System Description for T405 Team System

Author Name

Affiliation and Department
fei.su@whu.edu.cn

Abstract

This is a desciption for T405 Team System of Task 2 Multi-speaker VSR Open Track. Our system is built upon a pre-trained frontend and Transformer-based backend architecture. We investigate the use of AV-Hubert as a front-end to extract visual features, and utilizing E-Branchformer as the backend encoder.Due to our late entry into the competition, our model adjustments were not fully optimized. Although we participated in the Open Track, we did not utilize the CN-CVS2-P1 dataset for training. Consequently, we achieved results slightly above the baseline, reaching a CER of 58.1251 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*[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* 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 88 × 88 bounding box. We use the same video-based data augmentation methods referred to the [2] , including speed perturbation, frame-wise rotation, horizontal flip, frame-level cropping, color jitters, gray scaling and histogram equalization.

2.2. Model details

2.2.1. Pretrained Frontend

AV-Hubert: AVHubert is a unsupervised audio-visual model, utilizes audio and visual data to predict cluster labels, achieving effective unsupervised training. AV-Hubert builds upon the Audio Hubert framework [3], which is a self-supervised framework for training audio-based models. Hubert's training involves two stages: feature clustering and masked prediction. Predicting cluster labels for masked regions allows the model to leverage unmasked areas to learn local representations and long-range temporal dependencies among latent features. Iteratively, these stages enhance both the quality of clustering and the feature representation capability.

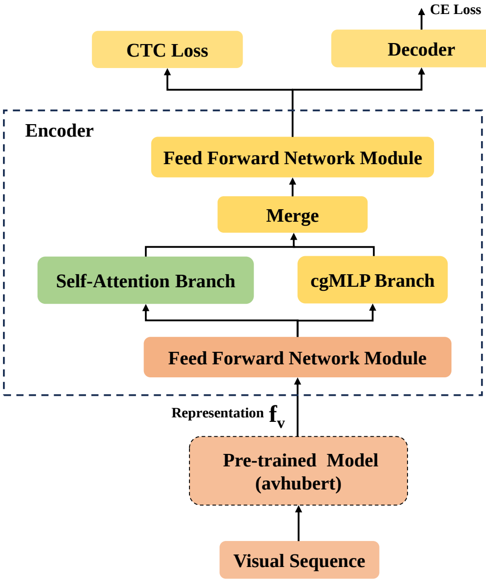
CE Loss
CTC Loss
Decoder
Encoder
Feed Forward Network Module
Merge
Self-Attention Branch
cgMLP Branch
Feed Forward Network Module
Representation $\mathbf{f}_v$
Pre-trained Model (avhubert)
Visual Sequence

Figure 1: The overview framework of our model.

2.2.2. Backend

E-Branchformer: We evaluate the performance of the E-Branchformer [4] architectures. The E-Branchformer is an enhanced variant of the Branchformer [5]. The Branchformer encoder consists of two parallel branches designed to capture a diverse range of contextual information. One branch employs self-attention mechanisms to capture long-range dependencies, while the other branch utilizes a multi-layer perceptron module with convolutional gating (cgMLP) to concurrently extract intricate local correlations. Moreover, [4] improves the Branchformer by incorporating a depth-wise convolution-based merging module and adding an additional pointwise feedforward module, leading to the development of the E-Branchformer.

As shown in Fig. 1,our model utilizes the avhubert for the frontend network and E-Branchformer blocks for the backend network. The frontend feature extraction network converts the 4-dimensional raw video frames (T, H, W, C) into 2-dimensional features (T, D) denoted as $X_{video} \in \mathbb{R}^{T\times D}$, representing T frames. This transformation is akin to the processing of speech-based features $X_{spec} \in \mathbb{R}^{T\times D}$ (such as MFCC, FBank, etc.). By converting the original video into a sequence of features, the network standardizes the input format, making

[1] http://aishell-cnceleb.oss-cn-hangzhou.aliyuncs.com/CNVSRC-2024/CNVSRC_2024_Evaluation_Plan.pdf

it suitable for the subsequent Transformer-based backend networks. We also incorporate Multi-layer Feature Aggregation by leveraging features from various layers of the pre-trained frontend, rather than relying exclusively on the output from the final layer. This approach is inspired by MFA-Conformer[6].

3. Results

Our model achieved a Character Error Rate (CER) of 58.1251 in the CNVSRC-Multi.Eval. We are confident that the designed VSR network has the potential to yield improved results. Due to the late entry into the competition, we were only able to use a subset of the data and conduct three rounds of training. Moving forward, we plan to further investigate VSR networks that leverage self-supervised pre-trained models for feature extraction in the frontend.

4. References

[1] Chen Chen, Dong Wang, and Thomas Fang Zheng, "Cncvs: 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] Haoxu Wang, Ming Cheng, Qiang Fu, and Ming Li, "The dku post-challenge audio-visual wake word spotting system for the 2021 misp challenge: Deep analysis," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*. IEEE, 2023, pp. 84–91.

[5] Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*. PMLR, 2022, pp. 17627–17643.

[6] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Proc. Interspeech*, 2022, pp. 306–310.