The VII-BUPT (T238) System for CNVSRC 2023

Zehua Liu, Xiaolou Li, Yiwei Sun, Zhenzhen Pan, Li Guo

Beijing University of Posts and Telecommunications

{liuzehua,lixiaolou}@bupt.edu.cn

Abstract

This report describes our system, developed for Task 1 Single-speaker VSR - Fixed track, in the first Chinese Continuous Visual Speech Recognition Challenge (CNVSRC 2023). Building upon the official baseline, we implemented two data augmentation strategies to expand the dataset. Subsequently, we introduced a multi-task training approach that incorporates character-phoneme multi-scale supervision, aimed at enhancing the VSR encoder's capacity. Following this, we integrated a Transformer-based language model for text error rectification. Finally, two representative systems are selected and fused to achieve the final result, achieving a Character Error Rate (CER) of 40.52%.

1. Data

In this section, we provide a detailed description of the data used in our VSR system.

For the Task 1 Single-speaker Fixed track, we used the CNVSRC-Single.Dev and CN-CVS [1] datasets to develop our system. To expand the data volume and enhance data variability, on one hand, we utilized facial images cropped from CNVSRC-Single.Dev videos and audio tracks from both the CN-CVS and CNVSRC-Single.Dev datasets, and fed into SadTalker [2] to generate videos; on the other hand, we applied speed perturbation to the CNVSRC-Single.Dev videos, altering their speeds to 0.9x and 1.1x of the original.

These data augmentation techniques ultimately produced approximately 372 hours of SadTalker-generated videos and about 200 hours of speed-perturbed videos. Including the original 100 hours of data from CNVSRC-Single.Dev, our training regimen thus encompassed a total of approximately 672 hours of video data.

Figure 1 illustrates our data augmentation process and Table 1 presents the data used in our experiments.

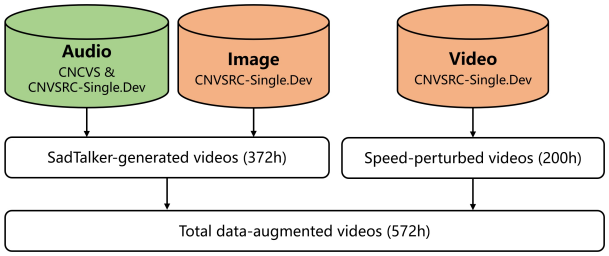

Figure 1: Illustration of our data augmentation process.

Table 1: The data used in our VSR system.

Type	Source	Hours
Original	CNVSRC-Single	100
SadTalker	CNVSRC-Single, CNCVS	372
SpeedPerturb	CNVSRC-Single	200
Total		672

2. Methods

Building upon the official baseline, we explored several methods to construct our VSR system.

2.1. Character-Phoneme multi-task training

Firstly, the Chinese language is characterized by its homophonic and polyphonic nature, wherein a single character may have multiple pronunciations, and different characters may share identical pronunciations.

To mitigate these linguistic ambiguities as well as bolster the model's capability in aligning lip movements with corresponding pronunciations, we extend the baseline character-level supervision by incorporating an auxiliary phoneme-level supervision, derived from Chinese Pinyin. This extension leads to a Character-Phoneme multi-task training framework.

In this framework, both the character sequence and phoneme sequence are simultaneously optimized through their respective classifiers. The classifier of both contains a linear layer followed by a CTC (Connectionist Temporal Classification) loss and a Transformer-based decoder with a CE loss. Figure.2 illustrates our proposed multi-task training framework.

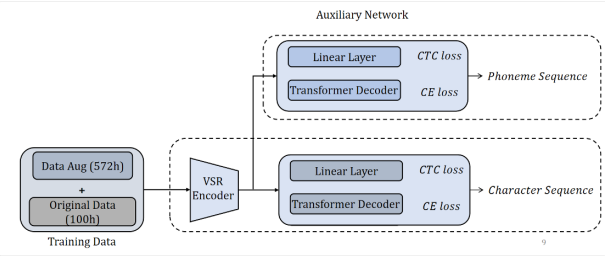

Figure 2: Our proposed character-phoneme multi-task training framework.

The loss function for this framework is formulated as follows:

$$L = a(mL_{\text{c-ctc}} + (1-m)L_{\text{c-att}}) + b(nL_{\text{p-ctc}} + (1-n)L_{\text{p-att}})$$

where $L_{\text{c-ctc}}$ and $L_{\text{p-ctc}}$ denote the CTC losses for the character and phoneme sequences, respectively, while $L_{\text{c-att}}$ and $L_{\text{p-att}}$ correspond to the CE losses for these sequences. The parameters a and b serve as hyper-parameters to balance the contribution between character-level and phoneme-level supervision. m and n are hyper-parameters to control the weights between the CTC and CE losses.

2.2. Transformer-based language model

Research has established that the integration of an additional language model (LM) can further enhance the performance of VSR systems. In this work, we introduce a Transformer-based LM [3] followed by [4]. Specifically, to train this LM, we amalgamated text data from both the CN-CVS and CNVSRC-Single.Dev datasets.

2.3. System fusion

Finally, to further improve system performance, a simple system fusion is implemented for the final decision-making process.

Specifically, two representative models are first selected from two different training phases. Then, in the Beam Search decoding phase, the character likelihoods from each model are averaged to determine the final character likelihoods. This straightforward fusion has proven to be effective in our experiments.

3. Settings

3.1. Multi-task training

For the multi-task training component, the official support model, pre-trained on CN-CVS, was adopted as the initial model. Then we followed the AdamW optimizer and learning rate scheduler with the baseline system to implement the training process. Finally, we carefully adjusted the values of the hyper-parameters to ensure the optimal balance among different loss components. In our experiments, the hyper-parameters $\{a, b, m, n\}$ were determined to be $\{0.2, 1.0, 0.5, 0.1\}$, respectively.

The training process in our experiments was structured into three phases:

Phase 1. In the first phase, we only used the training subset of CNVSRC-Single.Dev, complemented by the SadTalker-based data augmentation technique. This resulted in a total of 362 hours of video data, which was used to fine-tune the initial model. The initial learning rate was configured at 0.0011, accompanied by a warm-up scheduler extending over 2 epochs. The entirety of this phase encompassed 120 epochs of training.

Phase 2. In the second phase, we employed the entire (both the training and validation subsets) CNVSRC-Single.Dev dataset, and also involved the SadTalker-based data augmentation technique to augment the training data. Initialized by the model trained in Phase 1, we proceeded with an additional 42 epochs of training with the initial learning rate at 2e-5.

Phase 3. In the third phase, we added speed-perturbed videos based on *CNVSRC-Single.Dev*, and finally, there are 672 hours of video data in total. Initialized by the model trained in Phase 2, two additional epochs were trained with the initial learning rate at 3e-5.

3.2. Language model

We developed a Transformer-based Language Model (LM) with the neural backbone of 8 attention heads and 16 layers. Considering the limited volume of text data, we employed a warm-up strategy with StepLR to mitigate the risk of overfitting. Specifically, the learning rate was rapidly increased at the onset to expedite the model's convergence, and then followed by a gradual decrease to ensure stability in the training process. In addition, a dropout layer was adopted for each layer with a dropout rate of 0.1 to prevent overfitting.

3.3. System fusion

First, we conducted a simple average fusion of the models trained from Phase 2 and Phase 3 to obtain the character likelihoods. Then, during the Beam Search stage, we integrated the likelihoods predicted by the averaged acoustic model with those from the language model, applying a weight ratio of 1:0.1. Finally, the optimal search path determined was regarded as the final result.

4. Results

We submitted a total of four systems, which are delineated as follows:

- S1. The output of Phase 1 of multi-task training.
- S2. S1 with LM.
- S3. The output of Phase 2 of multi-task training with LM.
- S4. The fusion of S2 and S3.

Results are reported in Table 2 in terms of Character Error Rate (CER). Note that 'ST' denotes the SadTalker-based data augmentation; 'SP' denotes the speed-perturbation data augmentation; 'Phoneme' denotes the phoneme-level supervision; 'LM' denotes the Transformed-based language model.

Table 2: CER results of our four systems on Task 1 Single-speaker - Fixed Track.

	Data Aug		Phoneme	LM	CER %
	ST	SP			
Base	✗	✗	✗	✗	48.60
S1	✓	✗	✓	✗	41.62
S2	✓	✗	✓	✓	40.64
S3	✓	✓	✓	✓	40.72
S4	Fusion of S2 and S3				40.51

5. Conclusion

We developed a high-performance Single-speaker VSR system for the CNVSRC 2023 T1-Fixed track. Our results highlight several aspects: (1) Data augmentation techniques, including SadTalker and speed perturbation, proved to be significantly beneficial. (2) Involving an auxiliary phoneme-level supervision was found to enhance the VSR encoder's performance. (3) The integration of the language model was also identified as essential, further contributing to performance improvements.

6. 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] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang, "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," *arXiv preprint arXiv:2211.12194*, 2022.

[3] Kazuki Irie, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Language modeling with deep transformers," *arXiv preprint arXiv:1905.04226*, 2019.

[4] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.