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

Gaopeng Xu, Xianliang Wang, Sang Wang, Junfeng Yuan, Wei Guo, Wei Li, Jie Gao

NIO Co., Ltd.

ABSTRACT

This paper describes NIO system for audio-visual diarization and recognition in the Multimodal Information Based Speech Processing (MISP) Challenge 2022. In our system, we proposed combining end-to-end audio-visual neural speaker diarization model and Channel-wise Av-fusion encoder with speaker signature for multi-channel audio-visual speech diarization and recognition. Our system reduces the concatenated minimum permutation character error rate(cpCER) by 34.36% absolute compared to the baseline in track 2.

Index Terms— Multimodal, Channel-wise, Guided Source Separation, Audio-Visual Speech Recognition

1. INTRODUCTION

The Multi-modal Information Based Speech Processing (MISP) Challenge 2022 considers considers the problem of audio-visual distant multi-microphone signal processing in everyday home environments. The challenge consists of two tracks, Audio-Visual Speaker Diarization and Audio-Visual Diarization and Recognition. The training set is the same as the training set in the MISP2021 Challenge[1]. The whole audio set contains near-field data, middle-field microphone array data, far-field microphone array data. Video data are distributed in the form of MP4 files at a frame rate of 25 fps. Each class was recorded by a far-field wide-angle camera and a mid-field high-definition camera worn by each participant.

In this paper, we detail our proposed system for MISP2022 track2. Unlike MISP2021, the information about the speakers boundaries for utterances was not provided in Task 2 of MISP2022. Therefore, achieving accurate detection of speaker boundaries is one of the goals of track2. The combination of speech activity detection(VAD) and speaker recognition is a common method for speaker boundary detection. However, this method cannot deal with speaker overlap, so the detection is not accurate. Front-end and ASR performance will be affected if the speaker diarization is not accurate enough.

Figure 1 shows the framework of our systems. Our diarization system is consistent with the method proposed by the MISP 2022 organizer.it is an end-to-end audio-visual neural network that directly predicts speech probabilities for all speakers simultaneously and it has the abilities to handle overlap segments and distinguish between speech and nonspeech. For front-end, We used weighted prediction error (WPE) based dereverberation for multi-channel signals and multi-channel speech enhancement with guided source separation (GSS). For data augmentation(AUG), we applied 7 types of data augmentation, including speed perturbation, volume perturbation, reverberation simulation, babble noise augmentation, music noise augmentation, non-speech noise augmentation and pitch augmentation. For audio-visual speech recognition back-end, similar to our previous work, Channel-wise Av-fusion attention to learn the contextual relationship across channels and modals. Finally, we inject speaker information into the encoder to enable the model to learn speaker related information.

The rest of this paper is organized as follows. Section 2 contains a brief description of our system's front-end and audio-visual Diarization. Section 3 introduces our proposed model for the MISP 2022 track 2. Experimental results are reported in Section 4, while conclusions are given in Section 5.

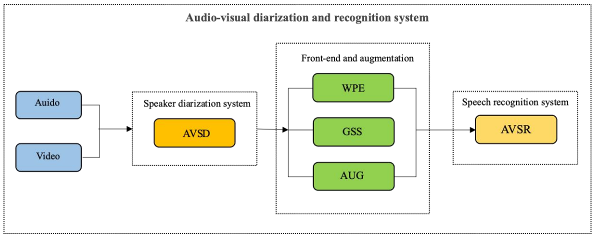
Audio-visual diarization and recognition system
Auido
Video
Speaker diarization system
AVSD
Front-end and augmentation
WPE
GSS
AUG
Speech recognition system
AVSR

Fig. 1. An illustration of overall framework.

2. FRONT-END AND DIARIZATION

2.1 Dereverberation and Denoising

We used the NARAWPE tool implementation of weighted prediction error (WPE) and and multi-channel speech enhancement with guided source separation (GSS) in our front-end module.

2.2 Audio-Visual Diarization

Our end-to-end audio-visual neural speaker diarization system is based on the offical baseline[2]. The differences is that we use a LARGE conformer encoder, which consists of 8-layer conformer (nhead = 8, dmodel = 512, dffn = 2048) and it was optimized using adamw and the learning rate is warmed up linearly in the first 10% updates.We present the audio-visual speaker diarization results on training sets. First, compared with the Baseline AVSD system DER of 13.09% in DEV set, we implemented two versions yield comparable DER, as listed in Tabel1. We see that using the single-channel based model M2 improved the baseline system's performance by about 0.85% absolute. Fusion of probabilities from 6-channels audio further improves speaker diarization.

Model	Architecture	DER
M1	baseline	13.09%
M2	our single channel	12.25%
M3	our 6-channels	11.68%

Table 1: The DER (%) results of AVSD systems.

3. AUDIO-VISUAL SPEECH RECOGNITION

3.1. Encoder

Similar to our previous work [3], we use channel-wise encoder to get channel-wise outputs per channel and Av-fusion attention to get audio-visual fusion output. We use the channel-wise output as Query, and the Av-fusion encoder outputs as Key, Value to conduct cross-attention operation. In this work, we use Channel-wise Av-fusion conformer (CFC) instead of Channel-wise transformer model (CFT). Meanwhile, conformer model equipped with a speaker signature module for speaker adaptation. The speaking styles of different speakers are obviously different. Therefore, we represent speaker information as i-vectors which will be served as an speaker signature information fed into the acoustic model. Speaker-specific input feature will be transformed by an affine layer before added to the CFC module. In this method, the AVSR model can utilize additional Speaker-specific information at both training and inference stage.

3.2 Decoder

During the inference stage, the CTC decoder generates n-best hypothesis. After CTC beam search decoding, a WFST-baseddecoder combines a 4-gram word-level LM to generate the N-best hypothesis. Then, hypotheses are rescored by the attention-decoder in the 2-pass stage with 0.5 rescore weight.

4. RESULTS

We perform our experiments on the MISP2022 audio-visual dataset, which contains 110+ hours of audio-visual data. All encoders contain 12 blocks, each with 512-dim, 8 attention heads, and 2048-dim feed-forward inner-layer. The Decoder includes 6 blocks with 8 heads, and the dimension of attention and the feed-forward layer was set to 512 and 2048. The CTC loss and attention loss are combined in the training stage, and CTC loss weight is set to 0.5. The model was optimized with Adam, and the learning rate was warmed-up for 25000 steps.

Model	Architecture	cpCER
M4	baseline	63.94%
M5	CFT	31.15%
M6	CFC	30.03%
M7	CFC+Speaker signature	29.58%

Table 2. The cpCER (%) results of AVSR systems.

Table 2 shows the result submitted for the MISP challenge 2022 track 2. The cpCER on the eval set of the baseline model was 63.94%. M5 shows the improvement gained using the CFT, achieving a cpCER of 31.15%. In M6, we use CFC, achieving a cpCER of 30.03%. In M7, we used CFC with speaker signature and the cpCER dropped from 30.03% to 29.58%.

5. CONCLUSION

We proposed combining end-to-end audio-visual neural speaker diarization model and Channel-wise Av-fusion encoder with speaker signature for multi-channel audio-visual speech diarization and recognition. Our best result achieved a cpCER of 29.58% for the evaluation set, with an absolute reduction of 34.36% compared to the baseline model.

6. REFERENCES

[1] Chen H, Zhou H, Du J, et al. The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 9266-9270.

[2] He M K, Du J, Lee C H. End-to-End Audio-Visual Neural Speaker Diarization[J]. Proc. Interspeech 2022, 2022: 1461-1465.

[3] Xu G, Yang S, Li W, et al. Channel-Wise AV-Fusion Attention for Multi-Channel Audio-Visual Speech Recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 9251-9255.