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

This paper introduces a novel architecture to enhance lip reading of Mandarin Chinese, combining Conformer networks with a Structured State Space Decoder. Mandarin, a tonal language, benefits significantly from visual cues to disambiguate spoken content. Our approach leverages the strengths of Conformers for robust feature extraction across both audio and visual modalities and integrates a bidirectional decoder within a structured state space framework to efficiently capture long sequences of temporal dynamics and complex dependencies. This method builds on recent advances in audio-visual speech recognition, emphasizing the importance of multimodal integration for improved linguistic understanding. Developed for the Chinese Continuous Visual Speech Recognition Challenge (CNVSRC 2024), our architecture achieved a Character Error Rate (CER) of 47.9259 on the *CNVSRC-Multi.Eval* dataset and 39.3110 on the *CNVSRC-Single.Eval* dataset, demonstrating the effectiveness of our method in this challenging task.

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*[1] 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

In video processing, we use RGB channels and crop the lip region to a size of 96x96 pixels. During training, we further randomly crop it to 88x88 pixels to achieve data augmentation, enhancing the model's robustness and generalization ability. Additionally, we apply temporal masking techniques to both the audio and video streams, randomly masking parts of the time sequences to prevent model overfitting and improve its adaptability to various temporal variations.To further enrich the diversity of the dataset, we apply speed perturbation to the audio-visual streams of the CN-CVS, CNVSRC-Multi, and CNVSRC-Single datasets. Specifically, we perform 1.1x and 0.9x speed changes on these data, simulating different playback speeds. This speed perturbation not only increases the diversity of the training data but also helps the model better understand and handle various speed variations that may be encountered in real-world applications, thereby improving the model's performance in actual scenarios.

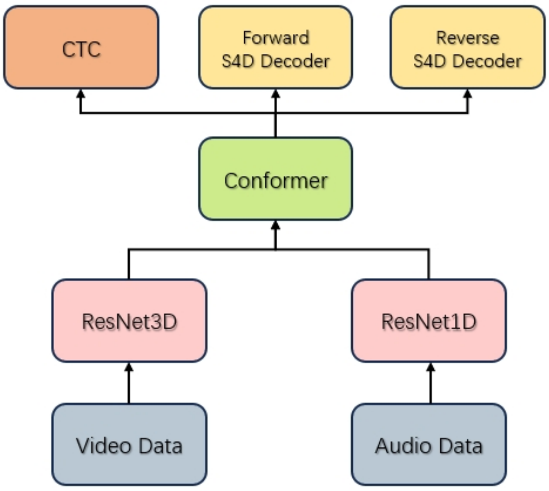

Figure 1: The overview framework of our model.

2.2. Model details

As shown in Figure 1, our multimodal lip recognition model extracts features from video and audio data through the ResNet3D and ResNet1D models, respectively. The video data are extracted with spatio-temporal features by ResNet3D to capture the dynamic changes of lip movements, while the audio data are extracted with temporal features by ResNet1D to obtain the spectral information of the audio signal. The processed audio and video features are thus fully encoded before entering the subsequent fusion module, ensuring the validity and integrity of the input features. We use the Conformer module to fuse the audio and video information, which can effectively fuse the audio and video features through the self-attention mechanism, and the Conformer module can capture both local and global

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

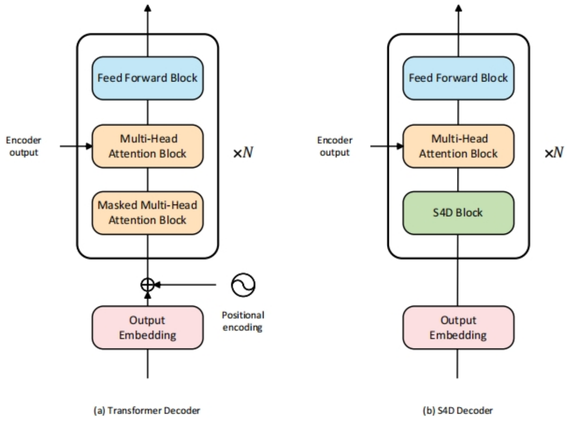

Figure 2: Structural Comparison Between S4D Decoder and Traditional Transformer Decoder.

information to enhance the richness and accuracy of the feature representation.

Given the excellent results that [2] has shown in speech recognition and speech synthesis tasks using the S4[3] model, we attempted to apply the state transfer space model to lip recognition tasks. Unlike previous work, we use the more memory-efficient S4D[4] model.Figure 2 presents a comparison between our proposed S4D Decoder and the traditional Transformer Decoder. The S4D Decoder replaces the masked multi-head attention mechanism with the S4D module, which interprets time series data bidirectionally, thereby augmenting the model's temporal modeling capabilities. Compared to the traditional Transformer Decoder, the S4D Decoder more effectively captures the intricate relationships between audio and video features, thereby enhancing the performance and accuracy of multimodal lip-reading tasks. This innovative design substantially improves the model's performance in multimodal applications, providing a distinct advantage in processing complex sequential data.

To ensure consistency between model outputs and the ground truth labels, we introduced CTC Loss and KLDivLoss during the decoding process. CTC Loss is primarily used to calculate the discrepancy between the forward decoder's outputs and the target sequence, thereby maximizing the sequence prediction probability and ensuring accurate generation of the target sequence. On the other hand, KLDivLoss is employed to measure the distributional difference between the outputs of the forward and reverse decoders. By minimizing the KL divergence between the forward and reverse decoder outputs, we can enhance the consistency of the decoder's output, ensuring that the model remains robust even when faced with noisy and variable data. Furthermore, the incorporation of KLDivLoss helps the decoder better learn dependencies between sequences during training, thereby improving the overall robustness and generalization ability of the model. The combined use of these loss functions significantly enhances the model's performance in multimodal lip reading tasks, ensuring excellent recognition accuracy even under complex and variable input conditions.

3. Results

Our architecture was specifically developed for the Chinese Continuous Visual Speech Recognition Challenge (CNVSRC 2024), and it has demonstrated outstanding performance in this competition. On the *CNVSRC-Multi.Eval* dataset, our model achieved a character error rate (CER) of 47.9259, while on the *CNVSRC-Single.Eval* dataset, it achieved a CER of 39.3110. These results clearly demonstrate the effectiveness of our approach in handling complex multimodal lip reading tasks. Especially when faced with real-world scenarios involving noise and diverse inputs, our model maintained a high recognition accuracy, showcasing its robustness and adaptability. Additionally, the low character error rate reflects the advantage of our architecture in accurately capturing and processing audio-visual features, further validating the exceptional performance of the S4D decoder in multimodal lip reading. These results highlight the practical application potential of our method, providing strong support for future research and development in related fields.

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] Koichi Miyazaki, Masato Murata, and Tomoki Koriyama, "Structured state space decoder for speech recognition and synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. 2023, pp. 1–5, IEEE.

[3] Albert Gu, Karan Goel, and Christopher Ré, "Efficiently modeling long sequences with structured state spaces," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022, OpenReview.net.

[4] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré, "On the parameterization and initialization of diagonal state space models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.