The Multimodal Information Based Speech Processing (Misp) 2022 Challenge Audio-Visual Diarization And Recognition

Wang, Zhe; Wu, Shilong ; Chen, Hang ; He, Mao-Kui ; Du, Jun ; Lee, Chin-Hui ; Chen, Jingdong ; Watanabe, Shinji ; Siniscalchi, Sabato Marco ; Scharenborg, Odette

DOI
10.1109/ICASSP49357.2023.10094836

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094836

THE MULTIMODAL INFORMATION BASED SPEECH PROCESSING (MISP) 2022 CHALLENGE: AUDIO-VISUAL DIARIZATION AND RECOGNITION

Zhe Wang[1], Shilong Wu[1], Hang Chen[1], Mao-Kui He[1], Jun Du[1,], Chin-Hui Lee[2], Jingdong Chen[6], Shinji Watanabe[3], Sabato Siniscalchi[2,4], Odette Scharenborg[7], Diyuan Liu[5], Baocai Yin[5], Jia Pan[5], Jianqing Gao[5], Cong Liu[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

ABSTRACT

The Multi-modal Information based Speech Processing (MISP) challenge aims to extend the application of signal processing technology in specific scenarios by promoting the research into wake-up words, speaker diarization, speech recognition, and other technologies. The MISP2022 challenge has two tracks: 1) audio-visual speaker diarization (AVSD), aiming to solve "who spoken when" using both audio and visual data; 2) a novel audio-visual diarization and recognition (AVDR) task that focuses on addressing "who spoken what when" with audio-visual speaker diarization results. Both tracks focus on the Chinese language, and use far-field audio and video in real home-tv scenarios: 2-6 people communicating each other with TV noise in the background. This paper introduces the dataset, track settings, and baselines of the MISP2022 challenge. Our analyses of experiments and examples indicate the good performance of AVDR baseline system, and the potential difficulties in this challenge due to, e.g., the far-field video quality, the presence of TV noise in the background, and the indistinguishable speakers.

Index Terms— MISP challenge, speaker diarization, speech recognition, multimodality

1. INTRODUCTION

Modern speech-enabled systems still suffer from performance degradation in real-world scenarios (e.g., at home and in meetings) due to factors associated with adverse acoustic environments and conversational multi-speaker interactions. Inspired by the finding that visual cues can help human speech perception [1], many researchers have proposed to use the visual modality to improve acoustic robustness [2, 3]. The MISP2021 challenge [4] released a large distant multi-microphone conversational Chinese audio-visual corpus, and some advanced audio-visual speech recognition (AVSR) systems have been proposed [5, 6]. However, these systems assume that the correspondence between speech segments and speakers is known in advance, which greatly limits its scope in real-world applications. For the second MISP challenge, we target the problem of audio-visual speaker diarization (AVSD), and audio-visual diarization and recognition (AVDR) in the home-tv scenarios. Specifically, the AVDR is an extended task from AVSR, replacing oracle speaker diarization results with AVSD results.

Many approaches have been proposed on speaker diarization and speech recognition under the audio-only condition. [7] utilized x-vector [8], agglomerative hierarchical clustering (AHC) and an LSTM-based overlap detector to get diarization results, which can be used for the guided source separation (GSS) and the deep neural network-hidden markov model (DNN-HMM). [9] proposed a novel system, which includes a speaker diarization module with target-speaker voice activity detection (TS-VAD), and a speech recognition module with self-attention. However, the audio-only speaker diarization and speech recognition task in the real scenes is still a huge challenge because of the potential strong background noise and high ratios of overlapping speech [10].

Facial behavior is highly correlated with speech activity [11], and visual modality is not disturbed by harsh acoustic environment. Researchers show great interest in audio-visual speaker diarization (AVSD) and audio-visual speech recognition (AVSR). For AVSD system, some related works have been proposed. [12] utilized mutual information to fuse the audio and video modalities, while [13] used a Bayesian method for audio-visual speaker diarization. In recent years, many deep learning methods have emerged. [14] used an audio-visual synchronization model, [15] proposed a diarization method using self-supervised learning, achieving positive results. For AVSR system, [3] proposed a 'Watch, Listen, Attend and Spell' (WLAS) network on the LRS data set and [2] adopted a Transformer-based model. [16] developed a CTC/Attention model based on conformer blocks. A DNN-HMM hybrid AVSR system with a gating layer [17] also showed good performance. Although AVSD and AVSR have received increased attention and have been shown to significantly outperform conventional audio-only methods, there is as yet little research done on audio-visual diarization and recognition (AVDR) which concentrates on AVSR with AVSD results.

The MISP2022 challenge includes two tracks: audio-visual speaker diarization (AVSD), and audio-visual diarization and recognition (AVDR). In this paper, we discuss the MISP2022 challenge, the data, tracks, and provide a detailed description of the baseline AVDR system, followed by a deep analysis. Besides, we point out the difficulties that participants may encounter in this challenge, including the low quality of far-field videos, the background noise in the home-tv scenarios, and the existence of indistinguishable speakers. To the best of our knowledge, we proposed a brand-new AVDR task, and our proposed AVDR baseline system is the first to concatenate the AVSD and AVSR into one large system. The resulting system has broad application prospects. More challenge details[1] and the baseline code[2] can be found on the websites.

[*] corresponding author

[1] https://mispchallenge.github.io/mispchallenge2022
[2] https://github.com/mispchallenge/misp2022_baseline

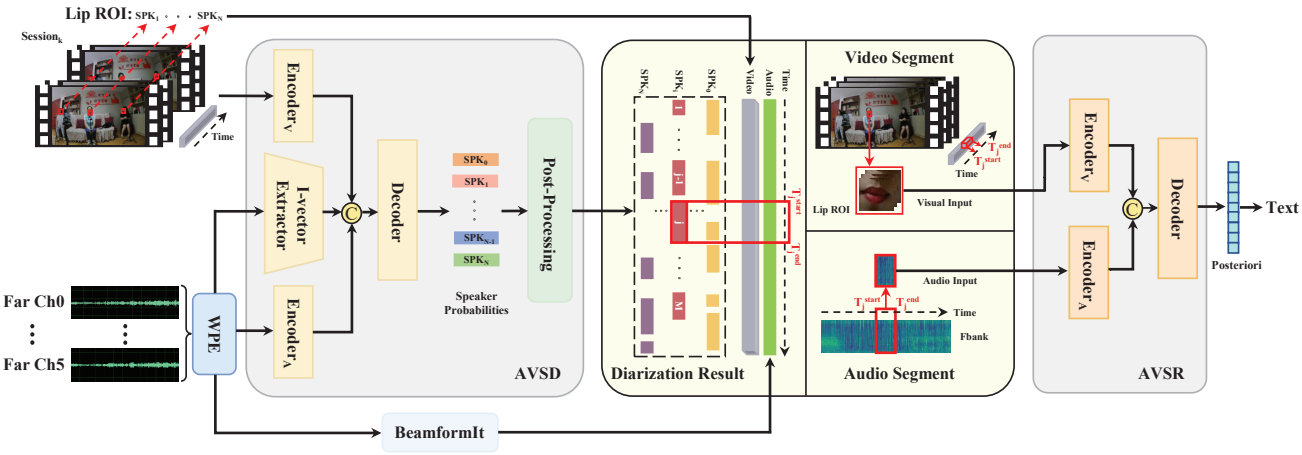

Fig. 1. The architecture of the audio-visual diarization and recognition baseline system

2. DATASET AND TRACKS

2.1. Training, Development, and Evaluation Sets

We adopted the same training set as in the updated AVSR corpus of the MISP2021 challenge [18] and picked a new 3h development set from the previous development and evaluation sets. The new development set consists of the audio and video recordings of 8 rooms and 26 participants, including 10 males and 16 females. In the future, to eliminate overlapping speakers in each subset, a new evaluation set will be released and used for final ranking. The evaluation set only contains the recordings from the far-field devices.

2.2. Track 1: Audio-Visual Speaker Diarization

Audio-visual speaker diarization aims to solve the "who spoke when" problem by labeling speech timestamps with classes that correspond to speaker identity using audio and video data. For evaluation, only the far-field audio and video data is available. We will provide the oracle speech segmentation timestamp. Participants need to submit a rich transcription time marked (RTTM) file for each session. RTTM files are text files containing one turn per line [10]. The start time (4th column), duration (5th column), and speaker ID (8th column) must remain in the same columns.

Diarization error rate (DER) [19] is adopted as the evaluation metric. The lower the DER value (with 0 being a perfect score), the higher the ranking. It is worth noting that we do not set the "no score" collar, and overlapping speech will be evaluated.

$$\mathrm{DER} = \frac{\mathrm{FA} + \mathrm{MISS} + \mathrm{SPKERR}}{\mathrm{TOTAL}} \quad (1)$$

where FA, MISS, SPKERR are the total durations of the false alarm, missed detection and speaker error, respectively, and TOTAL is the sum of durations of all reference speakers' speech.

In Track 1, external audio data can be used to train the AVSD model, such as VoxCeleb 1, 2 [20, 21], CN-Celeb [22], and other public datasets. Additional video data is also allowed to be used. However, participants should inform the organizers in advance about such data sources, so that all competitors know about them and have an equal opportunity to use them.

2.3. Track 2: Audio-Visual Diarization and Recognition

Track 2 moves beyond AVSD and also considers the task of speech recognition, i.e., transcribing the speech into its verbatim text. The same evaluation set is adopted as Track 1. Participants need to submit the RTTM file, and transcription files. In each session, participants should chronologically merge all utterances from one speaker and provide a transcription file. Transcription files contain two columns: the utterance ID (1st column), and the utterance (2nd column). The format of the utterance ID is $<$ speaker ID $>\_<$ session ID $>$.

With reference to the concatenated minimum-permutation word error rate (cpWER) in [23], we use concatenated minimum-permutation character error rate (cpCER) as the evaluation metric in Track 2. The calculation of cpCER is divided into three steps. First, recognition results and reference transcriptions belonging to the same speaker are concatenated on the timeline in a session. Second, character error rate (CER) of permutations of speakers is calculated as follows:

$$\mathrm{CER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \quad (2)$$

where S, D, I are the character number of the substitution error, deletion error, and insertion error. N is the total number of characters. Finally, select the lowest CER as the cpCER.

In Track 2, we restrict the rules of additional data usage. External audio data and video data are allowed to be used. Significantly, participants can utilize timestamps, speaker tags, and other information except for text contents. Participants should also inform the organizers in advance about such data sources.

3. BASELINE AVDR SYSTEM

Fig. 1 shows the baseline AVDR system, which consists of an AVSD module followed by an AVSR module. The AVSD module also serves as the baseline system for Track 1. In this section, we elaborate the architecture and training process of the AVSD and AVSR modules, and provide the details about joining the AVSD and AVSR modules for decoding.

3.1. Architecture and Training of the AVSD Module

We follow our previous work [24] as our baseline. The difference is that the preceding work used the data from the mid-field audio and video, while the current challenge focuses on the far-field audio and video.

As shown in the AVSD module in Fig. 1, our system has three encoder modules. In the visual encoder module, lip ROIs are used as

Table 1. **Speaker diarization results on Dev set (in %)**

System	FA	MISS	SPKERR	DER
ASD	0.01	19.88	11.36	31.25
VSD	6.64	8.17	3.89	18.69
AVSD	**4.01**	**5.86**	**3.22**	**13.09**

input of the network which consists of lipreading model [25], conformer blocks [26], and a BLSTM layer. The whole network can be regarded as a visual voice activity detection (V-VAD) model to generate visual embeddings and an initial diarization result. Next, we use the audio dereverberated by NARA-WPE [27] and the diarization result from the V-VAD model to compute i-vectors as speaker embeddings. Besides, through an FBank feature extractor and several 2D CNN layers, audio embeddings can also be extracted. In the decoder block, three types of embeddings are combined first and several BLSTM with projection (BLSTMP) layers are utilized to further extract features and get speech or non-speech probabilities for each speaker. In the post-processing stage, we first perform thresholding with the probabilities to produce a preliminary result and adopt the same approaches as in [28]. Furthermore, DOVER-Lap [29] is used to fuse the results of 6-channels audio.

The training process is as follows: first, we use the parameters of the pre-trained lipreading and train the V-VAD model. Then, we freeze the visual network parameters and train the audio network and decoder block. Finally, we unfreeze the visual network parameters, and train the whole network jointly.

3.2. Architecture and Training of AVSR Module

The AVSR model adopts a DNN-HMM hybrid system [18]. Firstly we apply the NARA-WPE [27] and BeamformIt [30] to the far-field 6-channel speech. Then, the FBank features extracted from the audio and the Lip ROIs cropped from the video were segmented on the basis of the speaker diarization results. The front-end module composed of 3D convolution and ResNet-18 is used to extract lip-movement information for the video modality and outputs $\mathrm{embedding_V}$. Meanwhile, the front-end module composed of 1D convolution and ResNet-18 is used to extract audio features and obtain $\mathrm{embedding_A}$. The audio-visual features, $\mathrm{embedding_{AV}}$, are extracted by the multi-stage temporal convolutional network (MS-TCN) [31] modules. Next, the posterior probabilities are obtained by the other MS-TCN modules. Finally, text is decoded from the posterior probabilities by using GMM-HMM, 3-gram model and DaCiDian.

During the training stage, oracle speaker diarization results are used. Kaldi [32] is applied to train a GMM-HMM system on all far-field audio data. The training of the DNN-based acoustic model uses Cross Entropy loss and Adam optimizer for 100 epochs with initial learning rate of 0.0003 and cosine scheduler. More details of the experiment can be found in [18].

3.3. Joint Decoding

During inference, the RTTM file as the output of the AVSD module contains the information of the $\mathrm{Session}, \mathrm{SPK}, \mathrm{T^{start}}$, and $\mathrm{T^{dur}}$. This information can be used for calculating DER in Track 1 and preprocessing far-field video and far-field 6-channel audio in Track 2. For $\mathrm{Session_k}$, a set of utterance identifier $(\mathrm{SPK_i}, \mathrm{T_j^{start}}, \mathrm{T_j^{dur}})$ are available, where $\mathrm{Session_k}$ and $\mathrm{SPK_i}$ denote k-th session and i-th speaker in this session, $\mathrm{T_j^{start}}$ and $\mathrm{T_j^{dur}}$ denote the start time and the duration of the j-th utterance for $\mathrm{SPK_i}$. For the far-field video in $\mathrm{Session_k}$, we first segment the whole video according to $\mathrm{T_j^{start}}$ and $\mathrm{T_j^{dur}}$ and crop the lip region of $\mathrm{SPK_i}$ in every frame as the visual input of the AVSR module. For the far-field 6-channel audio in $\mathrm{Session_k}$, we first perform WPE and BeamformIt for the raw 6-channel audio and segment the whole beamformed audio according to $\mathrm{T_j^{start}}$ and $\mathrm{T_j^{dur}}$ as audio input of the AVSR module. Finally, we concatenate the decoded text of each utterance belonging to $\mathrm{SPK_i}$ in $\mathrm{Session_k}$ according to time order.

Table 2. **Diarization and recognition results on Dev set (in %)**

System	S	D	I	cpCER
ASR(OS)	40.84	27.33	0.51	68.68
AVSR(OS)	35.78	27.82	0.36	63.96
ASD+ASR	31.83	44.34	4.27	80.44
VSD+ASR	39.25	31.22	0.66	71.13
VSD+AVSR	35.17	31.01	0.61	66.79
AVSD+AVSR	**35.94**	**29.45**	**0.68**	**66.07**

During the evaluation, due to the problem of permutation invariant training (PIT) and annotated segment text correspondence, we adopt cpCER as the final evaluation index.

4. RESULTS AND ANALYSIS

In this section, we first introduce the experimental results of the baseline systems. Next, we point out the difficulties of the MISP2022 challenge by providing examples, and analyze the good performance of AVDR system. Challenge participants can use this information to particularly focus on solving these issues in order to improve performance above the baseline.

4.1. Baseline Results

Table 1 shows the false alarm (FA) rate, missed detection (MISS) rate, speaker error (SPKERR) rate, and the DER for the audio-only speaker diarization systems (ASD), the visual-only speaker diarization system (VSD) and the audio-visual speaker diarization system (AVSD), where the latter is the baseline system of the AVSD track. For the ASD system, we use the VBx method [33]. For the VSD system, we use the result from the visual encoder module as described in Section 3.1. The ASD system has poor results, most likely due to the loud TV background noises and high speaker overlap ratios, resulting in high MISS and SPKERR rates. Because the visual modality is not disturbed by the acoustic environment, the VSD system outperforms the ASD system in terms of MISS, SPKERR, and DER. However, VSD system has a high FA rate, potentially due to the lip movement in the silent segments. Combining the audio and visual modalities in the AVSD system yields the best performance, showing that both modalities can be combined to overcome their individual weaknesses.

As shown in Table 2, we design 6 experiments for diarization and recognition system. The first two experiments are the speech recognition modules with the oracle speaker (OS) diarization results. The other experiments are the combinations of speaker diarization module and speech recognition module, e.g., ASD+ASR, VSD+ASR, VSD+AVSR, and AVSD+AVSR, where the latter is the baseline system of the AVDR track. For the ASD+ASR system, the high MISS and SPKERR rate results in a large number of deletion errors of target speakers. In addition, the high SPKERR rate leads to insertion errors of interfering speakers. Comparing the ASD+ASR system and the VSD+ASR system indicates that visual modality of speaker diarization module dominates the performance of the whole diarization and recognition system. In contrast to the VSD+ASR

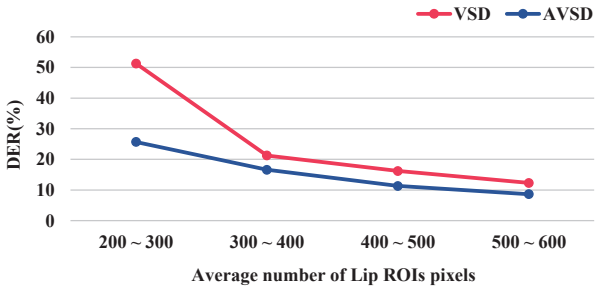

Fig. 2. The DER comparison between the VSD and AVSD systems for different pixel values of Lip ROIS in the conversations

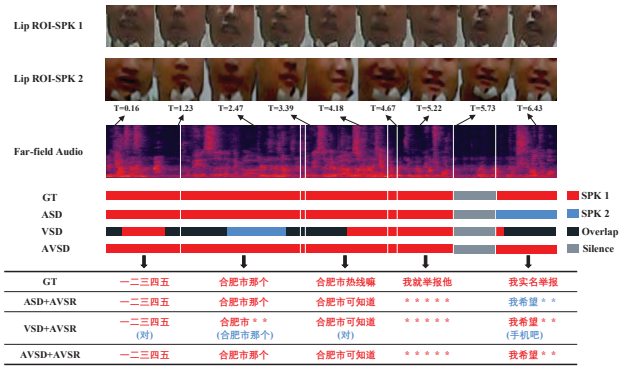

Fig. 3. An example in a session with the comparison of results from different systems

system, the visual modality in speech recognition module of the VSD+AVSR system provides distinguishable information that reduces substitution errors, which improves the whole system performance. In all experiments, it is the combination of the audio and visual modalities in both modules that yields the best system: AVDR.

4.2. Analyses of difficulties

In order to let challenge participants solve problems better, we point out the potential difficulties in this challenge. Meanwhile, we discuss the performance of different module combinations to further explore the impact of audio and visual modalities.

4.2.1. Far-field Video Quality

Due to the long distance between cameras and speakers, far-field video will result in a greatly reduced proportion of each speaker's lip ROI in the total image, especially in the scenes with more speakers. At the same time, lamplight, position, angle, occlusion, and other environmental factors may lead to the reduction of video quality. We explore how the number of lip ROIs pixels affects the performance of the VSD and AVSD systems, as shown in Fig. 2. It is found that as the average number of pixels decreases, the DER rises sharply. This will also affect the subsequent speech recognition task.

According to the example in Fig. 3, it can be seen that dim-light and far distance lead to low quality of the far-field lip ROIs, making lip movements detection wrong or missing. There are lots of overlapping segment false detections and speaker confusion in the VSD results. In fact, according to the ground truth (GT), only one speaker (SPK 1) is talking all the time. For the module of AVSR using VSD results, the existence of overlapping segments leads to more insertion errors for SPK 2, and the speaker confusion leads to deletion errors

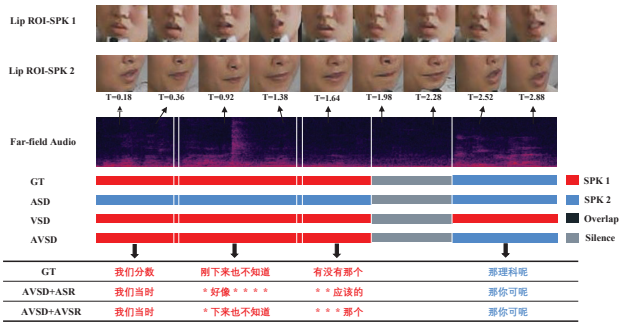

Fig. 4. Another example in a session with the comparison of results from different systems

for SPK 1. Because AVSD incorporates audio modality information to modify video modality, the results are significantly improved compared with VSD, and AVDR system results are also improved.

4.2.2. TV Background Noise

Since the TV is closer to the far-field microphone array, loud TV noise may cover the voice of the target speakers in the far-field audio. At the same time, due to the diversity of TV broadcast content, the audio may contain the voice of irrelevant speakers, which may interfere with the speaker diarization, and speech recognition. As shown in Fig. 3, in the fourth segment utterance, because actors on TV are talking loudly, noise received by the microphone completely covers the voice of the target speaker, making the AVDR system unable to recognize the speech content. Besides, in the last segment, due to the influence of TV background noise, ASD system wrongly assigns the segment of SPK1 to SPK2. Although the effect of AVDR system is better than that of single mode system, the TV background noise is also a big challenge in MISP2022.

4.2.3. Indistinguishable Speakers

Due to the diversity of speakers, it is possible that speakers with similar timbre appear in the same session. As shown in Fig. 4, the similar timbre leads to speaker confusion in ASD result. In addition, peristalsis of lips, namely lip-movement without utterance, occasionally occurs in video recordings. It is difficult for the model to distinguish whether a speaker is talking or just moving his lip. In the VSD process, due to peristalsis, speaker confusion also arises which causes the target speaker to have more deletion errors and the interfering speaker to have more insertion errors. However, in the AVSD process, through the information complementation between audio-visual modalities, we get the diarization result consistent with the ground truth, which corrects the speech recognition errors caused by the wrong diarization result.

5. CONCLUSIONS

This paper describes the MISP2022 challenge, which is the first to propose the audio-visual diarization and recognition (AVDR) task. We provide the analysis of this challenge, including the baseline results, the relationship between the diarization and the speech recognition modules, and the difficulties of the challenge. We believe that the research on audio-visual diarization and recognition can be better promoted through the MISP dataset and the MISP2022 challenge.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] Lawrence D Rosenblum, "Speech perception as a multimodal phenomenon," *Current directions in psychological science*, vol. 17, no. 6, pp. 405–409, 2008.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, et al., "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] Joon Son Chung, Andrew Senior, Oriol Vinyals, et al., "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[4] 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*. IEEE, 2022, pp. 9266–9270.

[5] Gaopeng Xu, Song Yang, Wei Li, et al., "Channel-wise av-fusion attention for multi-channel audio-visual speech recognition," in *Proc. ICASSP*. IEEE, 2022, pp. 9251–9255.

[6] Wei Wang, Xun Gong, Yifei Wu, et al., "The sjtu system for multimodal information based speech processing challenge 2021," in *Proc. ICASSP*. IEEE, 2022, pp. 9261–9265.

[7] Ashish Arora, Desh Raj, Aswin Shanmugam Subramanian, et al., "The JHU Multi-Microphone Multi-Speaker ASR System for the CHiME-6 Challenge," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments*, 2020, pp. 48–54.

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[9] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, et al., "The STC System for the CHiME-6 Challenge," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments*, 2020, pp. 36–41.

[10] Neville Ryant, Kenneth Church, Christopher Cieri, et al., "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[11] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[12] Athanasios Noulas, Gwenn Englebienne, and Ben JA Krose, "Multimodal speaker diarization," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 79–93, 2011.

[13] Israel D Gebru, Sileye Ba, Xiaofei Li, et al., "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[14] Rehan Ahmad, Syed Zubair, Hani Alquhayz, et al., "Multimodal speaker diarization using a pre-trained audio-visual synchronization model," *Sensors*, vol. 19, no. 23, pp. 5163, 2019.

[15] Yifan Ding, Yong Xu, Shi-Xiong Zhang, et al., "Self-supervised learning for audio-visual speaker diarization," in *Proc. ICASSP*. IEEE, 2020, pp. 4367–4371.

[16] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP*. IEEE, 2021, pp. 7613–7617.

[17] Fei Tao and Carlos Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, 2018.

[18] Hang Chen, Jun Du, Yusheng Dai, et al., "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," in *Proc. Interspeech*, 2022.

[19] Jonathan G Fiscus, Jerome Ajot, Martial Michel, et al., "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. MLMI*. Springer, 2006, pp. 309–322.

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[22] Yue Fan, JW Kang, LT Li, et al., "Cn-celeb: a challenging chinese speaker recognition dataset," in *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[23] Shinji Watanabe, Michael Mandel, Jon Barker, et al., "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments*, 2020, pp. 1–7.

[24] Maokui He, Jun Du, and Chin-Hui Lee, "End-to-end audio-visual neural speaker diarization," in *Proc. Interspeech*, 2022, pp. 1461–1465.

[25] Brais Martinez, Pingchuan Ma, Stavros Petridis, et al., "Lipreading using temporal convolutional networks," in *Proc. ICASSP*. IEEE, 2020, pp. 6319–6323.

[26] Anmol Gulati, James Qin, Chung-Cheng Chiu, et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[27] Lukas Drude, Jahn Heymann, Christoph Boeddeker, et al., "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[28] Maokui He, Xiang Lv, Weilin Zhou, et al., "The ustc-ximalaya system for the icassp 2022 multi-channel multi-party meeting transcription (m2met) challenge," in *Proc. ICASSP*. IEEE, 2022, pp. 9166–9170.

[29] Desh Raj, Leibny Paola Garcia-Perera, Zili Huang, et al., "Dover-lap: A method for combining overlap-aware diarization outputs," in *Proc. SLT*. IEEE, 2021, pp. 881–888.

[30] 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.

[31] Yazan Abu Farha and Jurgen Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.

[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al., "The kaldi speech recognition toolkit," in *Proc. ASRU*. IEEE, 2011.

[33] Federico Landini, Ján Profant, Mireia Diez, et al., "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.