THE XIAOMI-TALKFREELY SYSTEM FOR AUDIO-VISUAL SPEECH RECOGNITION IN MISP CHALLENGE 2021

Quandong Wang[1†], Xinyu Cai[2†], Weiji Zhuang[1], Yuxiang Kong[1*], Yongqing Wang[1*], Junnan Wu[1*], Dongbo Li[1], Zhiyong Yan[1],Mingshuang Luo[1], Xinyu Tang[1], Liyong Guo[1], Zhigao Chen[1], Yuquan Liang[1], Shijie Deng[1], Lichun Fan[1], Junbo Zhang[1], Peng Gao[1], Yujun Wang[1], Ying Huang[1], Zhiyong Wu[2]*

[1]Xiaomi Inc., Beijing, China
[2] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

ABSTRACT

This technical report describes our submitted system to task2 of the Multimodal Information Based Speech Processing (MISP) Challenge 2021. Task2 is audio-visual speech recognition with oracle speaker diarization. Our main technical points include the traditional and deep learning based speech enhancement and separation, training data augmentation via various kinds of techniques, acoustic model fusion, and K2 based WFST decoding and rescoring. Tested on the development set and the evaluation set respectively, our best system has yielded absolute Chinese character error rate (CCER) reduction of 37.7% and 35.6% compared to the official baseline system and ranked the second place among the submitted systems in the challenge.

Index Terms— speech recognition, neural network, signal processing, multi-modal

1. INTRODUCTION

In recent years, automatic speech recognition(ASR) for single-speaker and clean speech has substantially improved, and it has been widely utilized in meeting transcription, virtual voice assistant, and automatic captioning. However, there are still some challenges in applying ASR to real-world scenarios. Multiple interference, such as multi-speaker speech overlap, background noise, and reverberation, may occur in the far-field home and meeting interactive scenarios, causing the performance of the ASR system to decrease significantly.

Previous studies have mainly focused on extracting the target speaker's clean speech from the noisy signal. Multi-channel speech enhancement techniques such as BeamformIt [1], weighted prediction Error(WPE) [2] and guided source separation(GSS) [3] are widely used and show good performance.

Since the visual modal is not affected by the noise, authors have investigated the utilization of visual information in speech recognition. Traditional audio-visual ASR system follows a two-stage scheme. A feature extraction method [4] or a pretrained model [5] is used to obtain visual features, which are then combined with audio features and fed into the ASR model. Recently, some methods have been presented to train the extraction and recognition model jointly and got better results over the two-stage methods. Besides supervised methods, the self-supervised learning method achieves the state of the art performance by predicting the masked units to learn representation of audio-visual features[6].

However, there is no specific benchmark for audio-visual speech recognition(AVSR) in complex scenarios. In this context, the multimodal information based speech processing (MISP) challenge [7] provides the first AVSR dataset collected in everyday home environments, which aims to improve the performance of speech recognition systems by combining both audio and visual modal information.

During the challenge, we established an AVSR system including speech separation front-end and speech recognition back-end, investigated the impact of different front-ends, and compared the effects of audio-only and audio-visual recognition back-ends.

2. PROPOSED SYSTEM

2.1. system overview

The dataset used in MISP challenge task2 contains 110+ hours of audio-visual data. The videos in the dataset record multiple speakers chatting in the living room while the TV is playing, making speech recognition quite challenging. To solve this task, we made a rough problem analysis on task2. Firstly, the core problem of task2 is to recognize the target speaker's speech in the context of interfering speech overlap, reverberation, and background TV noise, so the front-end which aims to provide high-quality speech is needed. Secondly, the amount of data that can be used in this competition is limited, so data augmentation is needed. Thirdly, there exists time dislocation in far-field speech, so we did some data clean-up work to make sure time consistency between

* and † stand for equal contribution

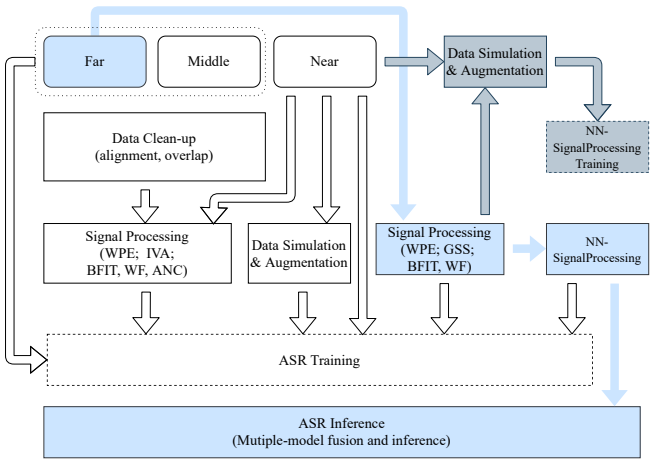

Fig. 1. An illustration of our proposed system on task2.

far and near field speech. And the solutions to the above are gathered up to build our system. The illustration of our system is displayed in Fig.1. It briefly depicts our training and inference process, in which arrows and blocks in white and gray represent the ASR training and neural network (NN) based signal processing training process respectively, while the blue ones are for the whole inference process.

During the inference stage, the far-field data is handled by traditional signal processing followed with NN based signal processing model which we chose to use SpEx+ [8]. After that, the output will be sent to the ASR model which includes both audio models and audio-video models. When training NN based signal processing model we use near-field speech and data provided by signal processing, like WPE, GSS, etc, for data augmentation and simulation. For concise presentation, we only report one of the trained four SpEx+ models in the experiment section. And for training ASR models, we also did a lot of data simulation and augmentation work, which includes augmentation based on original data, the data provided by traditional signal processing, and the data provided by NN based signal processing. Speed perturb as well as SpecAugment [9] are also taken into use when training the audio-only ASR model.

2.2. front-end

For the front-end of our system, traditional signal processing methods including multi-channel WPE (weighted prediction error) [2] and GSS (guided source separation) [3] were used.

2.2.1. WPE

Reverberation is inevitable in a far-field speech environment and can degrade the performance of microphone array processing methods and ASR. WPE is a commonly used dereverberation algorithm, which estimates the late reverberation and subtracts it from the observed signal. For simplicity, we used nara-wpe [10] in our system.

2.2.2. GSS

Source separation is essential in multi-speaker scenarios. Here we applied GSS as our source separation method because it had neither frequency permutation nor global source permutation problem. GSS is an offline source separation method, which is based on a complex Angular Central Gaussian Mixture Model (cACGMM) [11]. GSS avoids solving the permutation problems by exploiting the source activity information, which could be derived from the time annotations provided. The source activity information, set as one or zero depending on whether the speaker is active or not, is used to guide the parameters estimation of the mixture model. Furthermore, an additional class is used indicating the background noise, and its activity information was always set to be one. Finally, a context was used to reduce the permutation problem between the target source and the noise.

2.2.3. SpEx+

SpEx+ [8], a single-channel time-domain speaker extraction network, is used after GSS in our system to make further improvements on the separation effect. SpEx+ consists of speech encoder, speaker encoder, speaker extractor, and speech decoder as shown in Fig. 2.

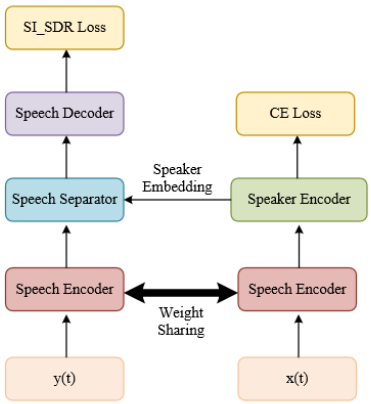

Fig. 2. Illustration of SpEx+ network

2.3. back-end

We built our back-end using the ESPnet toolkit [12]. The architecture of the back-end is shown in fig. 3. We focused on optimizing the audio-only back-end. Meanwhile, we explored the gains brought by the audio-visual back-end.

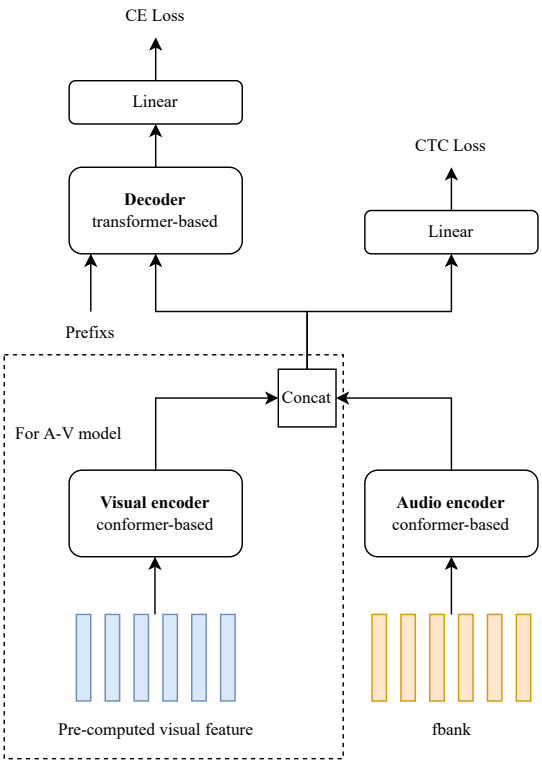

Fig. 3. The architecture of our speech recognition back-end

2.3.1. *end-to-end acoustic models*

Fig. 3 shows the whole architecture our end-to-end back-end models. Our audio-only speech recognition back-end has an encoder-decoder structure. We used convolution-augmented Transformer (Conformer) [13] with relative positional encoding-based self attention as basic encoder block, and followed the configuration in [12].

As for audio-visual speech recognition, inspired by [14], we add an extra conformer-based visual encoder to the audio-only back-end. For audio modal of our backend, we extracted 80-channel filterbanks features computed from a 25ms window with a stride of 10ms. For visual modal of our backend, we use a lipreading model pretrained on LRW-1000 dataset [15] to extract 512-dimensional visual features as input to the video encoder. The structure of the video encoder is similar to the audio encoder mentioned above. In the following audio-visual experiments, we simply concatenated visual and audio features and feed them into the decoder. We also adopt a simple model fusion strategy. We select the nine best models and get their utterance-level output scores. For each utterance, we choose the output text with the highest score as the final text.

2.3.2. *WFST based decoding*

Normally, we use beam search based decoding as in the ESPnet default setting, in which the final score is the weighted sum of encoder score, decoder score, and neural network language model score. To make full use of the n-gram model and improve the decoding speed. We introduce WFST (Weighted Finite-State Transducer) based decoding into the end-to-end model. The decoding WFST graph is made in the same way as[16]. We get n-best from WFST based decoding. After that, nbest will perform rescoring. Decoder score and neural network based language model score are added with AM score and LM score from the WFST to generate the final score.

The WFST based decoding was implemented by K2 framework [17]. The entire decoding process runs on the GPU fast. From Table 1 we can find the TLG based decoding with rescoring results is better than ESPnet default beam search decoding. Since the WFST based decoding introduces more decoding parameters, we do not have enough time for parameter search. So the final submitted result excluded WFST based decoding.

Decoding method	Far-field development set
beam search	30.9
TLG WFST	29.4

Table 1. Performance of our K2 WFST based decoding.

3. EXPERIMENTAL RESULTS

3.0.1. *data clean-up*

Our data clean-up contains overlap check and far-near alignment. Overlap check is to count the number of speakers in each utterance based on the textgrid the challenge offers. Doing far-near alignment, we filtered out the utterance whose far and near-field data were misaligned in time based on the cross-correlation function to ensure the audio for training ASR can correspond to the text.

3.0.2. *front-end comparison*

Here we compare three kinds of front-ends. The first one is multi-channel WPE followed by BeamformIt [1] (WPE + BeamformIT) which is the official baseline. The second one is GSS, while the third one is SpEX+ following GSS (GSS + SpEx+). Here GSS denotes the system in [3] containing WPE, separation and post-processing modules. The SpEx+ model was trained using 500-hour data generated by mixing the single speaker utterances processed by GSS. The speaker reference was also processed by GSS. Table 2 compares the CER results of the far-field development and evaluation sets using the same back-end model. The "GSS + SpEx+" method performed best and was used for submission.

Front-end	Development set	Evaluation set
WPE + BeamformIT	43.5	46.2
GSS	29.5	28.4
GSS + SpEx+	29.0	28.2

Table 2. CERs of three front-ends on the same back-end.

3.0.3. back-end comparison

For back-end training, data augmentation was applied. We simulated far-field speech with near-field data, enhanced or separated the middle- and far-field data with methods including IVA, BeamformIT, WPE, GSS as well as SpEx+. We conducted experiments on two datasets: *Train-base* and *Train-all*. *Train-base* is the official train dataset processed by WPE and BeamformIt. *Train-all* is a combination of all augmented data mentioned, with a total duration of 3000 hours. The former dataset is used to compare our systems with the official back-end, and the latter is used to explore the best performance of our proposed back-end.

Table 3 shows the performance of our audio-only(A) and audio-visual(A-V) back-ends. When trained on the same dataset *Train-base*, our A and A-V back-ends outperformed the official baseline back-ends by absolute 4.1 % and 3.7 %, respectively on the development set. Visual information helps the A-V back-end achieve a better result (58.1% vs. 58.9%) on the same set. When trained on *Train-all*, our results were significantly better than the baseline. Results on Eval-FE set show that our back-end could still gain 0.7 % from visual information (28.59% vs. 29.30%), but we did not get consistent results on Dev-FE set. It is probably due to the different distribution of the two sets. The best result we submitted is a fusion of 9-best models, which achieved 27.17% on the evaluation set, while the official baseline only got 62.74% on the same set.

Back-end	Data	Dev	Dev-FE	Eval-FE
Official A	*Train-base*	63.0	-	-
Official A-V	*Train-base*	61.8	-	-
Ours A	*Train-base*	58.9	49.2	46.16
	Train-all	41.9	27.0	29.30
	Train-all+Dev-FE	40.4	22.0	28.90
Ours A-V	*Train-base*	58.1	51.2	51.69
	Train-all	41.3	27.5	28.59
model-fusion	-	-	24.1	**27.17**

Table 3. Performance of our audio-only back-end (A) and audio-visual (A-V) back-ends in terms of CER[%]. "Dev" means development data processed by WPE and BeamformIt, while "Dev-FE" and "Eval-FE" means data processed by our best front-end. The last line was submitted.

To explore the impact of video quality on the A-V back-end, we also conducted experiments on middle-field *Train-base* dataset. As we can see from the results in Table 4, A-V back-end achieved a noticeable boost of 3.2% (4.0 vs. 0.8) on the visual information gains with middle-field video, which indicates that A-V ASR needs high-quality videos.

Back-end	Far-field	Middle-field
Ours Audio	58.9	49.1
Ours A-V	58.1	45.1

Table 4. Performance of A-V model on far-/middle- field datasets with official front-end of WPE and BeamformIT.

4. CONCLUSION

This technical report proposes our submission to task2 of the MISP 2021 challenge. Our work includes the investigation of different speech enhancement front-end and comparison of audio-only and audio-visual speech recognition back-ends. Our proposed AVSR system improves against the baseline with an absolute reduction of 35.6 % on the evaluation dataset and ranks 2nd out of 10 participated systems in the challenge.

5. REFERENCES

[1] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[2] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[3] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018, vol. 1.

[4] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[5] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 905–912.

[6] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[7] Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, Jia Pan, Jian-Qing Gao, and Cong Liu, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.

[8] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li, "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.

[9] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold 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.

[11] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

[12] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[14] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[15] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[16] Yajie Miao, Mohammad Gowayyed, and Florian 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.

[17] Daniel Povey, Piotr Żelasko, and Sanjeev Khudanpur, "Speech recognition with next-generation kaldi (k2, lhotse, icefall)," *Interspeech: tutorials*, 2021.