CHANNEL-WISE AV-FUSION ATTENTION FOR MULTI-CHANNEL AUDIO-VISUAL SPEECH RECOGNITION

Gaopeng Xu, Song Yang, Wei Li, Song Wang, Guo Wei, Junfeng Yuan, Jie Gao

NIO Co., Ltd.

ABSTRACT

In this paper, we present our work for automatic speech recognition (ASR) in the Multimodal Information Based Speech Processing (MISP) Challenge 2021. We proposed a combination of the guided source separation-based (GSS) speech enhancement technique and a novel Channel-wise Av-fusion encoder (CAE) based acoustic model and found that a kindly combination of these techniques provided essential accuracy improvements. Our ASR system reduces the Chinese Character Error Rate (CCER) by 37.67% absolute compared to the baseline in track 2, achieving first place in the evaluation period with the CCER of 25.07%.

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

1. INTRODUCTION

Far-field automatic speech recognition (ASR) is essential area of research and have many real-world applications such as transcribing meetings and shopping center conversations. Although end-to-end ASR approaches have achieved promising results for some open datasets such as AISHELL [1] and LibriSpeech [2], their performance remains unsatisfactory for far-field conditions in real environments. Far-field ASR can be challenging due to background noise and reverberation in the acoustic environments, conversational multi-speaker interactions with a large portion of speech overlap. In addition, due to the lack of a public dataset of far-field speech, it is more challenging to develop an ASR system in this field.

Over the past decade, some progress has been made in addressing the challenges posed by far-field speech. For ASR, this improvement can be attributed to effective data augmentation [3], advances in speech enhancement [4] and E2E-ASR [5-9] techniques. Some work has tried tackling reverberation and noise present in the far-field recording by training with room impulse responses and background noises [10]. Recently, spectral augmentation [11] has been successfully used for end-to-end and hybrid ASR systems. Adapting the acoustic model to the environment and speaker has also been studied. Another popular direction is frontend-based approaches such as dereverberation [12] and denoising through beamforming [13]. Far-field speaker diarization has also benefited from enhancement methods and techniques to handle overlapping speech. Guided source separation (GSS) [14] was proposed, which uses additional information such as time and speaker annotations for mask estimation. Some works [15, 16] have shown that additional multimodal information can significantly improve speech recognition robustness in noisy and far-field environments.

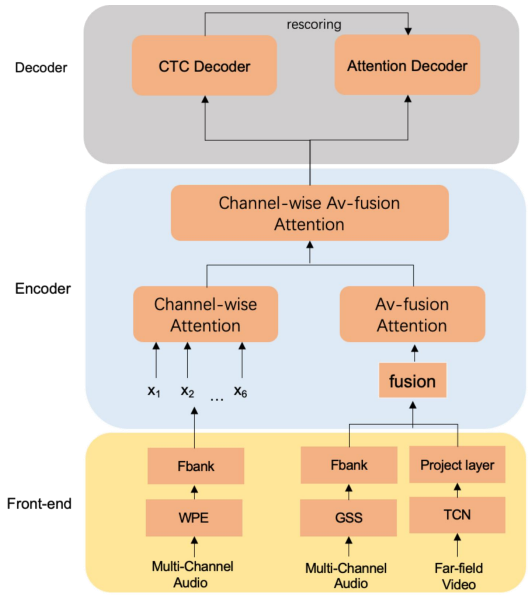

Fig. 1. An illustration of overall framework.

The Multi-modal Information Based Speech Processing (MISP) Challenge [17] considers the problem of distant multi-microphone conversational audio-visual wake-up and audio-visual speech recognition in everyday home environments. The challenge consists of two tracks, audio-visual wake word spotting track and audio-visual speech recognition with oracle speaker diarization. For track 2, all audio data are WAV files with a sampling rate of 16 kHz. Each session consists of the recordings made by the far-field linear microphone array with 6 microphones, the middle-field linear microphone array with 2 microphones and the high-fidelity near-field microphones worn by each participant. Video data are distributed as MP4 files with a frame rate of 25 fps, and each session consists of the recordings made by the far-field wide-angle camera and the

middle-field high-definition cameras worn by each participant. Our systems are designed for track 2. Figure 1 shows the framework of our systems. The front-end includes weighted prediction error (WPE) based dereverberation for multi-channel signals and multi-channel speech enhancement with guided source separation (GSS). Lipreading TCN [18] extracts 512-dimensional visual features and 80-dimensional FBANK as audio features. We use a project layer and sum-fusion for multi-modal features to get the audio-visual features. For the encoder module, we use channel-wise attention to get channel-wise outputs per channel, Av-fusion attention to getting audio-visual embeddings, and Channel-wise Av-fusion attention to learn the contextual relationship across channels and modals. For the decoder module, similarly to WENET [19], the CTC decoder consists of a linear layer, transforming the encoder output to the CTC activation and the attention decoder consists of multiple Transformer decoder layers. During the inference stage, the CTC decoder generates n-best candidates and then rescore the n-best candidates on the attention decoder to get the finally result. The detailed descriptions of the systems can be found in the following sections.

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

2. FRONT-END PROCESSING

2.1. Dereverberation and Denoising

We used the NARAWPE tool implementation of weighted prediction error (WPE) [20] based dereverberation for multi-channel signals.

2.2. Multi-channel guided source separation

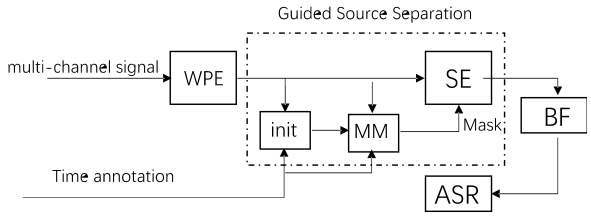

Fig. 2. An illustration of GSS framework.

GSS enhancement is a source separation technique originally proposed in CHiME-5. GSS aims to separate the sources using a pure signal processing approach. An Expectation-Maximization (EM) [21] algorithm estimates the parameters of a spatial mixture model, and the posterior probabilities of each speaker being active are used for mask-based beamforming. The overall framework of GSS is given in Fig. 2. The system consists of two stages: (1) a dereverberation stage (2) a guided source separation stage. The multiple input multiple output version of the WPE method was used in the dereverberation stage. Guided source separation consists of a spatial Mixture Model and a source extraction (SE) component. All the 6 channels audios are dereverbed with the WPE. SE masks combined with time annotations are served as an initialization of the Mixture Model (MM). After iterations, the masks representing the target speaker and the inference are used to do beamforming.

2.3. Front-end experiments

The whole audio set contains near-field data (NEAR), middle-field microphone array data (MIDDLE), far-field microphone array data (FAR), and multi-channel enhanced data (ENH), totally 4 parts. NEAR data is produced by the high-fidelity near-field microphones worn. MIDDLE and FAR data are made up of the original 6 microphones far-field and 2 microphones middle-field audios. ENH data is made up of MIDDLE and FAR data enhanced by GSS model. FAR_SP represents the training set obtained by a 3-fold speed perturbation of FAR and enhanced FAR. The details of the audio set are listed in Table 1.

Data	Description	Fold
NEAR	near-field audios	1
FAR	Original far-field audios	6
MIDDLE	Original middle-field audios	2
ENH	Augmented enhanced far-field and middle-filed audios	2
FAR_SP	3-fold speed perturbation FAR and enhanced FAR	7
ALL	NEAR+FAR+MIDDLE+ENH	11
ALL_SP	3-fold speed perturbation	33

Table 1. The details of the audio set, where 1-fold data is around 106 hours

We present the front-end results on several training sets. First, compared with the Baseline model CCER of 61.78%, We implemented five versions yield comparable CCER, as listed in Tabel2. We see that using the transformer-based model M2 improved the baseline system's performance by about 3% absolute. M3 shows the improvement gained by 11-fold data adjunction, a 6.3% decrease in the CCER was achieved by increasing the middle-field and near-field data. M4 shows the improvement gained by using the GSS. M5 is like M 4, but a 3-fold speed perturbation was used for all training data, and the CCER dropped from 34.21% to 32.51%. In M6, we used the M5 encoder to initialize the encoder of M6, and the model achieved a CEER of 32.02%. Data augmentation, GSS front-end and Av-fusion were individually effective for further improvement, and their combination provided a significant CCER improvement.

Model	Architecture	dataset	Front-end	CCER
M1	TDNN-F	FAR_SP	WPE+BF	61.78%
M2	speech-transformer	FAR_SP	WPE+BF	58.46%
M4	speech-transformer	ALL	GSS	34.21%
M5	speech-transformer	ALL_SP	GSS	32.51%
M6	AV-fusion-transformer (M5 init)	FAR_SP	GSS	32.03%

Table 2. comparison between different front-end on development set

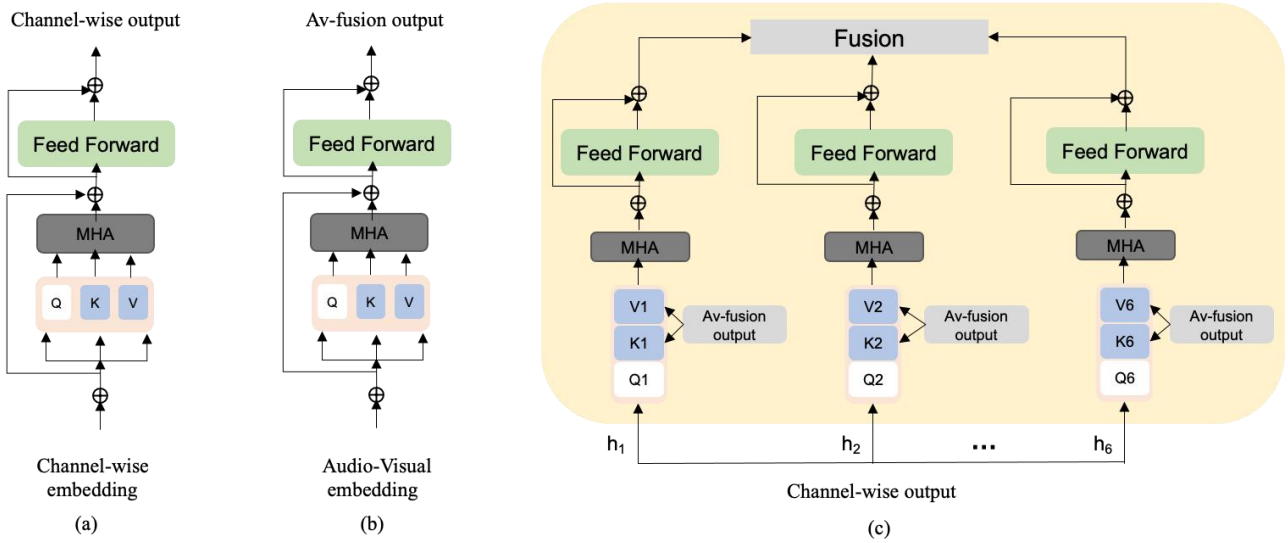

Fig. 3. An overview of Channel-wise Av-fusion Encoder (CAE). (a) Channel-wise encoder architecture. (b) Av-fusion encoder architecture. (c) Channel-wise Av-fusion Encoder architecture.

3. ACOUSTIC MODELS

3.1. Encoder

We propose an end-to-end channel-wise audio-visual fusion network that can synchronously use the audio and video feature and use the attention mechanism to learn the contextual relationship cross-channel and cross-modality. An illustration of Channel-wise Av-fusion Encoder (CAE) is given in Fig. 3. We use channel-wise [22] encoder to get channel-wise outputs per channel. In channel-wise encoder, we use the source channel embedded features plus the positional encoding are fed into a set of weight parameters to create Query (Q), Key (K), Value (V). Q and K compute the correlation across time steps within a channel multi-head attention. The attention matrix is then used to reweight the features of V in each time step followed by a feed-forward network to produce the channel-wise outputs. Similar to the channel-wise encoder, we use Av-fusion attention to get audio-visual fusion output. Given the channel-wise outputs per channel, the Channel-wise Av-fusion attention layers to learn the contextual relationship across channels and modals. We use the channel-wise output to create Q, and the Av-fusion encoder outputs to create K and V.

3.2 Decoder

For the decoder module, the CTC decoder consists of a linear layer, which transforms the encoder output to the CTC activation. The attention decoder consists of multiple Transformer decoder layers. As illustrated in Fig. 4, during the inference stage, the CTC decoder generates n-best hypothesis. After CTC beam search decoding, a WFST-based [23] decoder 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.

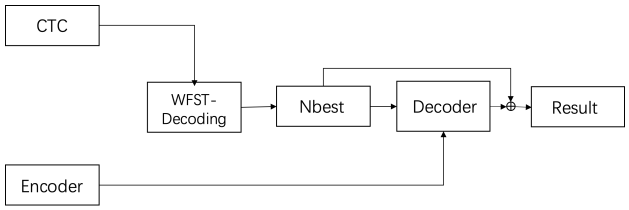

Fig. 4. An overview of decoder

4. EXPERIMENTS

We perform our experiments on the MISP2021 audio-visual dataset, which contains 110+ hours of audio-visual data. The dataset includes 340+ sessions. Each session consists of an about 20-minute discussion. The dataset has been split into training, development test and evaluation test. For audio features, the 80-dimensional FBANK are computed by Torchaudio with a 25ms window and a 10ms shift. SpecAugment [24] has applied 2 frequency masks with maximum frequency mask, and 2 times masks with maximum time mask to data augmentation. 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 05. The model was optimized with Adam, and the learning rate was warmed-up for 25000 steps. Finally, we obtain our final model by averaging the top-10 best models with a lower loss on the development set.

Model	Architecture	Front-end	CCER
M7	Baseline	WPE+BF	62.74%
M8	Av-fusion transfomer	GSS	31.94%
M9	Channel-wise Av-fusio transformer(M8 init)	GSS	25.07%

Table 3. Submitted results.

Table 3 shows the result submitted for the MISP challenge track 2. The CCER on the eval set of the baseline model was 62.74%. M8 shows the improvement gained using the Av-fusion transformer model, achieving a CCER of 27.6%. In M9, we used the M8 encoder to initialize the encoder of the Channel-wise Av-fusion encoder, and the WER dropped from 31.94% to 25.07%.

5. CONCLUSION

We proposed combining the guided source separation-based (GSS) speech enhancement technique and a novel Channel-wise audio-visual fusion encoder for multi-channel audio-visual speech recognition, which can learn the contextual relationship cross-modality and significantly improve speech recognition robustness in far-field environments. Our best result achieved a CCER of 25.07% for the evaluation set, with an absolute reduction of 37.67% compared to the baseline model.

6. REFERENCES

[1] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, 2017: 1-5.

[2] Panayotov V, Chen G, Povey D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 5206-5210.

[3] Kanda N, Boeddeker C, Heitkaemper J, et al. Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR[J]. arXiv preprint arXiv:1905.12230, 2019.

[4] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in CHiME5 Workshop, Hyderabad,India, 2018.

[5] L. Dong, S. Xu and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5884-5888.

[6] Y. zhao, J. Li, X. Wang and Y. Li, "The Speechtransformer for Large-scale Mandarin Chinese Speech Recognition," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7095-7099.

[7] K. Anjuli, D. Arindrima, N. Tara, "Large-scale multilingual speech recognition with a streaming end-to-end model," in Proc. ISCA Interspeech, 2019, pp. 2130–2134.

[8] A. Gulati, J. Qin, C. C. Chiu, et al. "Conformer: Convolution augmented transformer for speech recognition," arXiv preprint, arXiv:2005.08100, 2020.

[9] K. J. Han, J. Pan, V. K. N. Tadala, et al. "Multistream CNN for robust acoustic modeling," arXiv preprint, arXiv:2005.10470, 2020.

[10] Y. Wang, D. Snyder, H. Xu, V. Manohar, P. S. Nidadavolu,D. Povey, and S. Khudanpur, "The JHU ASR system for VOiCES from a Distance challenge 2019," Proc. Interspeech 2019, pp. 2488–2492, 2019.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.
[12] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "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.
[13] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2011–2022, 2007.
[14] C. Zorila, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach,"An investigation into the effectiveness of enhancement in ASR training and test for CHiME-5 dinner party transcription," arXiv preprint arXiv:1909.12208, 2019.
[15] Zhou P, Yang W, Chen W, et al. Modality attention for end-to-end audio-visual speech recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6565-6569.
[16] Yu J, Wu B, Gu R, et al. Audio-visual multi-channel recognition of overlapped speech[J]. arXiv preprint arXiv:2005.08571, 2020.
[17] Chen, Hang and Zhou, Hengshun and Du, Jun and Lee, Chin-Hui and Chen, Jingdong and Watanabe, Shinji and Siniscalchi, Sabato Marco and Scharenborg, Odette and Liu, Di-Yuan and Yin, Bao-Cai and Pan, Jia and Gao, Jian-Qing and Liu, Cong, The First Multimodal Information Based Speech Processing (MISP) Challenge: Data, Tasks, Baselines and Results, Proc. ICASSP 2022, 2022.
[18] Martinez B, Ma P, Petridis S, et al. Lipreading using temporal convolutional networks[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6319-6323.
[19] Zhang B, Wu D, Yang C, et al. WeNet: Production First and Production Ready End-to-End Speech Recognition Toolkit[J]. arXiv preprint arXiv:2102.01547, 2021.
[20] Drude L, Heymann J, Boeddeker C, et al. NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing[C]//Speech Communication; 13th ITG-Symposium. VDE, 2018: 1-5.
[21] Moon T K. The expectation-maximization algorithm[J]. IEEE Signal processing magazine, 1996, 13(6): 47-60.
[22] Chang F J, Radfar M, Mouchtaris A, et al. Multi-Channel Transformer Transducer for Speech Recognition[J]. arXiv preprint arXiv:2108.12953, 2021.
[23] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 167–174.
[24] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.