System Description for T244 Team System

Meng Miao , Feilong Bao , Guanglai Gao

College of Computer Science, Inner Mongolia University
National & Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
E-mail 32209013@mail.imu.edu.cn

Abstract

Visual Speech Recognition (VSR) is a technology that recognizes and understands speech by analyzing visible cues such as speaker's lip movements.In the case of using only the CN-CVS dataset, the multi-speaker visual speech recognition task trained with only visual information has some problems 1) it cannot recognize homophones 2) the individual differences in lip movements across speakers are too large. Therefore, we try to mitigate these problems by adding audio information along with speaker information during the training process.We use confomer as the backbone network for visual encoder and audio encoder, for speaker information we use a simple linear layer for encoding.Eventually, we reached a CER of 53.6815 on the *CNVSRC-Multi.Eval*.

1. Data

A complete description of the data profile used to model training. Specifically, for Task 1 Single-speaker VSR *fixed track*, ONLY *CN-CVS* and *CNVSRC-Single.Dev* are allowed to be used to perform system development. For Task 2 Multi-speaker VSR *fixed track*, ONLY *CN-CVS*[1] and *CNVSRC-Multi.Dev* are allowed to be used to perform system development. For Task 1 *open track* and Task 2 *open track*, ALL the used data sources except *CNVSRC-Single.Eval* and *CNVSRC-Multi.Eval* should be clearly stated in the system description. Again, all participants should *strictly* comply with the data rules mentioned in[1].

2. Models

2.1. Data preprocessing

For the visual stream,we extract the mouth region of interest (ROI) from each frame using a 96×96 bounding box. Subsequently, the frame is converted to grayscale and normalized by subtracting the mean and dividing by the standard deviation of the training set. For audio streams, we only perform normalization per utterance.In the realm of speaker information, a distinctive approach is employed: we leverage the folder name housing the video content, utilizing it as a unique identifier for the speaker. This identifier is then encoded into a one-hot representation. This innovative method not only simplifies the association of speakers with their respective content but also allows for an efficient encoding scheme that captures individual speaker distinctions. By extracting speaker IDs from folder names, our model becomes adept at recognizing and distinguishing between speakers, contributing to a more nuanced and contextually rich representation of speaker information. This nuanced representation is vital for tasks such as speaker recognition, fostering a system that is both adaptive and precise in handling a diverse range of speakers within various video contexts.

2.2. Model details

Our audio-only and visual-only Automatic Speech Recognition (ASR) models employ a module based on Conformer[2], pretrained on CNVSRC. This is followed by a Conformer encoder with 12 layers, featuring 768 input dimensions, 3,072 feed-forward dimensions, and 16 attention heads. Using Conformer as an extractor for audio and video allows for the unification of the model structure and simplifies overall system design and maintenance. This helps to reduce system complexity and improve scalability and maintainability.And it helps the model to integrate cross-modal information between different modes.The decoder is a 6-layer Transformer with identical dimensions and the same number of heads as the encoder.The speaker encoder is a crucial component in the system, and it is composed of a layer of Linear transformation. This linear layer is responsible for mapping the input features, which initially have 2382 dimensions, to a lower-dimensional space of 192 dimensions. This reduction in dimensionality is a key step in the process, as it helps in extracting essential speaker-related information while simultaneously reducing the computational complexity of subsequent operations. The resulting 192-dimensional representation serves as a more compact and informative embedding that encapsulates important speaker characteristics .The output of the audio encoder and the output of the visual encoder are fed to the cross-attention module for more coordinated cross-modal information fusion.After passing through the cross-attention module, its output is connected to the output of the speaker encoder, followed by the self-attention operation. This smooth and efficient process aims to achieve seamless integration of cross-modal information, and further enhance the model's ability to model complex context and speaker information by applying the self-attention mechanism after the connection.Note that once the speaker encoder has been trained, its parameters will be frozen and remain unchanged.model depicted in Figure 1.

2.3. Implementation details

For data augmentation,Horizontal flipping, random cropping, and adaptive time masking[3] are implemented on the visual in-

[1]http://aishell-cnceleb.oss-cn-hangzhou.aliyuncs.com/CNVSRC2023/CNVSRC_2023_Evaluation_Plan.pdf

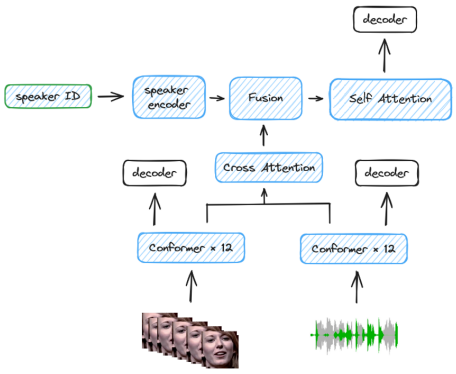

Figure 1: The overview framework of our model.

puts, whereas solely adaptive time masking is employed for the audio stream.For both streams, we opt for a quantity of masks that scales with the utterance length, along with a maximum masking duration of up to 0.4 seconds.For the target vocabulary, we use SentencePiece [4] subword units with a vocabulary size of 5 000.We pre-train the visual-encoder and audio-encoder on the CNVSRC dataset, and then train the two modules jointly.Regardless of the process of training in which module, we use AdamW [5] optimizer and train for 85 epoch.

During the training process, we used CTC Loss and Seq2Seq Loss to calculate six losses for video modality, audio modality and audiovisual modality simultaneously. This means that we considered the relationship between the speech and video modalities, as well as the overall consistency of the audiovisual modalities. Specifically, we train our model by jointly optimizing these six loss functions to ensure that good learning is achieved under multimodal inputs.However, in the inference process, we chose to utilize only the visual encoder and visual decoder. The purpose of this design may be to simplify the inference process and reduce the computational complexity, or it may be based on some a priori knowledge that makes it more efficient in practical applications.

3. Results

Our model achieved a Character Error Rate (CER) of 53.6815 in the CNVSRC-Multi.Eval. This notable performance underscores the effectiveness of our approach in accurately transcribing and recognizing speech in a multi-speaker setting.

We hold the belief that the Cross Attention module serves a crucial role in mitigating the semantic gap between visual and audio modalities. By incorporating speaker information, our aim is to optimize the reduction of variations in mouth shape across different individuals. This strategic inclusion enhances the overall synergy between visual and audio cues, fostering a more cohesive integration of information and contributing to improved performance in tasks requiring cross-modal understanding. The introduction of speaker information plays a key role in aligning visual and audio features, ultimately enhancing the model's ability to discern subtle differences in mouth shapes among speakers.

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] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[3] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual speech recognition for multiple languages in the wild," 02 2022.

[4] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao, Eds., Melbourne, Australia, July 2018, pp. 66–75, Association for Computational Linguistics.

[5] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.