Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 19(12): 13526–13540.
DOI: 10.3934/mbe.2022631
Received: 14 June 2022
Revised: 12 August 2022
Accepted: 24 August 2022
Published: 15 September 2022

Research article

Research on lip recognition algorithm based on MobileNet + attention-GRU

Yuanyao Lu* and Kexin Li

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

* **Correspondence:** Email: luyy@ncut.edu.cn.

Abstract: With the development of deep learning and artificial intelligence, the application of lip recognition is in high demand in computer vision and human-machine interaction. Especially, utilizing automatic lip recognition technology to improve performance during social interactions for those hard of hearing, and pronunciation is one of the most promising applications of artificial intelligence in medical healthcare and rehabilitation. Lip recognition means to recognize the content expressed by the speaker by analyzing dynamic motions. Presently, lip recognition research mainly focuses on the algorithms and computational performance, but there are relatively few research articles on its practical application. In order to amend that, this paper focuses on the research of a deep learning-based lip recognition application system, i.e., the design and development of a speech correction system for the hearing impaired, which aims to lay the foundation for the comprehensive implementation of automatic lip recognition technology in the future. First, we used a MobileNet lightweight network to extract spatial features from the original lip image; the extracted features are robust and fault-tolerant. Then, the gated recurrent unit (GRU) network was used to further extract the 2D image features and temporal features of the lip. To further improve the recognition rate, based on the GRU network, we incorporated an attention mechanism; the performance of this model is illustrated through a large number of experiments. Meanwhile, we constructed a lip similarity matching system to assist hearing-impaired people in learning and correcting their mouth shape with correct pronunciation. The experiments finally show that this system is highly feasible and effective.

Keywords: lip recognition; MobileNet; GRU; attention mechanism; deep learning

1. Introduction

Lip recognition technology [1] is utilized to recognize aspeaker's words as text content, only according to the visual information of the speaker's lip movement, which possesses considerable application value in lip interaction, speech recognition in noisy environments and silent video processing recognition [2]; and, it is also of great significance in the field of auxiliary authentication and public security [3]. In addition, with the increasing number of people with hearing impairment year by year, hearing impairment has gradually become one of the major global health problems [4]. According to the sixth national census, as of 2010, the number of people suffering from hearing impairment in China has reached 20.54 million, and the number of hearing-impaired children is also increasing year by year. These hearing-impaired people cannot hear anything for a long time, and their verbal communication ability will gradually deteriorate. Therefore, it is very necessary to prevent speech apraxia in hearing-impaired people as soon as possible [5]. However, the lack of hearing health care personnel and related resources makes it difficult to treat hearing impairment. Therefore, some researchers have put forward the use of automatic lip-reading technology to improve the social interactions of hearing-impaired people, to supplement auditory information by capturing visual information to make it more complete and even to obtain the language content of the speaker via the visual channel alone in certain cases [6]. This can be used as an effective means of communication between the hearing impaired and older people. These urgent needs enable lip recognition technology to continue to be studied and developed, and then achieve certain results.

In recent years, deep learning has developed rapidly [7]. In many fields of computer vision, deep learning has shown excellent performance. Deep learning-based lip recognition is composed of multi-layer convolutional networks, and features are learned through convolutional operations without manual design. Features are generally abstract and have low interpretability. Researchers have introduced a focus mechanism to convolutional neural networks (CNNs) to focus on areas of interest [8,9], and image classification and target detection have also achieved great success. For example, there is the CNN feature extraction method based on an attention mechanism [10] proposed by Yang et al. Schmidhuber proposed a recurrent structure based on long short-term memory (LSTM) [11,12]; it solved the problem of gradient disappearance in recurrent neural networks [13]. LSTM considers the influence of past information on current information, and Ma Ning et al. applied LSTM to lip recognition, effectively solving the problem of lip-reading information diversity [14]. However, gated recurrent units (GRUs) [15], as a variant of LSTM, can not only consider the impact of past information on current information and solve issues of lip-movement information diversity effectively, but it can also save much time in the case of large training data.

Inspired by the above analysis, we mainly focused on the application of lip recognition technology based on deep learning, specifically, the study of a speech-correction system for hearing-impaired people. The main contents of this paper are as follows: the research on extracting lip sequence image features with a lightweight network, the research on integrating an attention mechanism to obtain a high recognition rate and the construction of a lip recognition and correction system. The rest of the paper is organized as follows. In Section 2, we introduce the preparation and structure of the lip-reading model. In Section 3, the experimental results of our proposed method and the analysis of the correction system are presented. The fourth section gives the conclusions of this study.

2. Frame structure

In this section, we discuss the proposed framework and the main process according to the following three parts, as is shown in Figure 1.

First, we need to preprocess the lip-movement sequences and extract video frames with fixed length from videos. After that, we can determine the positions of the lips and reset the input size of the RGB color image with a size of 224 × 224.

Second, we input the sequences obtained in the first stage into MobileNet to generate high-dimensional feature vectors, which represents video spatial feature information.

Finally, we take advantage of a GRU network based on an attention mechanism [16] to extract the temporal features of images and adjust the weight parameters. Meanwhile, we predict the final recognition result through the use of a Softmax layer [17].

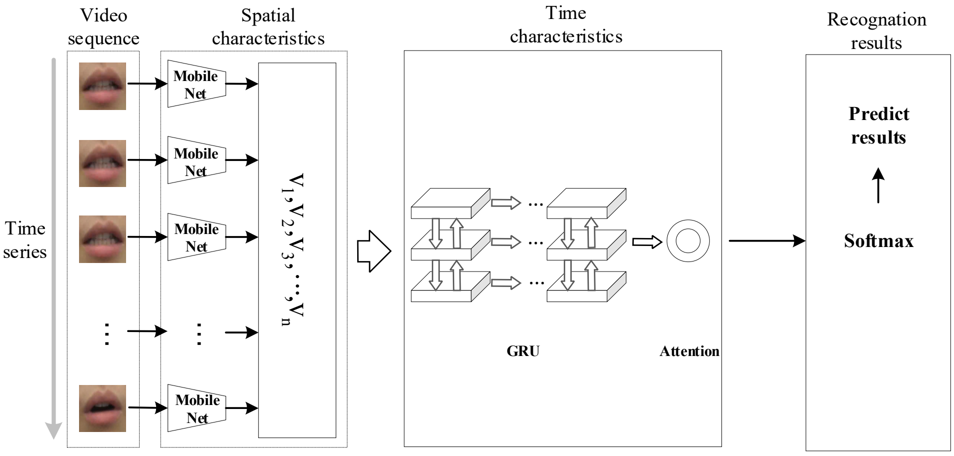


Figure 1. Structure of lip recognition model.

2.1. Preprocess

Our dataset is composed of six speakers (three male and three female), and the recorded content is the independent pronunciation of 10 English words 0–9, each pronounced 100 times, with a database sample size of 6000; each independent pronunciation lasts about 1–2 seconds and was recorded with a digital video camera. We divided 0–9 into separate parts and separated the pronunciation of the same number 100 times so that each video would be processed independently. We split the dataset into a training set which accounts for 80% and a test set which accounts for 20%.

Generally, there are approximately 25 frames per second in the data collected from the videos, and the pronunciation length of each word exists difference mostly. Furthermore, actually, the pronunciation of each word provides a series of images and redundant information representing lip movement, which indicates that it is extremely difficult to obtain the correlation between model feature extraction and image sequences in the process of training. Thus, we divided the videos into different regions according to the aggregate number of video frames, whose detail procedure is that we need to distribute the range of each block equally as much as possible so that the number of remaining frames x is less than that of the area blocks n under the condition that the fixed number of frames to be extracted is n given the premise of known prior conditions; and, the range of the first eight blocks is expanded by one video unit. As a consequence, the range of each block has been assigned semi-randomly. According to the positive order method, the calculation process of frames in the corresponding area blocks can be expressed in Eqs (2.1) and (2.2):

$$x = v - \left[\frac{v}{n}\right] \times n \tag{2.1}$$

$$F = A^{i}_{block_n} \tag{2.2}$$

where v means that there are v frames in the recognized video; F expresses the frame number extracted from each block and $A^{i}_{block_n}$ means that i frames will be extracted from the nth block.

Under the premise of a natural scene, where the system can grasp accurately the position of the lip, we implemented the preprocessing steps of lip image segmentation so that the lip recognition system can complete the task of lip recognition. First, we extracted the coordinates of the key points by using the key point detection technology with 68 faces from the Dlib library [18]. Then, according to the relative position information of facial features and the key point information of lips, lips can be precisely positioned. The serial number of the key points of Face 68 is shown in Figure 2.

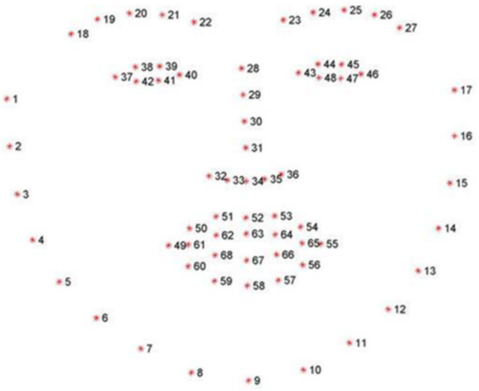


Figure 2. Dlib Face 68 key point distribution maps.

Furthermore, the Dlib library was used to recognize and locate the key points of a human face. Face key point detection takes a face image as input, and the face structure returned is simultaneously taken as the output, which is composed of markers that depend on different and specific face attributes. Additionally, we chose four key markers to locate the lip, whose labels are respectively 49, 52, 55 and 58. Then, through the coordinates of the four center points, we can segment the lip image and remove redundant information. Ultimately, we calculated the position of the lip center according to the lip coordinates of the boundary, which is denoted as $O(x, y)$; the positioning method can be expressed using Eqs (2.3) and (2.4):

$$L_1 = x_0 \pm \frac{w}{2} \tag{2.3}$$

$$L_2 = y_0 \pm \frac{h}{2} \tag{2.4}$$

where w and h represent the width and height of the mouth image, respectively, L_1 and L_2 represent the left, right and upper, lower dividing lines around the mouth, respectively.

This method possesses the characteristics of robustness, high efficiency and strong feature vector consistency. The lip movement sequencing process is shown in Figure 3.

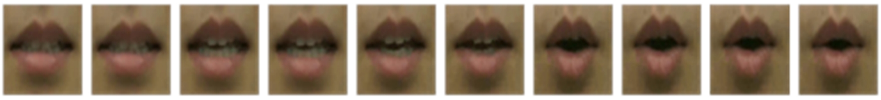


Figure 3. Lip movement diagram showing pronunciation of “zero” video sequences.

Additionally, there were 41 frames of video files with the pronunciation of “zero” from female volunteers; after extraction, the frame numbers were set as 4, 6, 13, 14, 17, 22, 25, 30, 35 and 38, respectively.

2.2. MobileNet

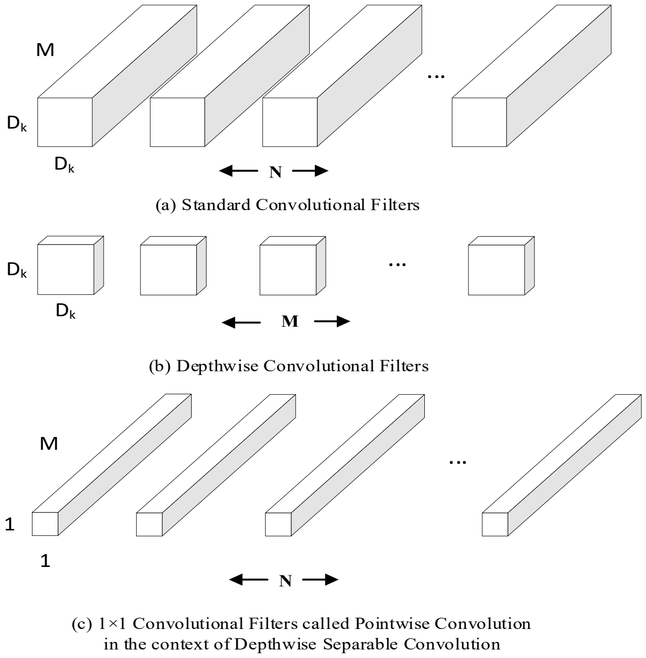
Currently, with the development of deep learning technology, not only should we consider the performance of network feature extraction, but we need to take the degree of dependence of the network on computer hardware into account during the process of designing a lip recognition system. Furthermore, the majority of recognition tasks require real-time performance. Thus, model compression [19] and knowledge distillation [20] were utilized to optimize models and a small network, i.e., MobileNet, was used to achieve the purpose of making models lightweight simultaneously. A comparison of frequently used feature extraction networks is shown in Table 1.

Table 1. Comparison of neural networks for feature extraction.

Model	Size	Top-1 accuracy	Top-5 accuracy	Number of parameters	Depth
Xception	88MB	0.790	0.945	22,910,480	126
VGG16	528MB	0.715	0.901	138,357,544	23
VGG19	549MB	0.728	0.910	143,667,240	26
ResNet50	99MB	0.759	0.929	25,636,712	168
InceptionV3	92MB	0.788	0.944	23,851,784	159
MobileNet	17MB	0.665	0.871	4,253,864	88

The results in the table above were evaluated using an ImageNet data set [21]. Additionally, the performance of basic feature extraction did not decrease sharply with the reduction of model parameters depending on the Top-5 accuracy. Thus, we adopted the model architecture of a feature extraction network based on MobileNet.

The basic unit of the MobileNet is depth-wise separable convolution (DSC), which solves the problems of low computational efficiency and large parameters in the convolutional network. Moreover, the idea of a MobileNet is to replace the original complete convolutional operator with the structural unit composed of deep convolution and point-by-point convolution, decomposing convolution into two separate layers. The first layer is shown in Figure 4(b), The anti-direction convolution completes the filtering operation via application of a convolutional filter in each input channel, which is lighter than the original convolution. The second layer is shown in Figure 4(c); point-by-point convolution consists of 1 × 1 convolution, which is utilized to calculate the linear combination of input channels and enhance the nonlinear fitting ability of the network. Furthermore, the original convolutional filter is shown in Figure 4(a), which suggests that the idea of the MobileNet is very ingenious by comparison.



(a) Standard Convolutional Filters

(b) Depthwise Convolutional Filters

(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 4. Comparison between DSC and standard convolution.

Thus, we utilized a MobileNet to extract the visual feature sequences of lips. More specifically, the feature sequencing length extracted by the MobileNet was 1 × 1024, which is one-fourth of the amount of feature parameters extracted by VGG. However, there is little performance degradation.

2.3. Attention-GRU

For the reasons that a MobileNet neural network cannot express the changes in time sequences effectively and lip recognition needs to pay attention to the relationship between image sequences and time simultaneously, we utilized deep learning so that the recurrent neural network can store the information of the previous sequences through a hidden state. Additionally, the GRU network is one of the recurrent neural networks, which can capture the dependence between images with a larger time step in image sequences by introducing reset gate and update gate structures. Furthermore, the input of the reset gate and update gate in the gating cycle unit are the current time step input x_i and the hidden state of the previous time step h_{t-1}, while the output is calculated by the full connection layer of the activation function σ (sigmoid function), as shown in Figure 5.

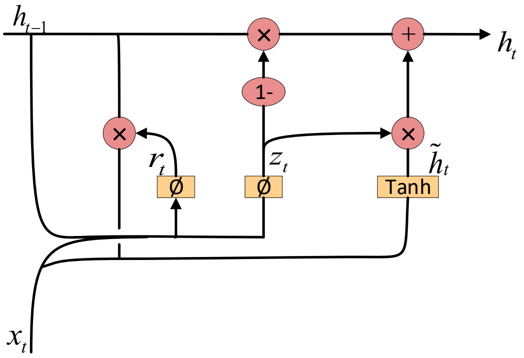


Figure 5. Structure of GRU network.

We set the number of hidden units as h, the small batch input of time step t to $x_t \in R^{n\times d}$ (sample number is n and input features number is d) and the hidden state of the previous time step as $h_{t-1} \in R^{n\times d}$. As a consequence, the mathematical expressions of the reset gate and update gate are listed as follows:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \tag{2.5}$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \tag{2.6}$$

where $W_{x\hbar}$, W_{xr}, $W_{xz} \in R^{n\times d}$, $W_{h\hbar}$, W_{hr} and $W_{hz} \in R^{h\times h}$ are weight parameters, $b_{\hbar}$, b_r and $b_z \in R^{1\times h}$ are bias parameters, $r_t \in R^{n\times d}$ is the reset door, the update door is $z_t \in R^{n\times h}$, x_t is the time step t of the input data and y_t is the output of the output layer.

Furthermore, the mathematical expression of the candidate hidden states $\hbar_t$ of time step t is shown as Eq (2.7), the hidden state of time step t is shown as Eq (2.8) and the output layer is shown as Eq (2.9), as follows:

$$\hbar_t = tanh \tag{2.7}$$

$$h_t = z_t \odot \hbar_t + \tag{2.8}$$

$$y_t = \sigma(W_0 \cdot h_t) \tag{2.9}$$

According to the expressions of the candidate hidden states, the reset gate controls how the hidden state of the previous time step flows into the candidate hidden state of the current time step. Meanwhile, the hidden state of the previous time step may contain all of the historical information of the time series up to the previous time step, and the calculation of the hidden state h_t shows that the update gate can control how the hidden state is updated by the candidate hidden state. Additionally, the reset gate captures short-term dependencies in the time series, while the update gate helps to capture long-term dependencies. Thus, through the current design of this structure, we can effectively capture the dependence of the time step distance in a time series, and the extraction of longtime step sequence relation information is particularly significant for lip recognition.

An attention mechanism is essentially used to calculate the probability distribution of attention [22], meaning that pivotal features will be assigned more attention, which is universally utilized to process sound signals with temporal relations. In other words, an attention mechanism has a good effect on the expression of sequences with temporal information. The structure of a GRU network integrated with an attention mechanism is shown in Figure 6.

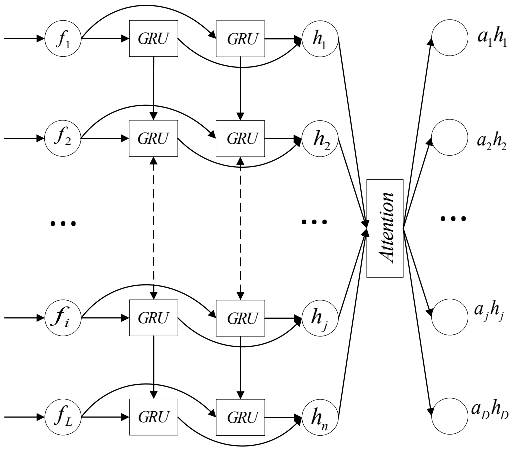


Figure 6. Schematic diagram of attention mechanism.

We input the temporal features into the GRU neural network with an attention mechanism to express them, assign different weights and highlight the impact of temporal features on the classification results, which are extracted by the feature extraction network. The feature expression of a signal passing through the feature extraction network is shown in Eq (2.10):

$$f_{res} = \{f_1, f_2, \dots, f_M\}, f_i \in R^D \tag{2.10}$$

where M represents the number of eigenvectors output by the MobileNet network, f_i represents the eigenvectors of $1 \times D$, D represents the dimension of features and f_{res} represents the extracted features, which will be input into the GRU network layer of the attention mechanism. Specifically, the GRU network first transforms the features f_{res} into hidden layer features $H = [h_1, h_2, \dots, h_D]$ ($H \in R^N \times D$, where H is the length of the hidden layer of the GRU network).

The attention mechanism generates an attention matrix on the basis of hidden variables, the process of which is as follows:

$$u_i = tanh(W_s h_i + b_s) \tag{2.11}$$

$$a_i = \frac{exp(u_i)}{\sum_i exp(u_i)} \tag{2.12}$$

$$f_{map} = \sum_i a_i h_i \tag{2.13}$$

where W_s represents the weight matrix of the hidden variables and possesses $N \times D$ dimensions, $h_i$represents the hidden layer state of input sequences and b_s denotes bias terms. Equation (2.11) maps the hidden variable characteristic h_i to the interval $[-1,1]$, while the weight coefficient α_i of each hidden variable is calculated using Eq (2.13), which is attention. Ultimately, we multiply and sum the hidden variables and weight coefficients to get the final feature expression f_{map}.

The optimizer we use is the adaptive moment estimation (Adam) optimizer, which has high computational efficiency, low memory requirements and needs virtually no adjustment of the hyperparameters, which is suitable for solving large-scale parameter optimization problems. In general, the Adam algorithm is preferred when the optimizer for the model and data are uncertain.

We chose SoftMax as the classifier and fully connected layer to complete the classification task. Since classification is an inverse operation, it is suitable to use SoftMax as the activation function, and the predicting result with the highest probability is the recognition result of the model.

3. Experimental data set and results

As deep learning continues to evolve in the industry and academia, its frameworks are also being updated and iterated. Currently, the mainstream frameworks for deep learning include TensorFlow, Keras, MXNet, Caffe and Pytorch, which provide convenient experimental tools for researchers with different purposes.

Since the requirement is based on the system, and in order to dig deeper and keep updating the model, we finally decided to use Pytorch, the mainstream framework in the industry, as the base framework through which the model is trained and deployed. Therefore, we chose a more stable version as the base library for training our model.

Network models require not only an excellent structure, but also training techniques to reach the optimal solution, so the setting of the hyperparameters and training methods are particularly important. Hyperparameters are parameter values set in deep learning or machine learning, and the goal of optimization is to reach the optimal hyperparameters. During the training and learning process, these parameters directly affect the final results, and there is even a possibility that the training process performs well but does not work. Therefore, the setting of the hyperparameters is particularly important. The hyperparameters were set as shown in Table 2 below.

Table 2. Hyperarameter setting.

Parameter	Value
Learning Rate	0.001
Activation Function	ReLU
Epoch	70
Optimizer	Adam
Learning Rate Decay	5

3.1. Data set

The experiment is implemented on our own data set, which consists of six speakers (three boys and three girls) giving out 10 isolated English numbers (from 0 to 9). Additionally, each pronunciation is an independent video clip in a natural environment, and participants face the video equipment and pronounce the corresponding words naturally. The resolution of the videos was 1920 × 1080, the frame rate was 25 frames per second and the duration of each independent voice is 1–2 seconds. Moreover, we labeled and segmented the original video seriatim so that the beginning and end of each pronunciation unit can accurately be located, and each isolated video can be extracted for a fixed 10-frame length. Finally, a 64 × 64 pixels window of the mouth region will be obtained after processing each video frame.

3.2. Research results

In this part, we chose accuracy and recall as the evaluation indexes to evaluate our neural network model and randomly divide the data set into 80% training data set and 20% test data set to analyze and compare the results. Furthermore, we chose PyTorch as the framework, and the initial learning rate was 0.001, which needs to decay in order to effectively approximate the optimal solution during the training. Thus, the learning rate l_r was set to decay by 10^{-1} when the loss of five consecutive epochs is no longer reduced; the activation function was set as a ReLU.

We tested the lip recognition performance of the constructed fusion network. Thus, 100 groups of lip-reading sequences were randomly selected for each number for detection. As the evaluation result of accuracy was established to reflect all 10 classification problems in the whole recognition system, the overall accuracy was very high, reaching more than 87%. The main reasons for the accuracy rate are as follows: 1) the lip region is small and the difference between lip movements is weak, and 2) there are not enough researchers and insufficient data sets for lip-reading recognition. Then, we determined the recall statistics for each English-language pure number pronunciation prediction result, as shown in Figure 7.

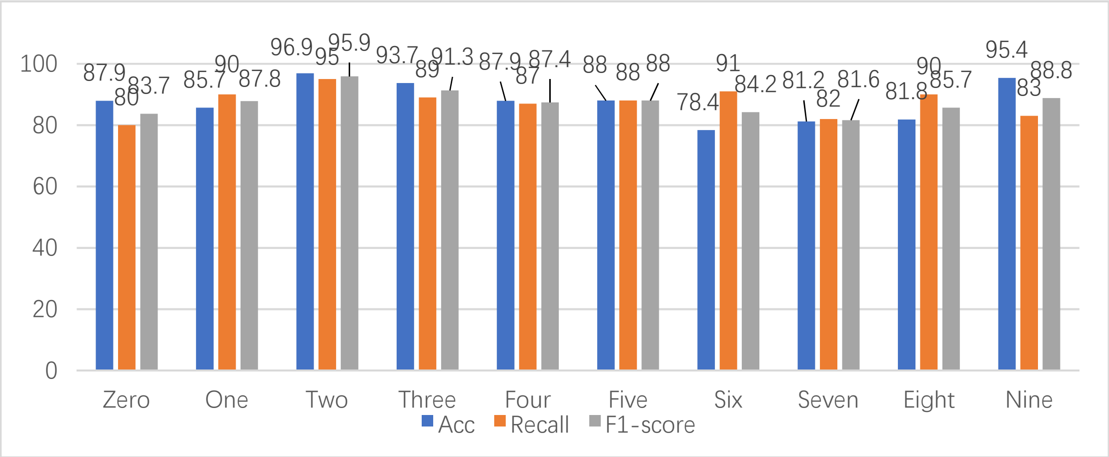


Figure 7. Prediction of pronunciation of English numbers.

According to the bar chart of accuracy and recall rate, it is not difficult to see that in terms of lip language recognition, the number 7 had the lowest recall rate, followed by the number 4 and number 9. However, in the accuracy evaluation index, the number 0 is obviously lower than the other nine numbers. The main reason for this is that the gestures of the numbers 0 and 7 are similar. In the system of lip recognition, the numbers 7 and 9 had low recall rates, and Numbers 6, 7 and 8 had low accuracy rates, which is mainly due to the similar lip-movement trends for the Numbers 6, 7 and 8 in English pronunciation. Therefore, in the future, we should increase the number of training samples of these kinds of numbers to improve their accuracy and recall rates.

To make the model more robust, we also introduced the hyperparameter α as the width factor to adjust the number of channels of the feature map, e.g., MobileNet 0.5× denotes a MobileNet model with a width factor of 0.5. The proposed algorithm in this paper has been compared with the current state-of-the-art algorithms in two dimensions: floating point computation (MFLOPs) and Top-1 accuracy; the results are shown in Table 3.

Table 3. Accuracy comparison for our approach and mainstream lightweight network models based on our data sets.

Models	MFLOPs	Top-1 Acc. (%)
ShuffleNet 1×	137	65.9
SD-ShuffleNet 1×	149	64.0
0.5 × MobileNetV1-224	149	63.7
0.6 × MobileNetV2-224	141	65.6
ours	144	65.3
ShuffleNet 0.5 ×	38	57.3
SD-ShuffleNet 0.5 ×	42	52.5
0.25 × MobileNetV1-224	41	50.6
0.3 × MobileNetV2-224	41	54.7
ours	40	56.2
ShuffleNet 0.25 ×	13	46.7
SD-ShuffleNet 0.25 ×	12	39.4
0.125 × MobileNetV1-224	12	39.6
0.15 × MobileNetV2-224	12	36.0
ours	12	45.1

From the above table, it can be seen that the proposed network is significantly better than the original MobileNet in terms of accuracy, given the limitation of floating-point computation. In the longitudinal analysis, our method requires only 40 MFLOPs to achieve about 63% recognition accuracy. Thus, the proposed network model requires much less floating point computation than the original MobileNet to achieve the same level of accuracy.

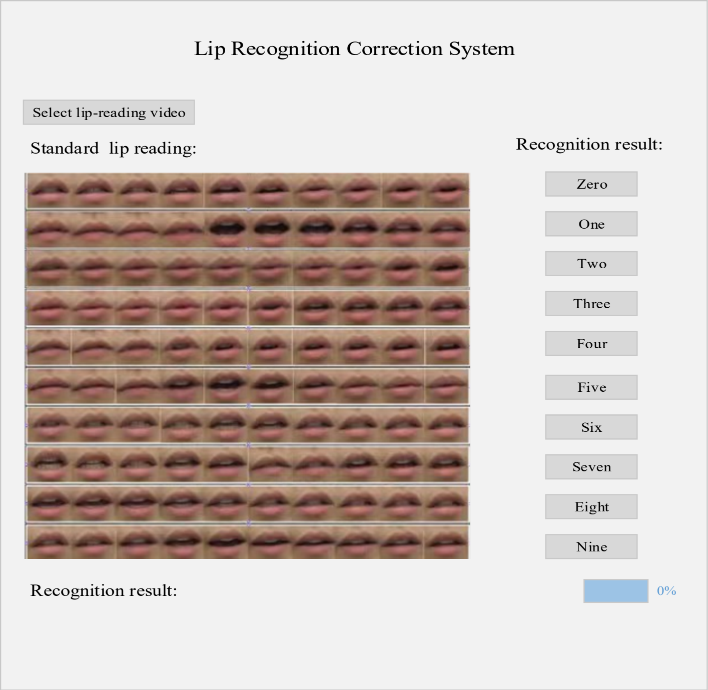


Figure 8. Lip recognition correction system.

Then, we built the lip recognition and correction system as shown in Figure 8. Two functional buttons can be seen on this interface, namely, those for selecting the lip-reading video. On the right side of the interface is the final result of the system recognition video, and the corresponding result is the sample lip sequence. At the bottom of the interface is the result of the lip recognition comparison.

In addition, people were asked to deliberately change their pronunciation and mouth shape to verify the effectiveness of our correction system. Figure 9 shows the simulation test results. The first line is a sample lip-reading sequence for the English pronunciation of the number 6. The second line shows the lip-reading sequence after deliberately changing the pronunciation. The third line is a lip-reading sequence after following the standard lip motion of the number 6. According to the system matching results, the matching degree as a result of deliberately changing pronunciation was 61.76%, while the matching degree for correct pronunciation was 80.83%.

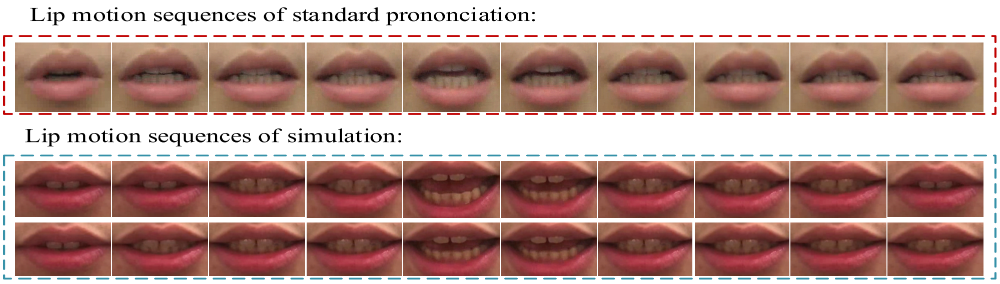

Lip motion sequences of standard prononciation:
Lip motion sequences of simulation:

Figure 9. Drawing of simulation.

In order to visualize the lip shape differences between words, we randomly selected videos of each word from the test set and regarded each word itself as a good sample and the rest of the words as bad samples; here, the lip shape differences between the words “three” and “China” and the rest of the words are shown in Figure 10.

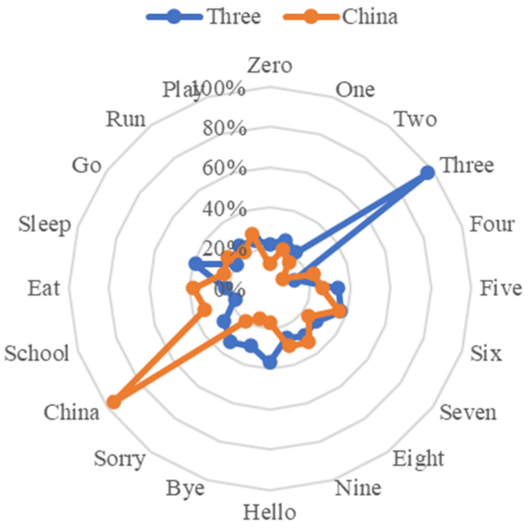


Figure 10. Lip shape differences between the words “three” and “China”.

As shown in the figure, the similarities between the words “three” and “China” and other words were limited to less than 30%, and the similarity with itself is greater than 80%, which shows that the proposed lip recognition network system has strong intra-class similarity and inter-class variability.

The experimental results show that the lip recognition application system designed in this study can effectively help people with hearing impairment to correct the pronunciation according to the standard lip-reading sequence and the lip- reading comparison results given by the system.

4. Conclusions

Lip recognition technology has profound development prospects and application demand in the field of computer vision and human-computer interaction. In particular, one of the most promising applications of artificial intelligence in healthcare and rehabilitation is the use of automatic lip recognition technology to improve the social interactions of people with hearing and pronunciation impairments. Therefore, the research of this topic focuses on the application system of lip recognition based on deep learning, namely, the speech correction system for the hearing impaired, which has more practical significance and aims to lay a foundation for the further implementation of automatic lip-reading recognition technology in the future. This study mainly used the combination of a MobileNet lightweight network and GRU network to identify the timing sequence of the speaker's lip movements and construct the lip similarity matching system to assist the hearing impaired with learning the correct lip pronunciation sequence image and correcting their mouth shape. In addition, in order to verify the effectiveness of the algorithm model and the practicability of the system, we reasonably applied the laboratory self-made test data set for test experiments. The experimental results show that the lip correction application system proposed and designed by us has high applicability and feasibility. The application system of lip recognition based on deep thought learning is based on a lightweight network, which not only has high recognition accuracy, but also can greatly shorten the running time of the system. The application system designed by us not only makes the lip recognition technology more practical, but can also effectively assist hearing-impaired people with learning to a certain extent and play a role.

Acknowledgments

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

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. M. H. Rahmani, F. Almasganj, Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features, in *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, (2017), 195–199. https://doi.org/10.1109/PRIA.2017.7983045

2. H. E. Cetingul, Y. Yemez, E. Erzin, A. M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, *IEEE Trans. Image Process.*, **15** (2006), 2879–2891. https://doi.org/10.1109/TIP.2006.877528
3. A. B. Hassanat, Visual passwords using automatic lip reading, preprint, arXiv: 1409.0924.
4. Y. Zhang, Similarity image retrieval model based on local feature fusion and deep metric learning, in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, (2020), 563–566. https://doi.org/10.1109/ITOEC49072.2020.9141871
5. I. Matthews, G. Potamianos, C. Neti, J. Luettin, A comparison of model and transform-based visual features for audio-visual LVCSR, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, *ICME 2001*, IEEE Computer Society, (2001), 825–828. https://doi.org/10.1109/ICME.2001.1237849
6. H. Ali, M. Hariharan, S. Yaacob, A. H. Adom, Hybrid feature extraction for facial emotion recognition, *Int. J. Intell. Syst. Technol. Appl.*, **13** (2013), 202–221. https://doi.org/10.1504/IJISTA.2014.065175
7. A. Rekik, A. Ben-Hamadou, W. Mahdi, An adaptive approach for lip-reading using image and depth data, *Multimedia Tools Appl.*, **75** (2016), 8609–8636. https://doi.org/10.1007/s11042-015-2774-3
8. K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, T. Ogata, Lipreading using convolutional neural network, in *fifteenth annual conference of the international speech communication association*, 2014. https://doi.org/10.1016/j.jvlc.2014.09.005
9. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., Mobilenets: efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
10. J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, et al., Neural aggregation network for video face recognition, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5216–5225. https://doi.org/10.1109/CVPR.2017.554
11. D. Lee, J. Lee, K. E. Kim, Multi-view automatic lip-reading using neural network, in *Asian Conference on Computer Vision*, Springer, Cham, (2016), 290–302. https://doi.org/10.1007/978-3-319-54427-4_22
12. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
13. A. Garg, J. Noyola, S. Bagadia, Lip reading using CNN and LSTM, in *Technical report*, *Stanford University, CS231 n project report*, 2016.
14. H. E. Romero, N. Ma, G. J. Brown, Snorer diarisation based on deep neural network embeddings, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2020), 876–880. https://doi.org/10.1109/ICASSP40776.2020.9053683
15. G. B. Zhou, J. Wu, C. L. Zhang, Z. H. Zhou, Minimal gated unit for recurrent neural networks, *Int. J. Autom. Comput.*, **13** (2016), 226–234. https://doi.org/10.1007/s11633-016-1006-2
16. L. Li, W. Jia, J. Zheng, W Jian, Biomedical event extraction based on GRU integrating attention mechanism, *BMC Bioinf.*, **19** (2018), 102–111. https://doi.org/10.1186/s12859-018-2275-2
17. H. Liu, L. F. Li, W. L. Zhao, L. Chen, Rolling bearing fault diagnosis using improved LCD-TEO and softmax classifier, *Vibroengineering Procedia*, **5** (2015), 229–234.
18. C. Krishna, Face recognition based attendance management system using DLIB, *Int. J. Eng. Adv. Technol.*, **8** (2019), 57–61.

19. S. Y. Nikouei, C. Yu, S. Song, R. Xu, T. R. Faughnan, Real-time human detection as an edge service enabled by a lightweight CNN, in *2018 IEEE International Conference on Edge Computing (EDGE)*, IEEE, (2018), 125–129. https://doi.org/10.1109/EDGE.2018.00025

20. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, preprint, arXiv:1503.02531.

21. Y. Wei, Z. Yao, C. Lu, S. Wei, L. Liu, Z. Zhu, et al., Cross-modal retrieval with CNN visual features: a new baseline, *IEEE Trans. Cybern.*, **47** (2017), 449–460. https://doi.org/10.1109/TCYB.2016.2519449

22. M. Song, H. Park, K. Shin, Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean, *Inf. Process. Manage.*, **56** (2018), 637–653. https://doi.org/10.1016/j.ipm.2018.12.005