SPEAKER-ADAPTIVE LIPREADING VIA SPATIO-TEMPORAL INFORMATION LEARNING

Yi He, Lei Yang, Hanyi Wang, Yun Zhu, Shilin Wang⋆, Senior Member, IEEE

Shanghai Jiao Tong University, China

ABSTRACT

Lipreading has been rapidly developed recently with the help of large-scale datasets and large models. Despite the significant progress made, the performance of lipreading models still falls short when dealing with unseen speakers. Therefore, it is necessary to utilize the speaker's videos for fine-tuning to obtain a speaker-adaptive model. However, this approach can result in high overheads, especially for full fine-tuning. To address this problem, we propose a novel parameter-efficient fine-tuning method based on spatio-temporal information learning. In our approach, a low-rank adaptation module which can influence global spatial features and a plug-and-play temporal adaptive weight learning module are designed in the front-end and back-end network, which can adapt to the speaker's unique features such as the shape of the lips and the style of speech, respectively. An Adapter module is added between them to further enhance the spatio-temporal learning. The final experiments on the LRW-ID and GRID datasets demonstrate that our method achieves state-of-the-art performance even with fewer parameters.

***Index Terms*—** Visual Speech Recognition, Speaker-Adaptive Lipreading, Parameter-Efficient Fine-Tuning

1. INTRODUCTION

Lipreading, also known as visual speech recognition, is a technology that recognizes speech content from the movements of a speaker's lip. It is useful in several real-world applications like assisting the hearing impaired, generating subtitles automatically and aiding audio speech recognition in noisy environments. This technique has developed rapidly in recent years by using larger models and more training data[1, 2, 3]. However, existing models degrade rapidly in performance when recognizing speakers who are absent in the training set. For instance, [3] recognize seen speakers' utterances with 18.0% Word Error Rate (WER), but 30.5% WER for unseen speakers. The variance in error rate can be attributed to the model's sensitivity towards personal characteristics of speakers, such as lip shape and talking style. This issue hinders practical application of the current lipreading model since it frequently encounter speakers who are not presented in the training dataset. To address this problem, a speaker-adaptive model is needed by fine-tuning with videos from the specific speaker. However, it is noteworthy that tuning the entire model not only introduces a significant computational burden due to the need to adjust all parameters, but also results in model forgetting the knowledge learned from extensive data. Hence, parameter-efficient speaker-adaptive fine-tuning methods are required.

A small amount of work[4, 5] has therefore begun to fine-tune lipreading networks to obtain speaker-specific models with as few training parameters as possible. For example, Kim et.al. [4] trained the convolutional layer's padding region in the front-end network to adapt to the speaker. It is important to note that the padding regions are situated at the edges of the feature map, and the changes of padding have to pass through several layers to impact the central area of the feature map. However, the most important part for lipreading task is the human lip situated in the center[6]. Therefore, this part requires more attention to appropriately conform to the speaker's distinctive lip shape and appearance. Recently, [5] extended [4] by injecting additional information into the back-end network through the prompt tuning technique[7]. However, this approach can only be applied to Transformer-based[8] models and can not be applied to other temporal models, such as the Temporal Convolutional Network (TCN)[9], which is the most commonly used network in word-level lipreading[10, 11].

It has been observed that the primary distinction between the speech videos of different speakers comes from two aspects: the spatial dimension features and the temporal dimension features. The former focuses on physiological characteristics such as the shape and appearance of the lips, and the latter focuses on individual talking habits including the duration of pronunciation and the linking sound[12]. Therefore, to fully adapt to the speakers' characteristics, it is necessary for the model to learn spatio-temporal personal patterns from the videos.

Considering the limitations of existing methods and the above observation, a novel speaker-adaptive network for lipreading is proposed. For the front-end network, the convolution-based low-rank adaptation is utilized to reduce the number of parameters during tuning while affecting the global features from shallow to deep level. In addition, a plug-and-play module is devised in the back-end temporal fusion

Shilin Wang⋆ is the corresponding author. This work was supported by the National Natural Science Foundation of China (62271307).

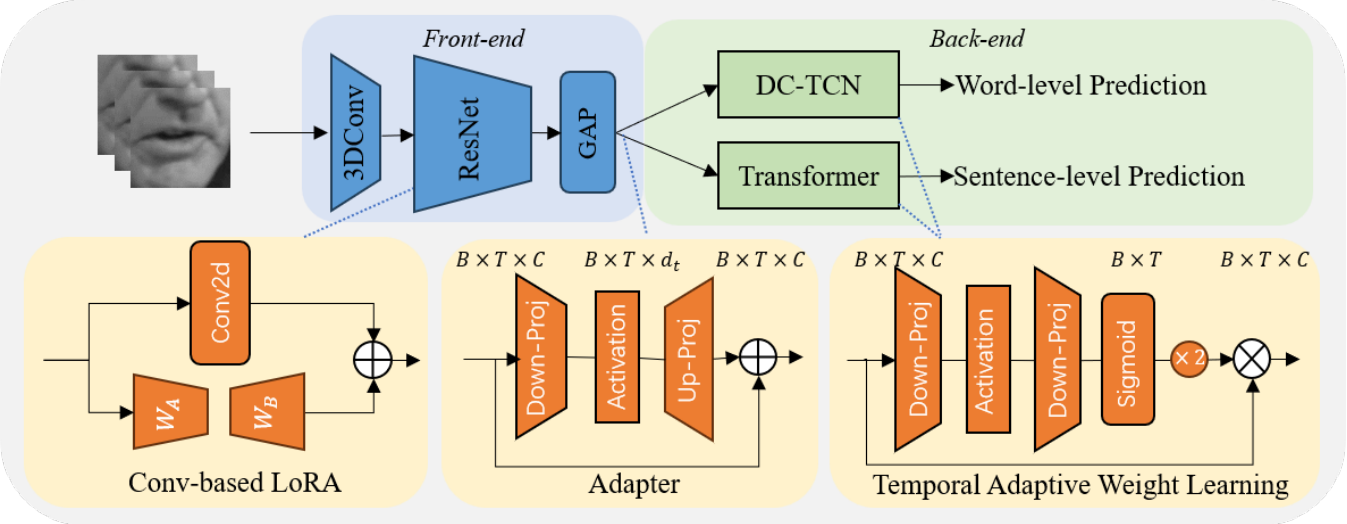

Fig. 1. Architectures of the baseline model and three kinds of parameter efficient fine-tuning methods. GAP means global average pooling.

network that modifies the feature amplitudes in each frame to adapt to the talking habits of a specific speaker. The spatio-temporal adaptation is enhanced through combining the two aspects with an Adapter module. Our approach enables the proposed network to integrate both the speaker's static physiological properties and their dynamic talking habits, therefore achieving speaker adaptation. Our contributions are mainly the following threefold: 1) A novel speaker-adaptive network for lipreading is proposed with three different adaption modules that learn the unique spatio-temporal features of the speaker; 2) A novel parameter-efficient fine-tuning module is proposed, which can be easily plugged in various back-end temporal fusion networks; 3) Extensive experimental results on the word-level LRW-ID dataset and the sentence-level GRID dataset demonstrate that the proposed network can achieve state-of-the-art performance on speaker-adaptive lipreading with fewer training parameters.

2. THE PROPOSED METHOD

The overall architecture of the proposed method is give in Fig.1. The network can be divided into a front-end feature extraction stage and a back-end temporal fusion stage. The sequence of lip region frames extracted from the speech video is the input to the model. The front-end network extracts visual feature sequence and the back-end transcribes it to recognition results. In this section, the detail of the baseline models and the proposed adaptation structures will be introduced.

2.1. Baseline Architecture

Lipreading tasks can be divided into two categories: word-level and sentence-level. The objective of word-level lipreading is to recognize isolated words, and a multi-classification network is typically used for the back-end. Sentence-level lipreading, on the other hand, necessitates the prediction of entire sentences. Therefore, temporal models such as Transformers are utilized to achieve sequence-to-sequence prediction.

In the proposed method, the commonly used front-end network in lipreading tasks is adopted, i.e. a single 3D convolution layer and a global average pooling layer following a ResNet[13] network, to extract visual features. For word-level lipreading, the state-of-the-art model, Densely-Connected Temporal Convolutional Network (DC-TCN)[14], is adopted as the front-end network. For sentence-level lipreading, an improved transformer-based network[15], which is optimized for lipreading tasks, is employed as a baseline.

2.2. Conv-based Low-rank Adaptation Module

Low-rank adaptation (LoRA)[16] is a currently widely used technique for fine-tuning large language models in a computationally efficient manner, which inject trainable layers (low-rank decomposition matrices) into projection layer in the Transformer's multi-head attention sub-layer. Inspired by LoRA, we introduce decomposition matrices into the 2D convolutional layer to adapt to the speaker's distinct space features. Specifically, the pre-trained weight matrix of the convolutional layer is denoted as $W \in \mathbb{R}^{C_{in} \times C_{out} \times k \times k}$, where C_{in}, C_{out} and k represent input channel, output channel and kernel size, respectively. The update of this matrix is constrained as $W + \Delta W = W + W_B W_A$, where $W_B \in \mathbb{R}^{C_{out} k \times rk}$ and $W_A \in \mathbb{R}^{rk \times C_{in} k}$ with the rank $r \ll min(C_{in}, C_{out})$. Note that W_A and W_B are trainable parameters and W is frozen

during tuning. For a specific input x, the output of convolutional layer $h = Wx + b$ can be modified as:

$$h = (W + s \cdot \Delta W)x + b = Wx + b + s \cdot W_B W_A x$$

where b is the bias value and s is the scaling hyperparameter. Note that this is a simplified representation that omits details unrelated to low-rank decomposition like filter sliding. Matrix W_B is initialized with a random Gaussian distribution and W_A is initialized with 0, so $W_B W_A = 0$ is guaranteed at the beginning of the tuning. By employing LoRA in the convolutional layer, the whole feature map can be affected and the speaker's lip shape and appearance can thus be considered during tuning.

2.3. Temporal Adaptive Weight Learning Module

In order to adapt to speaker's unique temporal characteristics while uttering, the temporal adaptive weight learning module (TAWL module) is designed in the back-end network. The input feature of TAWL module can be denoted as $f \in \mathbb{R}^{B \times T \times C}$, where B, T, C represent batch size, temporal length and channel, respectively.

The temporal weight is generated through two down-projection layers and non-linear activation. The first down-projection layer projects the feature to a low-dimensional space (with channel $d \ll C$). Following a nonlinear activation layer, the second down-projection layer is deployed to further project the feature to one dimension. As a result, the output feature is in the shape of $B \times T$. We then apply a sigmoid function element-wisely and multiply the result by 2, thereby obtaining weight values ranging from 0 to 2. These values represent the adaptive amplitude of the feature at each time step. The input feature f is finally multiplied by the weights to get the modified temporal feature. The weight matrices of the projection layer are initialized with a random Gaussian distribution having 0 mean and very small variance, resulting in the temporal weights being close to 1 at the start of training. By modifying the amplitude of the features in each frame, the model is able to learn the speaker's characteristics in temporal dimension, and therefore is better able to adapt to different talking styles.

2.4. Spatio-temporal Transition Module

In order to allow the back-end network to better adapt to the modification of features by the front-end adaptation network, we add a parameter-efficient spatio-temporal transition module before the features are fed into the back-end. This module can be regarded as a bridge connecting the adaptive knowledge learnt from the front and back ends. Specifically, we use a single Adapter[17] module, which firstly down-project the input dimension C to a low dimension d_t. The reduce rate l is defined as C/d_t. Then followed by a nonlinear activation function, the dimension is restored to the input dimension C by an up-projection layer.

3. EXPERIMENTS

3.1. Datasets

LRW-ID dataset[4] and GRID dataset[18] are used to evaluate the performance of speaker-adaptive lipreading. LRW-ID is a word-level lipreading dataset based on the LRW dataset[19], labelled with identity information. It consists of 500 English word classes and 17,850 speakers. 20 speakers (each with at least 900 videos) are selected to test the adaptation performance, and the others are used for training baseline models. GRID is a sentence-level lipreading dataset that includes 33 speakers, each speaking 1000 sentences. User 1, 2, 20, and 22 are used for performance evaluation and the rest for training baseline models.

3.2. Implementation Detail

Video frames of LRW-ID dataset and GRID dataset are cropped to 96×96 and 50×100, respectively. Random horizontal flip and time masking[14] are both employed for data augment. For back-end network in word-level baseline model, the DC-TCN structure is followed the setting in [14]. For sentence-level model, 4-layer improved Transformer[15] with embedding size of 256 is used. LoRA is applied to the second convolutional layer of each residual block in ResNet and TAWL modules are inserted between every encoding layers in back-end network. The rank r in conv-based LoRA is set to 2 and scaling parameter s is set to 16 empirically. The channel d in the TAWL module is 8 and the reduce rate l in the Adapter module is 32 .

3.3. Comparison with the State-of-the-Art

As in the previous approach[4, 5], for each speaker, we train the speaker-adaptive models with the baseline model using 1-minute, 3-minute and 5-minute video data, respectively. Note that the parameters of baseline models are frozen while fine-tuning. Regardless of the amount of training data, the same videos of the speaker are selected as the test set. Experiments were conducted on randomly selected data for five times and the average results are reported. Other methods are fine-tuned on the same baseline model, for a fair comparison.

In Table1, the speaker-adaptive performance tested on LRW-ID dataset is given. The accuracy of word prediction (ACC) is used as the metric and the number of parameters for fine-tuning is counted. It can be observed that our method outperforms the previous method when different amounts of training data are used and with fewer parameters. In Table 2, three more Transformer-based parameter-efficient fine-tuning methods, i.e. Adapter applied in Transformer[17], prompt tuning[5] and MLP-based LoRA[16], are further added for comparison on GRID dataset. WER is adopted to evaluate the performance of sentence prediction. The results illustrate that our model still outperforms all the other methods, which

Model	Params(M)	1min	3min	5min
Baseline	0	88.48	88.48	88.48
Padding[4]	0.108	89.41	90.34	90.72
Full Fine-tune	52.55	89.14	90.10	91.13
Proposed Method	0.099	**89.53**	**90.53**	**91.17**

Table 1. Speaker-adaptive performance (ACC in %) on LRW-ID dataset

Model	Params(M)	1min	3min	5min
Baseline	0	9.73	9.73	9.73
Padding[4]	0.053	6.54	4.91	4.45
Adapter in Transformer[17]	0.037	6.48	4.70	3.81
Prompt Tuning[5]	0.069	6.30	4.52	3.80
MLP-based LoRA[16]	0.051	5.14	3.64	2.90
Full Fine-tune	12.17	4.95	3.40	**2.51**
Proposed Method	0.035	**4.90**	**3.31**	2.73

Table 2. Speaker-adaptive performance (WER in %) on GRID dataset

proves the effectiveness of our spatio-temporal information learning approach.

On the other hand, when comparing with full fine-tuning, it can be observed that the proposed method maintains better performance when using 1-minute or 3-minute training data both on two datasets. This is because by training a small number of parameters, our model can not only adapt to the speaker's unique characteristics, but also retain the knowledge in the baseline model learnt from large-scale data, and is therefore more generalizable when using a small amount of training data. Although full fine-tuning can outperform our method on GRID when using more adaptive data, such as 5-minute videos, a large number of parameters need to be trained thus causing much more training overhead.

3.4. Ablation Study

To assess the effectiveness of each adaptation module, we conducted ablation studies on GRID dataset with 5 minutes of training data. It's shown in Table3 that increasing the trainable parameters, such as increasing the rank r of the LoRA module or the down-projection dimension d of the TAWL module, results in smaller performance gains. Consequently, we opted for r=2 and d=8 based on a balance between lipreading performance and number of parameters. It is also observed that the LoRA module provides a larger enhancement to adaptive capability when applied alone, whereas the TAWL module provides a comparatively smaller boost due to the greater ease of learning the shape and appearance of the lips compared to high-dimensional dynamic features like talking habits, which are more difficult to capture with a small amount of training data. When combined, the WER can be further reduced. Especially when

Model	Params(M)	WER
Baseline	0	9.73
Conv-based LoRA (r=2)	0.021	3.10
Conv-based LoRA (r=4)	0.042	3.07
TAWL (d=8)	0.010	7.89
TAWL (d=16)	0.019	7.50
Conv-based LoRA (r=2) + TAWL (d=8)	0.031	3.00
Conv-based LoRA (r=2) + TAWL (d=8) + Adapter (l=32) (Proposed Method)	0.035	**2.73**

Table 3. Ablation study on GRID dataset

Feature	Baseline Model	Adaptive Model
Front-end	0.892	0.922
Back-end	0.826	0.853

Table 4. Identity classification accuracy using front-end and back-end features

the Adapter is added, the error rate significantly decreases to as low as 2.73 despite that parameters only increase 0.004M. This indicates that the spatio-temporal adaptation is enhanced with the help of the Adapter module. Thus, through the above experiments, we demonstrate the effectiveness of each of the designed modules.

3.5. Effectiveness of Spatio-Temporal Information Learning

To confirm successful injection of the speaker's personal information via the adaptation modules, 10 sentences are randomly chosen from each test speaker's corpus in the GRID dataset to obtain the front-end output features and back-end output features of the adaptive model. Afterwards, these features are used to train an identity classifier consisting of one linear layer, and then this classifier is used to classify the identity of the remaining sentences. As in Table4, after adaptive training, the accuracy of both the front-end and back-end features for identity classification is improved. This suggests that the model has learnt spatio-temporal features containing the speaker's identity information.

4. CONCLUSION

In this paper, we propose a novel speaker-adaptive lipreading model. For front-end network, conv-based LoRA modules are used to adapt to speaker's space features. For back-end network, a plug-and-play TAWL module is designed to learn temporal characteristics. An Adapter module is finally employed to bridge the adaptation knowledge from front-end and back-end. The experiments show that the proposed method achieves the state-of-the-art performance on both word-level and sentence-level dataset with fewer training parameters.

5. REFERENCES

[1] Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al., "Synthvsr: Scaling up visual speech recognition with synthetic supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18806–18815.

[2] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[4] Minsu Kim, Hyunjun Kim, and Yong Man Ro, "Speaker-adaptive lip reading with user-dependent padding," in *European Conference on Computer Vision*. Springer, 2022, pp. 576–593.

[5] Minsu Kim, Hyung-Il Kim, and Yong Man Ro, "Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition," *arXiv preprint arXiv:2302.08102*, 2023.

[6] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 356–363.

[7] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[10] Weidong Tian, Housen Zhang, Chen Peng, and Zhong-Qiu Zhao, "Lipreading model based on whole-part collaborative learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2425–2429.

[11] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2857–2866.

[12] Shi-Lin Wang and Alan Wee-Chung Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328–3335, 2012.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, and Maja Pantic, "Training strategies for improved lip-reading," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8472–8476.

[15] Xingxuan Zhang, Feng Cheng, and Shilin Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proceedings of the IEEE/CVF International conference on Computer Vision*, 2019, pp. 713–722.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[18] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer, 2017, pp. 87–103.