THE CQJTU SYSTEM FOR MULTIMODAL INFORMATION BASED SPEECH PROCESSING CHALLENGE 2022

Yi Chengjiang, Peng Yi
Chongqing Jiaotong University

ABSTRACT

This technical report describes our submitted system to track1 and track2 of the Multimodal Information Based Speech Process- ing (MISP) Challenge 2022. Track1 is Audio-Visual Speaker Diarization and track2 is Audio-Visual Diarization and Recognition. Our main works include the speech enhancement , training data augmenta tion , Audio-Visual acoustic model and WFST based decoding. our best track1 system diarization error rate (DER) on the eval set was 13.05%. and our best track2 system concatenated minimum permutation character error rate (CPCER) was 30.01% and ranked the second place among the submitted systems in the challenge.

***Index Terms*—** speaker diarization, speech recognition, multi-modalgnition

1. INTRODUCTION

In recent years, speaker diarization and automatic speech recognition(ASR) for single-channel speech has substantially improved, and it has been widely utilized in meeting transcription, voice assistant, and automatic captioning. However, there are still some challenges in applying to multi-channel scenarios. Multiple interference, such as multi-speaker speech overlap, background noise, and reverberation, may occur in the far-field scenarios, causing the performance of the system to decrease significantly. The multimodal information based speech processing (MISP) challenge 2022 provides a AVSR dataset collected in everyday home environments, which aims to improve the performance of speech diarization and recognition systems by combining both audio and visual modal information.During the challenge, we established an AVSR system including speaker diarization, speech enhancement front-end and audio-visual speech recognition model. The dataset of MISP challenge contains 110+ hours of audio-visual data.All audio data are distributed as 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 near-field high-fidelity microphones worn by each participant. All video data are distributed as MP4 files with a frame rate of 25 fps.

2. PROPOSED SYSTEM

2.1 Front-end

For the front-end of our system, traditional signal processing methods including multi-channel WPE and GSS were used. WPE is a commonly used dereverberation algorithm, which estimates the late reverberation. Source separation is essential in multi-speaker scenarios. . GSS is an offline source separation method, which is based on a complex Angular Central Gaussian Mixture Model (cACGMM). 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.

2.2 Speaker Diarization

We used audio-visual neural speaker diarization model [1] provided by MISP 2022 organizer. It takes Fbank features, multi-speaker lip regions of interest (ROIs) and multi-speaker i-vector embeddings as multimodal input. A set of binary classification output layers produces the activity for each speaker. With a well-designed end-to-end structure, it can explicitly handle overlapping speech and utilize multimodal information to accurately distinguish speech from non-speech. I-vectors are the key point to solve alignment problems caused by visual modality errors.

ID	Model	DER
ID1	baseline	13.88%
ID2	Far0 channel	13.68%
ID3	6-Channels Fusion	13.05%

Table 1 Diarization evalset results for Track 1

Our speaker diarization result in Tabel1. ID1 is baseline result. ID2 is Far0 channel result and ID3 improved performance about 0.63% absolute by fusion 6-channels audio.

2.2 Acoustic models

We use Wav2vec as audio encoder. In Wav2Vec encoder pre-training step, we use about 1200 hours open-source corpora of OpenSLR (Aishell1,aidatatang and MAGICDATA) and misp2022 dataset to training Wav2Vec encoder , Wav2Vec encoder consists of a feature encoder, context network and quantization module. Our model contains 18 layers of transformer block with 16 heads, and the hidden dimension is 1024, and the FFN-dim is 4096. The model was optimized with Adam. In Wav2Vec encoder fine-tuning step, We fine-tuned on the misp2022 data with pre-training model. The model was optimized with Adam, and the learning rate was warmed-up for the first 1k of steps. For the first 1000 updates, only the final output layer was trained, after Transformer block was also trained. The feature encoder was frozen during fine-tuning step. For visual encoder we use an conformer-based visual encoder. For audio encoder we extracted 80-channel filterbanks features computed from a 25ms win- dow with a stride of 10ms. For visual encoder, we use a lipreading model pretrained on LRW-1000 dataset to extract 512-dimensional visual features as input to the video encoder. Final we concatenated visual and speech features and feed them into the decoder.

3. RESULTS

3.1. data

To increase the amount of data and enhance the robustness of the model against different speaking styles, we applied 6-fold data augmentation, include 3-fold speed perturbation, volume perturbation, noise augmentation, reverberation augmentation.

3.2 Result

Table 2 shows our system result for the MISP challenge 2022 track 2. The baseline model CPCER result was 63.94% in eval set . ID5 shows our conformer audio encoder result , CPCER was 39.12%. , we used Wav2ver audio encoder instead of conformer audio encoder in ID6 and the CPCER reduced from 39.12% to 30.01%.

ID	MODEL	CPCER
ID4	baseline	63.94%
ID5	Conformer audio encoder	39.12%
ID6	Wav2vec audio encoder	30.01%

Table 2 speech recognition results for Track2

4. CONCLUSION

This technical report describes our submission to the MISP 2022 challenge. Our work includes speech enhancement front-end , training data augmenta tion, audio-visual speaker diarization and audio-visual speech recognition model. In audio-visual speech recognition model,We use Wav2vec as audio encoder. Wav2Vec encoder firstly pre-training in open-source corpora and misp dataset. Our system improves against the baseline with an absolute reduction of 33.93 % on the eval dataset and ranks 2nd in the misp2022 challenge track2.