Human Factors and Systems Interaction, Vol. 192, 2025, 1–9
https://doi.org/10.54941/ahfe1006622

A Lip-Reading Recognition System Based on SimAM and TCN

Yi Liu and Yuanyao Lu

School of Information Science and Technology, North China University of Technology, Beijing, 100144, China

ABSTRACT

Lip-reading recognition is a technology that converts the visual information of a speaker’s lip movements into corresponding textual content. It has broad applications in fields such as national defense, healthcare, and public safety, and holds significant academic value. In recent years, with the rapid advancement of deep learning, lip-reading technology has made notable progress, achieving numerous innovative and breakthrough results. This paper proposes a novel lip-reading recognition architecture that integrates a Residual Network (ResNet) with a Temporal Convolutional Network (TCN), and introduces a simple yet highly effective attention mechanism—Simple Attention Module (SimAM). The key components of the proposed approach are as follows: (1) Feature Extraction: ResNet is employed to extract spatial features from lip images. By introducing residual connections into conventional convolutional neural networks, ResNet effectively alleviates information loss and mitigates the vanishing gradient problem, allowing for more efficient utilization of deep-layer features. (2) SimAM: Traditional attention mechanisms often focus on enhancing features along either the spatial or channel dimension, limiting their ability to learn complex, multi-dimensional attention weights, and typically incurring high computational costs. To address these limitations, SimAM is incorporated. It leverages a spatial suppression mechanism to compute attention weights for each neuron, requiring no additional parameters, while simultaneously attending to both spatial and channel dimensions. (3) Temporal Modeling: TCN is adopted for sequence modeling, applying convolutional operations along the temporal axis. Unlike recurrent networks, TCN enables parallel computation, captures long-range dependencies effectively, and offers a simpler architecture with faster training and greater stability—particularly well-suited for large-scale lip-reading datasets. To validate the effectiveness of the proposed model, experiments were conducted on the largest publicly available lip-reading dataset, LRW, which features diverse pronunciation scenarios and a large number of samples. Comparative experiments with various state-of-the-art architectures demonstrate that the proposed model achieves significant improvements in both recognition accuracy and computational efficiency.

Keywords: Lip reading, Simple attention module, Temporal convolutional networks

INTRODUCTION

Human-Computer Interaction (HCI) (Card et al., 1983) is one of the most cutting-edge research fields in computer science. It refers to the process of information exchange between humans and computers through a specific interaction language and method to accomplish defined tasks. With the advancement of various innovative technologies, HCI has evolved from humans adapting to computers to computers adapting to humans, with speech interaction technologies being the most mature. In quiet environments, Audio Speech Recognition (ASR) (Jelinek, 1976) has achieved remarkable performance, reaching accuracy rates of 95% or higher. However, in acoustically complex environments, such as crowded public spaces or negotiation meetings with intense discussions, background noise and overlapping conversations significantly interfere with audio input, reducing recognition accuracy and failing to meet desired outcomes. This makes the effective utilization of visual speech information increasingly critical. Automatic Lip Reading (ALR), also known as Visual Speech Recognition (VSR) (Petajan, 1984), involves extracting visual information from a speaker's lip movements and converting it into human-readable text. With the emergence of deep learning and advancements in traditional methods, lip reading technology has made significant strides and demonstrated immense potential for further development.

The development of lip-reading recognition technology has gone through several important stages, with the proposal of various networks and models. In the early stages, researchers primarily used Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) to model the relationship between lip movements and speech. With the rise of deep learning, Convolutional Neural Networks (CNN) became widely used for feature extraction. Yue et al. (2015) proposed an end-to-end lip-reading recognition system based on deep convolutional networks, which significantly advanced the field. To better handle temporal sequence features, Graves et al. (2013) introduced Long Short-Term Memory (LSTM) networks and applied them to lip reading recognition. Subsequently, Tran et al. (2015) proposed 3D Convolutional Neural Networks (3D-CNN), which capture both spatial and temporal information when processing video sequences, improving the model's ability to model lip movement variations. In recent years, the Transformer architecture proposed by Vaswani et al. (2017), with its self-attention mechanism, effectively captures long-range dependencies. Zhou et al. (2020) applied this architecture to lip reading recognition, further improving recognition accuracy and robustness. In the domain of multimodal learning, Duan et al. (2019) introduced a deep neural network architecture that fuses visual information with audio signals. By combining CNN, LSTM, and acoustic models, they enhanced the performance of lip-reading recognition in complex environments. The introduction of these methods has significantly advanced lip-reading technology, leading to substantial breakthroughs in both accuracy and application domains. In summary, our main contributions are:

- We introduce a lightweight Simple Attention Module to compute attention weights, accelerating the weight calculation process and enhancing the flexibility of computing attention weights across both channel and spatial dimensions.

- We compare and evaluate the proposed model against several well-known networks, achieving further improvements in accuracy.

ARCHITECTURE DESIGN

The lip-reading recognition system architecture in this paper primarily consists of an input module, a feature extraction module, and an output module. The input module processes video or image sequences. The feature extraction approach is divided into image feature extraction and temporal feature modeling. Image feature extraction mainly utilizes Convolutional Neural Networks (CNNs) (LeCun et al., 1998), along with popular models such as ResNet (He et al., 2016), VGG (Simonyan and Zisserman, 2015), MobileNet (Howard et al., 2017), and ShuffleNet (Zhang et al., 2018). Temporal feature modeling primarily employs Recurrent Neural Networks (RNN) (Bai et al., 2018), Long Short-Term Memory Networks (LSTM) (Li et al., 2015), Gated Recurrent Units (GRU) (Chung et al., 2014), and Temporal Convolutional Networks (TCN) (Bai et al., 2018). In this work, we propose a lip-reading recognition system that combines Residual Networks (ResNet) with Temporal Convolutional Networks (TCN), while also incorporating the Simple Attention Module (SimAM) (Yang et al., 2021). The overall architecture is shown in Figure 1.

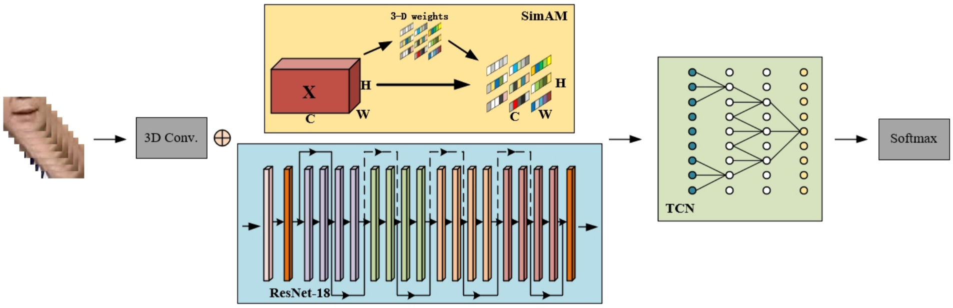

Figure 1: The architecture of a lip-reading recognition system.

Image feature extraction in this system is performed using ResNet, whose core innovation lies in the introduction of residual modules. By incorporating "skip connections" (identity mapping) that directly link inputs to outputs, ResNet effectively mitigates issues such as vanishing or exploding gradients in deep neural networks. This design allows for stacking significantly deeper layers, such as ResNet-18 and ResNet-50. The structure of ResNet is composed of multiple residual blocks, each containing convolutional layers, batch normalization layers, and ReLU activation functions, with the skip connections preserving the input features. ResNet has been widely applied in tasks like image classification, object detection, and semantic segmentation, profoundly influencing the design of deep learning models and becoming a foundational model for modern deep networks.

SimAM is a lightweight attention mechanism module whose core idea is to model the saliency of neuron activations, simulating the response characteristics of activation functions in neuroscience. This approach enhances the representational power of Convolutional Neural Networks (CNNs) in a simple, efficient, and parameter-free manner. In visual neuroscience, the neurons that carry the most information are typically those that exhibit distinct firing patterns compared to surrounding neurons. These active neurons are also capable of suppressing the activity of neighboring neurons, giving them higher priority. To identify these neurons, the following energy function is defined for each neuron:

$$e_t\left(\omega_t, b_t, y, x_i\right) = (y_t - \widehat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \widehat{x}_i)^2 \tag{1}$$

here,$\widehat{t} = \omega_t t + b_t, \widehat{x}_i = \omega_t x_i + b_t$, where t and x_i represent the target neuron and other neurons within a single channel of the input features $X \in R^{C\times H\times W}$. $M = H \times W$ denotes the number of neurons in this channel. The objective is to minimize the energy function when $\widehat{t} = y_t$. Minimizing this equation is equivalent to finding the linear separability between the target neuron t and the other neurons within the same channel. The labels for y_t and y_0 are binary (e.g., 1 and -1), and regularization is applied to obtain:

$$e_t\left(\omega_t, b_t, y, x_i\right) = (1 - (\omega_t t + b_t))^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (\omega_t x_i + b_t))^2 + \lambda\omega_t^2, \tag{2}$$

Theoretically, each channel has M energy functions, and the iterative computation can be computationally expensive. Therefore, a fast closed-form solution for ω_t and b_t is introduced:

$$\omega_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda}, \tag{3}$$

$$b_t = -\frac{1}{2}(t + \mu_t)\omega_t, \tag{4}$$

here, $\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$, $\sigma_t^2 = \frac{1}{M-1}\sum_{i}^{M-1}(x_i - \mu_t)^2$. Each channel follows the same distribution, avoiding the need for iterative computation of μ and σ at each position. Therefore, the minimum energy function can be computed as follows:

$$e_t^* = \frac{4(\widehat{\sigma}^2 + \lambda)}{(t-\widehat{\mu})^2 + 2\widehat{\sigma}^2 + 2\lambda}, \tag{5}$$

here, $\widehat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\widehat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \widehat{\mu})^2$, Equation (5) indicates that the smaller e_t^* is, the greater the distinction between neuron t and its surrounding neurons, making it more important for visual processing. Therefore, the importance of each neuron can be calculated as $\frac{1}{e_t^*}$. Based on the principle that attention modulation typically manifests as scaling the neuron responses, a scaling operator is used to refine the features:

$$\widetilde{X} = sigmoid(\frac{1}{E}) \odot X, \tag{6}$$

here, E represents the aggregation of e_t^* across all channels and spatial dimensions, and the sigmoid function ensures that the output is confined to the range [0, 1], without affecting the relative importance of each neuron.

The structure of SimAM is based on calculating attention scores for each pixel within a channel. It uses a simple mathematical formula to measure the distinguishability between pixels and the background, thereby enhancing key features while suppressing irrelevant information. Compared to feature refinement that focuses only on the channel or spatial dimensions, SimAM improves the flexibility of computing attention weights that span both channel and spatial variations. Since SimAM does not require additional learnable parameters, it is easy to integrate into various network architectures and has achieved notable performance improvements in tasks such as image classification and object detection.

Temporal feature modeling in this system is achieved using TCN, a convolutional neural network architecture designed for processing sequential data. The core idea of TCN combines causal convolution and dilated convolution. Causal convolution ensures the temporal causality of the sequence, while dilated convolution expands the receptive field, enabling the effective capture of long-term dependencies. Compared to traditional RNN-based approaches (such as LSTM and GRU), TCN offers advantages such as higher parallel computation efficiency and more stable gradients. Its architecture consists of multiple stacked convolutional layers, integrated with residual connections to mitigate gradient vanishing issues, along with pooling and normalization operations. TCN has demonstrated exceptional performance in tasks such as time series forecasting, speech recognition, and natural language processing, making it a powerful alternative in the field of sequence modeling.

EXPERIMENTAL ANALYSIS

In this study, we use the LRW dataset (Chung et al., 2017). The LRW dataset was introduced by the Visual Geometry Group at the University of Oxford in 2016 and is sourced from BBC broadcast programs rather than recordings made by volunteers or experimental subjects. It contains speech from hundreds of different speakers, consisting of 500 distinct words. All videos are 29 frames in length (1.16 seconds), with the word appearing in the middle of the video. The dataset includes over 550,000 speech instances, fulfilling the data volume requirements for deep learning to some extent.

Table 1: The Top-1 and Top-5 accuracy rates for ResNet-18 combined with various attention modules.

Method	Top-1 Accuracy(%)	Top-5 Accuracy(%)
ResNet-18	70.33%	89.58%
ResNet-18+SE	71.19%	90.21%
ResNet-18+CBAM	71.24%	90.04%
ResNet-18+ECA	70.71%	89.85%
ResNet-18+**SimAM**	71.31%	89.88%

During the preprocessing phase, we use the Dlib library's face detector and the 68-point landmark model to identify the lip region. Landmark information is used to extract facial key points, such as those of the lips and eyes, from each frame, returning the (x, y) coordinates for each key point. Missing frames in the landmark data are interpolated, and after processing the video frame by frame, the key points for all frames are saved. Based on the key point data, several feature points around the mouth (e.g., points 48 to 67) are selected to define the boundaries of the lips. The cropped lip region is then normalized to a fixed size of 96×96 and converted to grayscale. This process results in a standardized Region of Interest (ROI) for the lips. Before training, we also combine ResNet-18 with SimAM and other attention modules for comparison, and calculate the Top-1 and Top-5 accuracy on ImageNet-1000, as shown in Table 1.

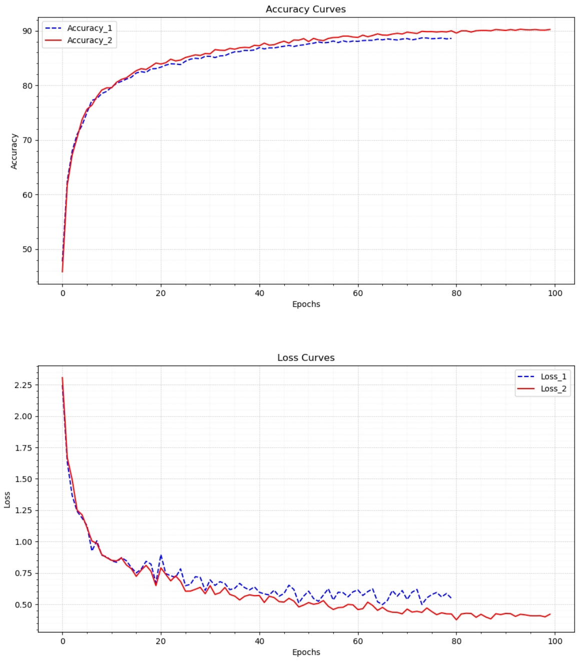

Figure 2: Loss and accuracy curves.

The training phase was conducted on a laboratory server equipped with a GeForce RTX 3090 Ti. The AdamW optimizer was used, with a batch size of 96 and 100 epochs. The initial learning rate was set to 0.0003. The entire training process was based on the LRW dataset, followed by final validation. The resulting loss curve and accuracy curve, as shown in the figure below, were compared with the scenario where the SimAM module was not added.

Figure 2 shows the loss curves before and after adding the attention module on the LRW dataset and displays the accuracy curves under the same conditions. A comparison reveals a significant reduction in loss and an increase in accuracy, with the loss decreasing by 0.12 and accuracy improving by 2%. Additionally, it is clear that increasing the number of epochs has a significant impact on the results.

Table 2: Compare with advanced models on LRW datasets. The author, architecture composition and accuracy are indicated respectively.

Authors	Method	Accuracy(%)
Chung et al.	CNN	61.10
Chung et al.	CNN+LSTM+attention	76.20
Stafylakis et al.	3D-CNN+ResNet-34+Bi-LSTM	83.00
Petridis et al.	3D-CNN+ResNet-34+Bi-GRU	83.39
Courtney et al.	Res-Bi-Conv-LSTM	85.20
Weng et al.	3D-CNN+Bi-LSTM	84.11
Wang et al.	3D-CNN+Bi-Conv-LSTM	83.34
Zhao et al.	3D-CNN+ResNet-18+Bi-GRU	84.41
Xu et al.	3D-CNN+P3D-ResNet50+TCN	84.80
Martinez et al.	3D-CNN+ResNet-18+MS-TCN	85.30
Kim et al.	3D-CNN+ResNet-18+Bi-GRU+VAM	85.40
Ma et al.	3D-CNN+ResNet-18+MS-TCN+KD	88.50
Kim et al.	3D-CNN+ResNet-18+MS-TCN+MH-VAM	88.50
Koumoaroulis et al.	3D-CNN+EfficientNetV2 +Transformer+TCN	89.52
Ryumin et al.	AVCRFormer	89.57
	Ours	**89.63**

Subsequently, we evaluated various studies using different deep learning networks for lip reading recognition and compared their results with ours, as shown in Table 2. The comparison indicates that our approach indeed enhances the accuracy of lip-reading recognition.

CONCLUSION

Lip-reading recognition is a crucial technology in the field of intelligent human-computer interaction and has demonstrated broad application value across various domains. This study builds on a lip-reading architecture that combines ResNet and TCN, further integrating the lightweight SimAM module. Significant advancements have been achieved in both model design and performance optimization. Specifically, we leveraged the efficiency and effectiveness of the SimAM module in capturing critical features across channel and spatial dimensions to enhance ResNet's ability to extract spatial features from lip images. Meanwhile, TCN contributed to improved training efficiency and prediction performance by leveraging parallel computation and its capacity to capture long-term dependencies in sequence modeling. Experimental results on the LRW dataset demonstrate that the proposed model achieves improvements in both accuracy and efficiency compared to existing methods, validating its effectiveness and practical value. In the future, we aim to conduct further research to enhance the model's speed and accuracy.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (61971007 and 61571013).

REFERENCES

Bai, S., Kolter, J. Z., Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling [C]. arXiv preprint arXiv:1803.01271, 2018.

Card, S. K., Moran, T. P., Newell, A. The Psychology of Human-Computer Interaction [M]. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

Chung, J, Gulcehre, C, Cho, K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. 2014.

Chung, J. S., Senior, A., Vinyals, O., et al. Lip Reading Sentences in the Wild [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 3444–3453.

Duan, X., Sun, X., Liu, H., et al. Multi-modal lip reading [C]. Proceedings of the 27th ACM International Conference on Multimedia (ACM MM), 2019: 1267–1275.

Graves, A., Mohamed, A.-R., Hinton, G. Speech recognition with deep recurrent neural networks [C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013: 6645–6649.

He, K. M., Zhang, X. Y., Ren, S. Q., et al. Deep Residual Learning for Image Recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770–778.

Howard, A. G., Zhu, M., Chen, B., et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [C]. arXiv preprint arXiv:1704.04861, 2017.

Jelinek, F. Continuous speech recognition by statistical methods [J]. Proceedings of the IEEE, 1976, 64(4): 532–556.

LeCun Y., Bottou L., Bengio Y., et al. Gradient-Based Learning Applied to Document Recognition [C]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

Li, J, Mohamed, A, Zweig, G, et al. LSTM time and frequency recurrence for automatic speech recognition [C]. Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 2015: 187–191.

Petajan, E. Automatic Lipreading to Enhance Speech Recognition [C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1984: 46–48.

Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition [C]. Proceedings of the International Conference on Learning Representations (ICLR), 2015.

Tran, D., Wang, L., Torresani, L. Learning spatiotemporal features with 3D convolutional networks [C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 4489–4497.

Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need [C]. Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017: 5998–6008.

Yang, L., Zhang, R.-Y., Li, L., et al. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks [C]. Proceedings of the 38th International Conference on Machine Learning (ICML), 2021: 11863–11874.

Yue, X., Wang, Y., Xie, L., et al. Lip reading with deep learning: A comprehensive review [C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 3388–3396.

Zhang, X., Zhou, X., Lin, M., et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6848–6856.

Zhou, W., Wu, Y., Zha, H. Lip reading using transformer [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2020: 1–15.