THE FLYSPEECH AUDIO-VISUAL SPEAKER DIARIZATION SYSTEM FOR MISP CHALLENGE 2022

Li Zhang[1†], *Huan Zhao*[1†], *Yue Li*[1], *Bowen Pang*[1], *Yannan Wang*[2],
Hongji Wang[2], *Wei Rao*[2], *Qing Wang*[1], *Lei Xie*[1,*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University (NPU), Xi'an, China
[2]Tencent Corporation, Shenzhen, China

ABSTRACT

This paper describes the FlySpeech speaker diarization system submitted to the second **M**ultimodal **I**nformation Based **S**peech **P**rocessing (**MISP**) Challenge held in ICASSP 2022. We develop an end-to-end audio-visual speaker diarization (AVSD) system, which consists of a lip encoder, a speaker encoder, and an audio-visual decoder. Specifically, to mitigate the degradation of diarization performance caused by separate training, we jointly train the speaker encoder and the audio-visual decoder. In addition, we leverage the large-data pretrained speaker extractor to initialize the speaker encoder. Then we particularly explore the effectiveness of different frameworks of audio-visual decoder, including transformer, conformer, and cross-attention mechanism. In the decoding phase, we adjust the decoding frame shift to reduce boundary error and increase the stability of decoding results. Moreover, we introduce median filtering and secondary speaker verification as post-processing methods to smooth the predicted probabilities of speaker activities and to reduce speaker prediction errors on the single-speaker segments, respectively. With the above contributions, our best system achieves 10.90% diarization error rate (DER) on the MISP evaluation set, being ranked 3rd on the diarization track of this challenge.

Index Terms— audio-visual, speaker diarization, MISP

1. INTRODUCTION

Speaker diarization (SD), sometimes called active speaker detection, is the task of determining "who spoke when?" in an audio or video recording [1]. Its application is indispensable in multimedia information retrieval, speaker turn analysis, and multi-speaker speech recognition. With the emergence of wide and complex application scenarios of SD, uni-modal (audio- or visual-based) SD encounters performance bottleneck [2, 3]. Therefore, the MISP challenge 2022 [4, 5] particularly set up a multi-modal SD track to promote effective integration of audio and visual information, aiming to obtain better environmental and speaker robustness in realistic applications.

In this challenge, we focus on building a robust end-to-end neural audio-visual speaker diarization (AVSD) system, which is composed of a (visual) lip encoder, an (audio) speaker encoder as well as an audio-visual decoder. Importantly, we take a large-data pre-trained speaker extractor to initialize the speaker encoder and train the speaker-encoder-joint-decoder system in an end-to-end manner using the challenge data. Leveraging the pre-trained speaker extractor and the joint training strategy leads to improved system robustness and substantial SD performance gain. In the decoder module, we introduce different neural network frameworks (transformer [6], conformer [7], cross-attention [8]) to effectively fuse audio and visual representations. In addition, we adjust the decoding frame shift and make median filtering [9] on predicted probabilities to smooth the predicted results. Finally, we leverage the pre-trained speaker encoder to perform secondary speaker verification of single-speaker speech segments in the predicted Rich Transcription Time Masked (RTTM). Together with the above strategies, our superior system achieves 10.90% DER on the evaluation set at the AVSD track of MISP challenge 2022.

2. SYSTEM DESCRIPTION

An overview of our system is illustrated in Fig 1. It consists of visual and audio feature extractors (F_v and F_a), uni-modal encoders (lip encoder and speaker encoder), an audio-visual speaker decoder, and post-processing on decoding results. Specifically, the lip encoder and pretrained speaker extractor in the grey boxes are pretrained and do not participate in audio-visual decoder training. They are used to extract lip and speaker embeddings. Meanwhile, the parameters of the speaker encoder are initialized by the pretrained speaker extractor. The embeddings from lip encoder, speaker encoder, and pretrained speaker extractor are fused and fed into the audio-visual speaker decoder for speaker activation probability prediction. The audio-visual speaker decoder is to decode speaker activities of each frame into probability matrices which represent whether each speaker speaks at each frame. In the post-processing module, we first apply median filtering to smooth the predicted probabilities of speaker activity detection. Then we use secondary speaker verification (SV) on the single-speaker segments particularly to reduce speaker prediction errors in RTTM, where the pretrained speaker extractor serves as the SV model.

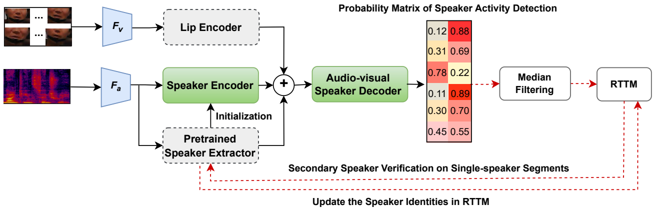
Probability Matrix of Speaker Activity Detection
F_v
Lip Encoder
F_a
Speaker Encoder
Initialization
Pretrained Speaker Extractor
Audio-visual Speaker Decoder
0.12 0.88
0.31 0.69
0.78 0.22
0.11 0.89
0.30 0.70
0.45 0.55
Median Filtering
RTTM
Secondary Speaker Verification on Single-speaker Segments
Update the Speaker Identities in RTTM

Fig. 1: Overview of the submitted system. The encoders in the grey boxes are only used to extract embeddings and do not involved in optimization during joint training. The red dashed line represents post-processing.

Audio and Visual Feature Extraction (F_a and F_v): For audio feature extraction, 80-dimensional log Mel-filter bank features

[†]Equal contribution.

*Corresponding author.

are extracted from input speech with 25ms window size and 10ms window shift. The visual features are the regions of interest (ROI) of the lips generated from the position detection results provided by the challenge organizers [5, 3].

Speaker and Lip Encoders: In the speaker encoder module, we exploit two kinds of neural networks – ResNet34-SE [10] and ECAPA-TDNN [11]. ResNet34-SE is configured with 256 channels of squeeze-and-excitation (SE) attention modules and the channel configurations of residual blocks are $\{32, 64, 128, 256\}$. The embedding dimension of ResNet34-SE is 128. ECAPA-TDNN [11] is configured with 1024 channels. In addition, the speaker encoder is initialized by the pretrained speaker extractor which leverages a large training dataset. But differently, the pretrained speaker extractor extracts utterance-level speaker embeddings, while the speaker encoder, joint trained with the audio-visual decoder, is used to extract frame-level speaker embeddings. The lip encoder is configured as [3] and it is pretrained using the lip feature and the speaker activity status provided by the RTTM .

Audio-visual Speaker Decoder: For the audio-visual speaker decoder, we explore transformer [6], conformer [7] and cross-attention [8] respectively. The transformer decoder and conformer decoder is configured as [6, 7]. For the cross-attention decoder, we first perform cross-attention between lip and utterance-level speaker embeddings extracted from the prertained speaker extractor. Then another cross-attention is further performed between the output of the first cross-attention and frame-level speaker embeddings from speaker encoder, which jointly train with audio-visual decoder. Finally, we use a set of binary classification layers to predict the speaker activation probability from the fused representation.

Post-processing: In the decoding stage, we adjust the frame shift from 600 frames [3] to 100 frames to reduce boundary error and get stable decoding results. Then we use median filtering [9] to smooth the predicted probabilities of speaker activity detection. Furthermore, for the single-speaker segments produced by RTTM, we use the pretrained speaker extractor to cluster the single-speaker segments to recheck and correct the speaker identities in RTTM.

Datasets: Table 1 details the specific data for training each module in Fig. 1. All datasets except MISP are open-sourced on openSLR.

Table 1: The training data used for each module in Fig. 1.

Module	Training data (from MISP & OpenSLR)
Lip Encoder	MISP training set
Pretrained Speaker Extractor (ResNet34-SE)	CN-CELEB 1 & 2, VoxCeleb 1 & 2, MISP training set, MUSAN, RIRs
Pretrained Speaker Extractor (ECAPA-TDNN)	CN-CELEB 1 & 2, VoxCeleb 1 & 2, Primewords, DataTang 200, Aishell-1, ST-CMDS, MISP training set, MUSAN, RIRs
Speaker Encoder (ResNet34-SE/ECAPA-TDNN)	MISP training set
Audio-visual Speaker Decoder	MISP training set

3. EXPERIMENTAL RESULTS

Results are reported in diarization error rate (DER) which considers false alarm (FA), missed detection (MISS), and speaker error (SPKERR) simultaneously.

In this challenge, we build four AVSD systems. The first three systems are **ResNet-Transformer**, **ResNet-Conformer**, and **ResNet-CrossAttention**, whose speaker encoders are ResNet34-SE and decoders are transformer, conformer, and cross-attention, respectively. The last system is **ECAPA-Transformer**, which consists of ECAPA-TDNN speaker encoder and transformer decoder. The experimental results of these systems and post-processing methods are illustrated in Table 2. The second and the third rows in Table 2 show that transformer is the best choice for an audio-visual decoder compared with others. With the help of post-processing strategies (decoder frame shift, median filtering, and secondary speaker verification), ECAPA-Transformer is the superior system, which obtains 8.53% of DER on the development set. Finally, this system is submitted to the leaderboard, which achieves 10.90% DER on the evaluation set.

Table 2: The experimental results (%) on the development set and evaluation set (The last line is the result on the evaluation set).

System	FA	MISS	SPKEER	DER (Dev/Evl)
Baseline AVSD [3]	4.01	5.86	3.22	13.09/-
ResNet-Transformer	1.36	6.23	1.92	9.54/-
ResNet-Conformer	2.01	5.50	2.10	9.61/-
ResNet-CrossAttention	1.35	6.26	1.95	9.57/-
ECAPA-Transformer	1.64	5.77	1.89	9.30/-
+ Decoding Frame shift (100)	1.64	5.62	1.89	9.15/-
+ Median Filtering	2.31	4.75	1.86	8.92/-
+ Secondary SV	**1.95**	**4.79**	**1.78**	**8.53/10.90**

4. CONCLUSION

In the AVSD track of MISP challenge 2022, we explore three strategies to improve the end-to-end AVSD system. First, we jointly train the speaker encoder and the audio-visual decoder instead of separate training. Then we explore the effectiveness of different structures for the audio-visual decoder. To further improve the AVSD performance, we adopt useful post-processing tricks, which are adjusting frame shift, median filtering on predicted probabilities, and secondary speaker verification on single-speaker segments. Together with the above strategies, our submission achieves 10.90% DER on the evaluation set in the AVSD track, reaching the third place.

5. REFERENCES

[1] Xavier Anguera, Simon Bozonnet, and Evans Nicholas et al., “Speaker diarization: A review of recent research,” *TASLP*, 2012.

[2] Shinji Watanabe, Michael Mandel, and Barker Jon et al., “Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” *Interspeech*, 2020.

[3] Mao-kui He, Jun Du, and Chin-Hui Lee, “End-to-end audio-visual neural speaker diarization,” in *Proc. Interspeech 2022*, 2022, pp. 1461–1465.

[4] Hang Chen, Hengshun Zhou, and Du Jun et al., “The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results,” in *ICASSP*. IEEE, 2022.

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

[6] Narla John Metilda Sagaya Mary and Umesh Srinivasan et al., “S-vectors and tesa: Speaker embeddings and a speaker authenticator based on transformer encoder,” *TASLP*, 2021.

[7] Yang Zhang, Zhiqiang Lv, and Wu Haibin et al., “Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification,” *Interspeech*, 2022.

[8] Ruijie Tao, Zexu Pan, and Das Rohan Kumar et al., “Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection,” in *ACM MM*, 2021.

[9] Ivan Medennikov, Maxim Korenevsky, and Prisyach Tatiana et al., “Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario,” *Interspeech*, 2020.

[10] Hee Soo Heo, Bong-Jin Lee, and Huh Jaesung et.al., “Clova baseline system for the voxceleb speaker recognition challenge 2020,” .

[11] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *Interspeech*, 2020.