THE XMU SYSTEM FOR AUDIO-VISUAL DIARIZATION AND RECOGNITION IN MISP CHALLENGE 2022

Tao Li[1]*, Haodong Zhou*[2]*, Jie Wang*[2]*, Qingyang Hong*[1]*, Lin Li*[2]

[1]School of Informatics, Xiamen University, Xiamen 361005, China;

[2]School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China;

{qyhong,lilin}@xmu.edu.cn

ABSTRACT

In this paper, we present our efforts in track 2 of the Multimodal Information based Speech Processing (MISP) 2022 Challenge. We built a cascaded system and explored different acoustic front-ends and end-to-end speech recognition back-ends based on multimodal. To promote effective fusion between the different modalities, we introduced a multi-level feature fusion network. By utilizing several additional strategies, our system achieved 31.88% in the concatenated minimum permutation character error rate (cpCER) on the evaluation set, earning us the 3th place ranking in the competition.

***Index Terms*—** Multimodal, Multi-talker, Far Field, Speech Recognition

1. INTRODUCTION

With the advancement of multimodal technology, there is an increasing focus on audio-visual speech recognition. The MISP Challenge seized this trend and released the first multi-microphone conversational Chinese audio-visual corpus and a large vocabulary continuous Chinese lip-reading dataset in home-TV scenarios[1]. Specifically, the MISP Challenge has two tracks, among which track 1 is Audio-Visual Speaker Diarization (AVSD) and track 2 is Audio-Visual Diarization and Recognition (AVDR). Our approach for track 2 primarily involves data augmentation, exploration of front-end and back-end processing, experimentation with multimodal fusion methods, and system fusion. To further improve the modal fusion, we tried to fine-tune the visual feature extractor and proposed a multilevel feature fusion network, furthermore, we also apply some optimization tricks such as intermediate CTC and test-time augmentation, which can yield positive results, our final system achieved cpCER of 31.88% on the evaluation set. The following sections of our paper will be organized as follows: Section II provides an overview of our system, Section III presents our experimental results and analysis, and Section IV concludes the paper.

2. SYSTEM OVERVIEW

Our system is designed as a cascaded system with several subsystems, including an official baseline of track 1 to produce speaker diarization, an offline speech front-end that separates the target speaker's speech signal using speaker diarization, visual and audio feature extractors that extract features from each modality, a feature fusion module, and a speech recognition module as the back-end.

2.1. Data Augmentation

The audio training data is primarily derived from the far-field and middle-field data after undergoing GSS processing. We employed various techniques to augment audio data, including noise interference, speed, pitch and volume perturbation, and SpecAugment. To further improve the diversity of training data, we simulated multi-channel data using the official room information provided. Additionally, we also utilized WPE and BeamformIT to enhance the raw far-field data and middle-field data. We ended up with nearly 2000 hours of training data. For visual data, we only performed simple augmentations such as random cropping, rotation, and brightness interference.

2.2. Acoustic Front-end and Back-end

The front-end includes two parts: WPE and guided source separation (GSS). WPE uses linear prediction to estimate and remove the reverberation component in the signal, while GSS leverages complex angular central Gaussian distributions, MVDR beamforming, and time annotation to separate the speech signal of the target speaker. Essentially, GSS introduces speaker diarization into the blind source separation process, thus addressing the issue of speaker arrangement after separation. We adopted a Conformer encoder with 16 blocks for the back-end. To further improve performance, we introduced the intermediate CTC[2] which acts as a regularization mechanism. The well-trained Conformer encoder was used to initialize the encoder in our multimodal model.

2.3. Multimodal Fusion

We tried three different modal fusion methods, which are concatenation, cross-attention, and the proposed multi-level feature fusion network, as shown in Fig. 1, the multi-level feature fusion network takes the audio and video features as input and then applies the Conformer encoder with 6 blocks to ex-

Table 1. Performance of our system on development and evaluation sets.

System	Back-end	Front-end	Fusion Method	Fine-tune	Dev(cpCER%)	Eval(cpCER%)
baseline	MS-TCN	WPE+BF	Concat	-	66.07	63.03
M1	Conformer-A	GSS	-	-	42.1	-
M2	Conformer-A*	GSS	-	-	32.7	37.22
M3	Conformer-A*+inter CTC	GSS	-	-	31.5	35.98
M4	Conformer-AV*+inter CTC	GSS	Cross Attention	No	33.4	-
M5	Conformer-AV*+inter CTC	GSS	Concat	No	32.5	36.79
M6	Conformer-AV*+inter CTC	GSS	Concat	Yes	31.0	35.53
M7	Conformer-AV*+inter CTC	GSS	Multi-level Feature Fusion	Yes	29.4	33.94
ROVER	-	-	-	-	**27.5**	**31.88**

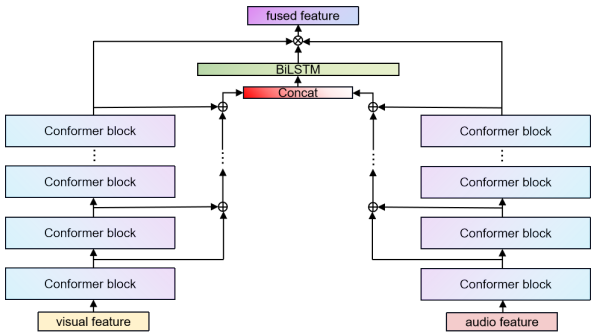

Fig. 1. Illustration of proposed multi-level fusion network.

tract features at different levels for both modalities. The features of different levels are then concatenated and fed into a 3-layer BiLSTM network with 512 cells to capture the correlations between the audio and video modalities. The correlative scores between the modalities are then obtained through a linear layer and a sigmoid function, producing the final fused features via dot product. The idea behind this approach is to utilize both high-level semantic information and low-level detailed information from both audio and video modalities to achieve better fusion performance.

2.4. Visual Feature Extractor Fine-tune

We utilized a lip-reading model pre-trained on LRW1000[3] as the visual feature extractor, but encountered some issues during the experiment. The model was initially trained for isolated word recognition, which is a multi-classification task, and was inconsistent with the training objective of the end-to-end continuous speech recognition model. To address this issue, we replaced the cross-entropy loss with CTC loss and fine-tuned it using far-field video data from MISP.

3. RESULTS AND ANALYSIS

Table. 1 displays the results of our system for both the development set and the evaluation set, where "A" stands for audio, "V" stands for video, and the * indicates data augmentation. By comparing the performance of the single-modality systems M1 and M2, we discovered that data augmentation significantly enhanced the model. Subsequently, intermediate CTC further improved the model's performance. We then evaluated various multimodal fusion strategies, and based on a comparison of M5 and M4 in the development set, we found that concatenation outperformed cross-attention. Furthermore, a comparison between M5 and M3 revealed that the performance of the multimodal model was even worse than that of the single-modality model, before fine-tuning the visual feature extractor. After fine-tuning, the model's performance improved slightly. M7 demonstrates the advantage of our proposed multimodal fusion approach. Finally, we utilized ROVER tools and test time augmentation to integrate the decoding results of all single and multimodal conditions, resulting in a cpCER of 31.88% on the evaluation set.

4. CONCLUSION

In this paper, we explored acoustic front-end, multimodal fusion methods, etc. Our best outcome involved a 32.06% of cpCER decrease against the baseline on the evaluation set, ranking third out of ten teams.

5. REFERENCES

[1] Hang Chen, Jun Du, Yusheng Dai, Chin Hui Lee, Sabato Marco Siniscalchi, and Shinji Watanabe, "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022, vol. 2022, pp. 1766–1770.

[2] Jaesong Lee and Shinji Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6224–6228.

[3] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.