THE NPU-ASLP SYSTEM FOR AUDIO-VISUAL SPEECH RECOGNITION IN MISP CHALLENGE 2022

Pengcheng Guo[†], He Wang[†], Bingshen Mu, Ao Zhang, Peikun Chen

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xian, China

ABSTRACT

This paper describes our NPU-ASLP system for the Audio-Visual Diarization and Recognition (AVDR) task in the Multimodal Information Based Speech Processing (MISP) Challenge 2022. Specifically, the weighted prediction error (WPE) and guided source separation (GSS) techniques are used to reduce reverberation and generate clean signals for each single speaker first. Then, we explore the effectiveness of Branchformer and E-Branchformer based ASR systems. To better make use of the visual modality, a cross-attention based multi-modal fusion module is proposed, which explicitly learns the contextual relationship between different modalities. Experiments show that our system achieves a concatenated minimum-permutation character error rate (cpCER) of 28.13% and 31.21% on the Dev and Eval set, and obtains a second place in the challenge.

Index Terms— Multimodal, Audio-Visual Speech Recognition

1. INTRODUCTION

With the advances of deep learning, lots of progress has been achieved for automatic speech recognition (ASR) and its performance has been improved significantly [1, 2]. However, ASR systems are still susceptible to performance degradation in real-world far-filed scenarios like meetings or home parties, due to the background noise, inevitable reverberation, and multiple speakers overlapping [3]. To achieve a robust ASR system in such challenging acoustic environments, plenty of studies focus on combining a separate speech enhancement module with the ASR model or an end-to-end optimization of all components. In addition, audio-visual based ASR (AV-ASR) training has also drawn immense interest when both auditory and visual data are available, since the additional visual cues, such as facial/lip movements, could provide complementary information and increase the model's robustness, especially in noisy conditions.

Inspired by this, the first Multimodal Information Based Speech Processing (MISP) Challenge [4, 5] was launched, which targeted exploring the usage of both audio and video data in distant multi-microphone conversational wakeup and recognition tasks. Different from the first MISP Challenge that provides oracle speaker diarization results, this year, the MISP Challenge 2022 removes such prior knowledge and extends previous tasks to more generic scenarios, which are audio-visual speaker diarization (AVSD), and audio-visual diarization and recognition (AVDR).

This study describes our system for the AVDR task (Task2) of the MISP Challenge 2022. To develop a robust AV-ASR system, we first explore several commonly used data processing techniques, including the weighted prediction error (WPE) [6] based dereverberation, guided source separation (GSS) [7] and data simulation.
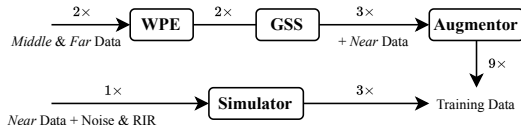
---

† Equal contribution.

Fig. 1: The flow chart of data processing and simulation. N× refers to N times the original 106 hours *Near* data provided by the challenge.

Then, advanced end-to-end architectures like Branchformer [8] and E-Branchformer [9] are used to build the basic ASR systems with the joint connectionist temporal classification (CTC)/attention training. To better make use of the visual modality, we propose a cross-attention based multi-modal fusion module, which explicitly learns the contextual relationship between different modalities. After combining the results from various systems by the Recognizer Output Voting Error Reduction (ROVER) technique, we achieve a final concatenated minimum permutation character error rate (cpCER) of 28.13% and 31.21% on the Dev and Eval set, obtaining a second place in the challenge competition.

2. PROPOSED SYSTEM

2.1. Data Processing and Simulation

Fig. 1 shows our data processing and simulation progress. Both *Middle* and *Far* data are first pre-processed by the WPE [6] and GSS [7] algorithms to obtain the enhanced clean signals of each speaker. Then, an Augmentor module conducts speech perturbation on the combination of enhanced data and original *Near* data, resulting in about 9-fold training data[1]. For the simulation part, the MUSAN corpus [10] and the open-source pyroomacoustics toolkit[2] are applied to generate background noises and room impulse responses (RIRs). The total training data is about 1300 hours.

2.2. Audio based Speech Recognition

For the audio based ASR systems, we investigate the effectiveness of the recently proposed Branchformer [8] and E-Branchformer [9] architectures. The Branchformer [8] encoder adopts two parallel branches to capture various ranged contexts. While one branch employs self-attention to learn long-range dependencies, the other branch utilizes a multi-layer perceptron module with convolutional gating (cgMLP) to extract fine-grained local correlations synchronously. In [9], Kim *et al*. enhanced Branchformer by applying a depth-wise convolution based merging module and stacking an additional pointwise feed-forward module, named E-Branchformer.

2.3. Audio-Visual based Speech Recognition

Fig. 2 shows the overall framework of our AV-ASR model. In detail, each modality is first processed by a frontend module to extract features. The visual frontend is a 5-layer ResNet3D module, while the

---

[1] N-fold data refer to N times the original 106 hours *Near* data.
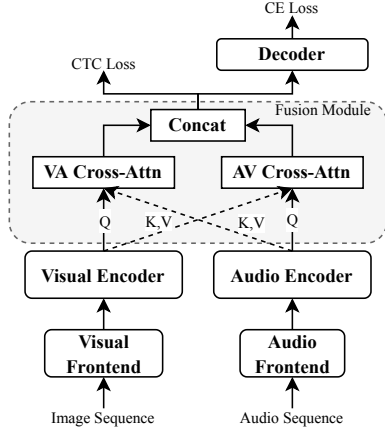[2] https://github.com/LCAV/pyroomacoustics

Fig. 2: An overview of the proposed AV-ASR model.

audio frontend is a 2-layer convolutional subsampling module. Following the frontends, two modal-dependent Branchformer encoders are used to encode input features as latent representations. The proposed fusion module consists of 2 cross-attention layers, each of which takes one modality as the Query vector and the other modality as Key/Value vectors. With the help of cross-attention layers, each modality could learn the related and complementary context from the other modality. Finally, representations from different modalities are concatenated together to compute CTC loss and fed into the Transformer decoder to compute cross-entropy (CE) loss.

2.4. Inference Procedure

During the inference, the Eval set is first segmented by a speaker diarization (SD) model, enhanced by WPE and GSS, transcribed by our ASR or AV-ASR models, and rescored by a Transformer based language model (LM). Our SD model is implemented based on the released baseline system[3] by replacing the long short-term memory (LSTM) based diarization module with Transformer and the speaker embedding module with an ECAPA model [11] pre-trained on all available speaker recognition and verification data in the challenge homepage[4]. The diarization error rates (DERs) of baseline SD and our SD are 13.09% and 9.43% on the Dev set, respectively. Finally, results from different systems are fused by the ROVER technique.

3. EXPERIMENTS

3.1. Setup

All of the models are implemented with ESPnet [12]. For the audio based ASR systems, we follow the ESPnet recipe to set the framework: Enc = 24, Dec = 6. Since the additional modules increase the parameters of E-Branchformer, we also train an E-Branchformer Small (Enc = 16) for a fair comparison. For AV-ASR systems, the visual front is a 5-layer ResNet3D module, whose channels are 32, 64, 64, 128, 256 and kernel size is 3, the visual encoder is a 12-layer Branchformer, and others are the same as ASR systems. During the training, the audio branch is initialized by well-trained ASR models.

3.2. Results

Table 1 presents the cpCER results of various ASR systems. It can be seen that all of our models gives better results over the official baseline and achieve up to 40% absolute cpCER improvement due to the enhanced dataset and advanced model architectures. Comparing M1 and M2, the data simulation gives a noticeable gain. Besides, E-Branchfomer Small (M3) obtains similar results with Branchformer

---

[3]https://github.com/mispchallenge/misp2022_baseline
[4]https://mispchallenge.github.io/mispchallenge2022/extral_data.html

Table 1: The cpCER (%) results of ASR systems on the Dev set. The Dev set is segmented by oracle timestamps, baseline SD model, and our SD model.

Sys.	Model	Oracle Timestamps	Base SD (DER=13.09%)	Our SD (DER=9.43%)
B1	Official Baseline	66.07	N/A	N/A
M1	Branchformer	26.60	34.04	30.67
M2	- remove Simu Data	27.90	35.00	31.72
M3	E-Branchformer Small	26.80	33.83	30.61
M4	E-Branchformer	**26.50**	**33.73**	**30.49**

Table 2: The cpCER (%) results of AV-ASR systems on the Dev set.

Sys.	Model	Oracle Timestamps	Our SD (DER=9.43%)
M5	AV-Branchformer (init by M2)	26.90	30.85
M6	AV-Branchformer (init by M1)	**25.70**	**29.73**

(M1) when model sizes are the same, and increasing the encoder block of E-Branchformer (M4) gives the best results. When comparing the results in each row, we find that a better SD model could bring consistent performance improvement. Table 2 shows the cpCER results of our AV-ASR systems. For AV-ASR, a good initialization of the ASR branch gives better performance (M5 vs. M6). Comparing M6 and M1, the incorporation of visual modality gives about 0.9% cpCER improvement. After fusing systems of M1, M3, M4, M5 and M6, we obtain a cpCER of 28.13% and 31.21% on the Dev and Eval set (w/ our SD), achieving second place in the challenge.

4. CONCLUSION

In this study, we describe our system for the Task2 of the MISP 2022 Challenge. Our effort includes data processing and simulation strategies, investigation of advanced architectures, and a novel cross-attention based multi-modal fusion model. By combining various systems, we get a cpCER of 31.21% on the final Eval set, obtaining a second place in the challenge.

5. REFERENCES

[1] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, et al., "A comparative study on Transformer vs RNN in speech applications," in *Proc. ASRU*. IEEE, 2019, pp. 449–456.

[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, et al., "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. Interspeech*. ISCA, 2020, pp. 5036–5040.

[3] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME-6 Workshop*, 2020, pp. 1–7.

[4] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, 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] Hang Chen, Jun Du, Yusheng Dai, Chin-Hui Lee, Sabato Marco Siniscalchi, et al., "Audio-visual speech recognition in MISP2021 challenge: Dataset release and deep analysis," in *Proc. Interspeech*. ISCA, 2022, pp. 1766–1770.

[6] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE/ACM TASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[7] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, et al., "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME-5 Workshop*, 2018.

[8] Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*. PMLR, 2022, pp. 17627–17643.

[9] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, et al., "E-Branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*. IEEE, 2023, pp. 84–91.

[10] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[11] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*. ISCA, 2020, pp. 3830–3834.

[12] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*. ISCA, 2018, pp. 2207–2211.