System Description for T267 Team System

Shunfei Chen, Xinkang Xu, Xinhui Hu

Hithink RoyalFlush AI Research Institute, Zhejiang, China
chenshunfei@myhexin.com

Abstract

Our team's submission system for the CNVSRC2023 Challenge is presented in this technical report. The challenge focuses on continuous visual speech recognition (VSR) for specific speaker and non-specific speakers. Our system is composed of training data augmentation, modeling unit design, transformer and LSTM language model (LM) training, and model fusion with the ROVER tool. In comparison to the official baseline system for the Fixed Track in Single-speaker VSR (T1), our system achieved a 6.98% absolute Character Error Rate (CER) reduction on the evaluation set. Moreover, for the Fixed Track in Multi-speaker VSR (T2), our system obtained a 3.82% absolute CER reduction on the evaluation set.

1. Data

CNVSRC 2023 defines two tasks: Single-speaker VSR (T1) and Multi-speaker VSR (T2)[1]. For the T1 task and T2 task, each task has two defined tracks based on the data used for system development: Fixed Track and Open Track. For the Fixed Track, only the CNCVS dataset and the development set of T1 or T2 are allowed for training or fine-tuning the system. Therefore, In our system of the T1 fixed track, We only use the CNCVS corpus contains visual-speech data from over 2,557 speakers with more than 300 hours of data and CNVSRC-Single dev set includes audio and video data from a single speaker with over 94 hours of data. For the T2 fixed track, we only use the CNCVS corpus and CNVSRC-Multi dev set includes audio and video data from 43 speakers with over 29 hours.

To improve the robustness of the model, we use some video data augmentation methods[1] to process CNVSRC-Single dev and CNVSRC-Multi dev, such as adding salt and pepper noise, and video augmentation with Gaussian blur. On the other hand, because our system uses Chinese characters as the modeling unit, we use the text of the training set to construct the modeling unit.

2. Models

2.1. pre-precessing

To process video data, we use the methods in the baseline system to perform the following preprocessing operations: The first step is to perform face detection using the retinaface predictor model[2]. Following that, each frame is aligned to a referenced frame[3], commonly known as the mean face, to normalize rotation and size differences across frames. The final step in the pre-processing module is to crop the mouth region from the aligned mouth image. We crop the mouth region of interests (ROIs) using a bounding box of 96 × 96.

2.2. model structure

We do training or fine-tuning based on the officially provided pretrained model *model_avg_14_23_cncvs_4s.pth*, which is trained using a data set is shorter than 4s in CNCVS corpus. The model consists of a ResNet-based front-end module and is followed by a standard conformer based on a jointing CTC/Attention model. The Conformer encoder with 12 layers, 768 input dimensions, 3072 feed-forward dimensions with 12 attention heads, and the kernel size of convolution is 31. Then the decoder is a 6-layer Transformer with the same dimensions and number of heads as the encoder. We set the CTC weight to 0.1 in the training and decoding phases.

Our system is different from the baseline system in two main aspects. On the one hand, we use Chinese character modeling instead of bpe modeling in the baseline system. On the other hand, we have performed data augmentation on the dev set. The two methods will be proven effective in subsequent experimental results.

The model training is divided into two stages. The first stage is to use CNCVS corpus for course learning based on *model_avg_14_23_cncvs_4s.pth*. The second stage is to use *CNVSRC-Single dev* or *CNVSRC-Multi dev* corresponding augmentation data to continue fine-tuning training.

2.3. language model

In our system, a recurrent neural network(RNN) LM and transformer-based LM were trained by the Espnet[2] using the transcripts of CNCVS training set and CNVSRC-Single dev set or CNVSRC-Multi dev set. The RNN LM consists of 2 LSTM layers with 1024 dimensions. The transformer-based LM consists of a 16-layer encoder with 2048-dim feed-forward and 8 heads attention with 512 dimensions. The model unit of both models is character.

2.4. model fusion

To fuse different models, the fusion strategy ROVER[3] is used. It is a system developed at NIST to produce a composite automatic speech recognition(ASR) output when the outputs of multiple ASR systems are available.

3. Results

The results of our system are demonstrated in Table 1. Compared with the results of baseline and using character as a modeling unit, we can find that the character model unit can significantly improve the performance. The third row of Table 1 shows the CER when using video augmentation on *CNVSRC-Single dev* or *CNVSRC-Multi dev*, which achieves 1.29% and 2.04% absolute reduction in single-speaker dev and multi-speaker dev compared with the CER of the second row. Furthermore, based on the trained model with video augmentation, we use the trans-

[1] https://github.com/okankop/vidaug

[2] https://github.com/sectum1919/face_detection

[3] https://github.com/sectum1919/face_alignment

Table 1: The results of our system

method	training set(hours)	finetuning set(hours)		single-speaker dev CER	multi-speakers dev CER
		T1	T2		
baseline	287	83.7	18	48.57%	58.77%
char model unit	287	83.7	18	43.59%	56.77%
+video aug	287	166	36	42.30%	54.74%
+RNN LM				42.18%	
+transformer LM				42.16%	
+model fusion				**41.50%**	

former LM and the RNN LM to decode respectively, so that the recognition results are slightly improved. Finally, we used ROVER to fuse the results from the second row to the fifth row in Table 1, and obtain the final result in single-speaker dev, which achieved 7.05% absolute reduction compared with the result of baseline.

We also submitted our recognition results on the Eval set. The CER in *CNVSRC-Single eval* is 41.62% and The CER in *CNVSRC-Multi eval* is 54.55%.

4. References

[1] Chen Chen, Dong Wang, and Thomas Fang Zheng, "Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] Shinji Watanabe, Takaaki Hori, Shigeki Karita, and et al., "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[3] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 2002.