diarization results from the V-VAD model to compute i-vectors as speaker embeddings, while also generating audio embeddings through CNN layers. The decoder combines those embeddings and uses BLSTM with projection layers to output speech/non-speech probabilities. Furthermore, DOVER-Lap [13] is used to fuse the results of 8-channels audio. Training occurs in three stages: pre-training the visual network for V-VAD, training the audio network and decoder, and joint fine-tuning of the entire network.

Diarization error rate (DER) [14] is adopted as the evaluation criterion. The lower the DER value (with 0 being a perfect score), the higher the ranking. DER is calculated as: the summed time of three different errors of speaker confusion (SC), false alarm (FA), and missed detection (MD) divided by the total duration time, as shown in

$$\mathrm{DER} = \frac{T_{\mathrm{SC}} + T_{\mathrm{FA}} + T_{\mathrm{MD}}}{T_{\mathrm{total}}} \tag{1}$$

where T_{SC}, T_{FA} and T_{MD} are the time duration of the three errors, and T_{total} is the total time duration. It is worth noting that we do not set the "no score" collar, and overlapping speech will be evaluated.

3.2. Task 2: Audio-Visual Speech Recognition (AVSR)

The challenge of AVSR is to handle overlapped segmentation and recognize the content of multiple speakers. The provided evaluation set is the same as Task 1, and the ground truth diarization information is available. Participants are required to transcribe each speaker. As for the baseline system, we adopt that from our previous work [15]. The module uses a hybrid DNN-HMM architecture, combining audio and visual modalities for speech recognition. It starts with preprocessing, where far-field 8-channel audio is dereverberated using GSS [16], and lip ROIs are extracted from video based on diarization results. Lip movement is processed with 3D convolutions and ResNet-18 to generate visual embeddings. Log Mel-filterbank features are passed through 1D convolutions and ResNet-18 to produce audio embeddings. Audio-visual embeddings are extracted using multi-stage temporal convolutional networks. Posterior probabilities are computed with additional MS-TCN modules.

The performance is measured by character error rate (CER). The CER compares, for a given hypothesis output, the total number of characters, including spaces, to the minimum number of insertions (Ins), substitutions (Subs) and deletions (Del) of characters that are required to obtain the reference transcript. Specifically, CER is calculated by:

$$\mathrm{CER} = \frac{N_{\mathrm{Subs}} + N_{\mathrm{Del}} + N_{\mathrm{Ins}}}{N_{\mathrm{total}}} \times 100 \tag{2}$$

where N_{Subs}, N_{Del} and N_{Ins} are the character number of the three errors, respectively, and N_{total} is the total number of characters. The lower the CER value (with 0 being a perfect score), the better the recognition performance. For such speech overlap segments, we calculate all errors based on the recognition results and the ground truth for each speaker based on the oracle speaker diarization results.

3.3. Task 3: Audio-Visual Diarization and Recognition (AVDR)

Task 3 aims to directly solve the "who spoke what when" problem and can be seen as a combination of Task 1 and 2. The evaluation set is the same as Task 2, but using ground truth diarization information is prohibited. We cascade the baseline systems from Tasks 1 and 2 to form the baseline system for Task 3, following the same cascading process described in our previous work [17]. During inference, the AVSD module outputs an RTTM file with session, speaker, start time (Tstart), and duration (Tdur) information. Each session and speaker's utterances are segmented based on Tstart and Tdur. The video frames are cropped to extract lip ROIs, while the audio is dereverberated, beamformed, and segmented. AVSR generates transcriptions, which are concatenated in chronological order.

With reference to the concatenated minimum-permutation word error rate (cpWER) in [6], we use concatenated minimum-permutation character error rate (cpCER) as the evaluation criterion in Task 3. The calculation of cpCER in a session is divided into three steps:

1. Recognition results and reference transcriptions belonging to the same speaker are concatenated on the timeline in a session.
2. CERs between the reference and all possible speaker permutations of the hypothesis $\{\boldsymbol{s}_i | i = 0, 1, \cdots, \mathrm{P}_{N_{\mathrm{spk}}}^{N_{\mathrm{spk}}}\}$ are calculated as Eq. (2), where N_{spk} is the total number of speakers in the session.
3. The lowest CER as the cpCER, the process is as follows:

$$\mathrm{cpCER} = \min_{\{\boldsymbol{s}_i | i=0,1,\cdots,\mathrm{P}_{N_{\mathrm{spk}}}^{N_{\mathrm{spk}}}\}} \mathrm{CER}_i \tag{3}$$

4. Results and Analysis

4.1. AVSD

Table 2 presents a summary of the methods and results for Track 1 (Audio-Visual Speaker Diarization) in the MISP 2025 Challenge. The table lists the participating teams, their proposed methods, pre-training strategies, external datasets used, the modality employed, and the corresponding Diarization Error Rate (DER) achieved.

The top four teams in Task 1 (Audio-Visual Speaker Diarization) of the MISP 2025 Challenge showcased a range of innovative approaches. The winning team, DKU-WHU, extended the single-channel SSND model[28] to multi-channel audio with a channel attention mechanism, achieving a DER of 8.09%. The second-place team, XMUSPEECH, leveraged cross-modal integration through their CASA-Net architecture with cross-attention and self-attention mechanisms, resulting in a DER of 8.18%. NJUAALAB, in third place, combined end-to-end neural diarization with clustering techniques and advanced models like WavLM-Large[23] and Conformer[12], yielding a DER of 8.33%. FOSAFER, the fourth-place team, used a hybrid approach combining traditional methods with WavLM-based segmentation, achieving a DER of 8.88%.

The teams shared a focus on advanced neural architectures such as Conformer and WavLM, along with multi-channel audio processing. DKU-WHU and NJUAALAB emphasized multi-channel fusion, effectively utilizing channel attention and clustering techniques for complex environments. In contrast, XMUSPEECH and FOSAFER explored multimodal integration with varying success. XMUSPEECH achieved notable

Table 2: *Results of submissions of track 1 (Audio-Visual Speech Diarization) in the MISP 2025 challenge*

ID	Team	Method	Pre-train	External Datasets	Modality	DER(%)
1	DKU-WHU	MC-SSND	by VoxBlink2[18]	VoxCeleb2[19]+VoxBlink2[18]+ KeSpeech[20]+3D-Speaker[21]	Audio-only	**8.09**
2	XMUSPEECH	CASA-Net	ECAPA-TDNN[22]	MISP 2022[8]	Audio-visual	8.18
3	NJU-AALAB	DiariZen	WavLM-large[23]	AMI[24]+AISHELL-4[3]+ AliMeeting[2]	Audio-only	8.33
4	Fosafer Research	Hybrid-system	WavLM-large[23]	AISHELL-4[3]+AliMeeting[2]+ MISP2021[7]+VoxCeleb2[19]	Audio-only	8.88
5	Baseline	MISP-Baseline	LRW[25]	-	Audio-visual	15.52

Table 3: *Results of submissions of track 2 (Audio-Visual Speech Recognition) and track 3 (Audio-Visual Diarization and Recognition) in the MISP 2025 challenge*

ID	Team	Method	Pre-train	External Datasets	Modality	CER(%)	cpCER(%)
1	Fosafer Research	ASR-Aware OA System	WavLM[23]	AliMeeting[2]+KeSpeech[20]+ MISP2021[7]+VoxCeleb2[19]+ AISHELL-4[3]	Audio-only	**9.48**	**11.56**
2	DKU-WHU	MC-SSND	Whisper[26]	-	Audio-only	15.85	16.88
3	Baseline	MISP-Baseline	WenetSpeech[27]+ LRW[25]	-	Audio-visual	20.10	84.05

improvements through cross-modal attention, while FOSAFER found visual data less impactful when audio alone was sufficient. Post-processing techniques like DOVER-Lap[13] and temporal alignment were common to all teams, highlighting the importance of refining raw predictions.

The comparison suggests several promising directions for future research: first, further exploration of cross-modal attention could enhance audio-visual data integration in noisy or low-quality environments. Second, FOSAFER's dynamic hybrid approach indicates potential for adaptive systems tailored to varying acoustic conditions, such as different levels of overlapping speech. Third, NJUAALAB's use of advanced clustering and fusion techniques could improve the aggregation of diarization results, particularly in multi-channel settings. Finally, the limited impact of visual data in FOSAFER's approach suggests a need for more sophisticated video-based correction methods to complement audio in multimodal systems.

4.2. AVSR & AVDR

Table 3 provides a summary of the participant methods and results for Tracks 2 (Audio-Visual Speech Recognition) and 3 (Audio-Visual Diarization and Recognition) in the MISP 2025 Challenge. The table includes the team names, methods, pre-training strategies, external datasets used, modalities employed, and the corresponding Character Error Rate (CER) and character-phoneme CER (cpCER) values.

Both the FOSAFER and DKU-WHU teams demonstrated state-of-the-art performance in the MISP 2025 Challenge, yet their approaches exhibit distinct methodologies, and shared insights. The FOSAFER team's ASR-Aware Observation Addition (ASR-Aware OA) system focuses on addressing low SNR challenges by fusing noisy speech, Mossformer2-separated speech, and GSS-separated speech through a weighted fusion mechanism, guided by a sentence-level bridging module leveraging ASR recognition information. This approach, combined with Mossformer2[29] and Paraformer[30] models, achieved a remarkable CER of 9.48%, highlighting the effectiveness of audio-centric strategies. Similarly, the DKU-WHU team used a pretrained Whisper model as backbone, integrating WPE[31] and GSS[16] for speech enhancement and separation, achieving a CER of 15.85% and a cpCER of 16.88%. Both teams explored visual information but found it insufficient to enhance performance, underscoring the dominance of audio signals in their systems. However, the FOSAFER team's hybrid approach to speaker diarization, combining traditional methods with a WavLM-based model, contrasts with the DKU-WHU team's Multi-Channel Sequence-to-Sequence Neural Diarization (MC-SSND) framework, which refines diarization accuracy using multi-channel audio. While both systems excel in handling overlapping speech and speaker identification, the FOSAFER team's ASR-Aware OA system demonstrates superior performance in speech recognition, likely due to its innovative fusion strategy and ASR-driven optimization. Further exploration of advanced fusion techniques and domain-specific pretraining could enhance robustness in challenging acoustic environments, paving the way for more accurate and adaptable speech processing systems.

5. Conclusion

In this paper, we presented the MISP 2025 Challenge, a multimodal speech processing benchmark focused on meeting scenarios. We introduced the large-scale audio-visual meeting dataset (MISP-Meeting), provided comprehensive baseline systems for audio-visual speaker diarization, speech recognition, and joint diarization and recognition, and analyzed participant submissions. Future work will focus on improving multimodal fusion, tackling domain adaptation challenges, and expanding the dataset to cover more diverse meeting scenarios. These efforts will significantly advance multimodal speech processing.

6. References

[1] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia *et al.*, "The chil audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language resources and evaluation*, vol. 41, pp. 389–407, 2007.

[2] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.

[3] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv preprint arXiv:2104.03603*, 2021.

[4] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[6] S. Watanabe, M. Mandel, J. Barker *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. CHiME 2020*, 2020, pp. 1–7.

[7] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin *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.

[8] Z. Wang, S. Wu, H. Chen, M.-K. He, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. Siniscalchi, O. Scharenborg *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.

[9] S. Wu, C. Wang, H. Chen, Y. Dai, C. Zhang, R. Wang, H. Lan, J. Du, C.-H. Lee, J. Chen *et al.*, "The multimodal information based speech processing (misp) 2023 challenge: Audio-visual target speaker extraction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8351–8355.

[10] H. Chen, C.-H. H. Yang, J.-C. Gu, S. M. Siniscalchi, and J. Du, "MISP-Meeting: A real-world dataset with multimodal cues for long-form meeting transcription and summarization," in *Proceedings of the 63st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025, pp. 1–14.

[11] M. He, J. Du, and C.-H. Lee, "End-to-end audio-visual neural speaker diarization," in *Proc. Interspeech 2022*, 2022, pp. 1461–1465.

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[13] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 881–888.

[14] J. G. Fiscus, J. Ajot, M. Michel *et al.*, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. MLMI 2006*, 2006, pp. 309–322.

[15] Y. Dai, H. Chen, J. Du *et al.*, "Improving audio-visual speech recognition by lip-subword correlation based visual pre-training and cross-modal fusion encoder," in *Proc. ICME 2023*, 2023, pp. 2627–2632.

[16] D. Raj, D. Povey, and S. Khudanpur, "Gpu-accelerated guided source separation for meeting transcription," *arXiv preprint arXiv:2212.05271*, 2022.

[17] Z. Wang, S. Wu, H. Chen *et al.*, "The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[18] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," *arXiv preprint arXiv:2407.11510*, 2024.

[19] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[20] Z. Tang, D. Wang, Y. Xu, J. Sun, X. Lei, S. Zhao, C. Wen, X. Tan, C. Xie, S. Zhou *et al.*, "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] S. Zheng, L. Cheng, Y. Chen, H. Wang, and Q. Chen, "3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement," *arXiv preprint arXiv:2306.15354*, 2023.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.

[25] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2857–2866.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[27] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6182–6186.

[28] M. Cheng, Y. Lin, and M. Li, "Sequence-to-sequence neural diarization with automatic speaker detection and representation," *arXiv preprint arXiv:2411.13849*, 2024.

[29] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, and B. Ma, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 356–10 360.

[30] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[31] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.