applied sciences

MDPI

Article

Automatic Lip Reading System Based on a Fusion Lightweight Neural Network with Raspberry Pi

Jing Wen and Yuanyao Lu *

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; wwwinj@163.com

* Correspondence: luyy@ncut.edu.cn

Received: 11 November 2019; Accepted: 10 December 2019; Published: 11 December 2019

Abstract: Virtual Reality (VR) is a kind of interactive experience technology. Human vision, hearing, expression, voice and even touch can be added to the interaction between humans and machine. Lip reading recognition is a new technology in the field of human-computer interaction, which has a broad development prospect. It is particularly important in a noisy environment and within the hearing- impaired population and is obtained by means of visual information from a video to make up for the deficiency of voice information. This information is a visual language that benefits from Augmented Reality (AR). The purpose is to establish an efficient and convenient way of communication. However, the traditional lip reading recognition system has high requirements of running speed and performance of the equipment because of its long recognition process and large number of parameters, so it is difficult to meet the requirements of practical application. In this paper, the mobile end lip-reading recognition system based on Raspberry Pi is implemented for the first time, and the recognition application has reached the latest level of our research. Our mobile lip-reading recognition system can be divided into three stages: First, we extract key frames from our own independent database, and then use a multi-task cascade convolution network (MTCNN) to correct the face, so as to improve the accuracy of lip extraction. In the second stage, we use MobileNets to extract lip image features and long short-term memory (LSTM) to extract sequence information between key frames. Finally, we compare three lip reading models: (1) The fusion model of Bi-LSTM and AlexNet. (2) A fusion model with attention mechanism. (3) The LSTM and MobileNets hybrid network model proposed by us. The results show that our model has fewer parameters and lower complexity. The accuracy of the model in the test dataset is 86.5%. Therefore, our mobile lip reading system is simpler and smaller than other PC platforms and saves computing resources and memory space.

Keywords: mobile lip reading system; lightweight neural network; face correction; virtual reality (VR)

1. Introduction

Lip reading refers to recognition of what people are saying by catching the speaker's lip motion. Especially in a noisy environment of voice superposition, or people with hearing impairment, the system will automatically detect lip area and identify the information [1]. Lip reading technology can supplement speech information by visual perception based on enhanced learning. Meanwhile, automatic lip reading technology can be widely used in Virtual Reality (VR) systems [2], information security [3], speech recognition [4] and auxiliary driving systems [5]. The lip reading system is mainly divided into the lip reading system based on traditional methods and the lip reading system based on in-depth learning. Traditional lip reading systems usually include two aspects: feature extraction and classification. For feature extraction, there are two kinds of methods: pixel-based and model- based. Pixel-based feature extraction uses the pixel values extracted from the

interested mouth Region of Interest (ROI) as visual information. Then, the abstract image features are extracted by Discrete Cosine Transform (DCT) [6], Discrete Wavelet Transform (DWT) [7], and Principal Component Analysis (PCA) [8]. The method based model is to express the lips by a mathematical model, approximate the lip contour infinitely with curves and special features, and obtain the lip geometric features. For classification, the extracted features are sent to the classifier for classification. The commonly used classifiers are Artificial Neural Network (ANN) [9], Support Vector Machine (SVM) [10], and Hidden Markov Models (HMM) [11]. The breakthrough of in-depth learning also affects the development of lip reading technology. It has changed from the research direction of combining artificial design-based features with traditional classification model to an end-to-end complete system based on a deep-level neural network [12].

In recent years, researchers of the Google team proposed the MobileNets. MobileNets model to be an efficient model for mobile and embedded visual applications; it can combine depth separable convolution to construct a lightweight depth neural network [13]. This type of network offers an extremely efficient network architecture that can easily be matched to the requirements for mobile and embedded applications [14]. Considering that lip feature extraction has voice information and visual perception, we propose a hybrid neural network which combines MobileNets and LSTM to build a mobile lip reading system based on Raspberry Pi. Raspberry Pi is a credit card sized microcomputer which can do everything a normal PC can do and is widely supported by a large number of users. For example, it can be embedded in VR wearable devices. The whole lip reading recognition system runs on Raspberry Pi which is based on the Linux system; we deployed our project to the destination folder (/home/pi/lip-recognition) of our Pi to realize the self-startup. In this article, we realized self-startup by adding a script file. Also, we compared the Raspberry Pi with android smartphones and computers. Smartphones are limited by the space of PCBA (Printed Circuit Board Assembly), therefore they cannot allow the corresponding USB, HDMI and other interfaces. The low hardware cost of smartphones leads to low software adaptability. Although the computer has powerful process capability, it is inconvenient to move and cannot be deployed in simple devices. In contrast, Raspberry Pi has the advantages of small size, easy to carry, and low cost.

Our lip reading recognition system on mobile devices can be divided into the following stages: First, a lip reading video is obtained by a camera connected to the Raspberry Pi, and frames are extracted by using our own design rules to reduce the complexity of redundant information [15]. In the second stage, the multi-task cascade convolution network (MTCNN is used to correct the face and extract the key points of lip region [16]. Then MobileNets are used to extract lip features. After this, the attention-based LSTM network is used to learn the sequence information and attention weight between key frame features of the video. Finally, the final recognition results are predicted by two full connection layers and softmax. The softmax function converts the prediction results into probability [17]. The advantages of this mobile lip reading system are: (1) Face correction and lip key point detection using the MTCNN network can improve the accuracy of feature extraction. (2) Compared with PC-based mobile devices, Raspberry Pi has the advantages of small size, low power consumption and low cost. It can also accomplish some PC tasks and applications as usual. (3) Hybrid neural networks based on MobileNets and LSTM can reduce the number of parameters, the model complexity and the interference of invalid information.

The rest of this paper is organized as follows: In Section 2, we introduce the preparation and architecture of mobile lip reading system. Section 3 contains the analysis and experimental results of our proposed method. Section 4 provides conclusions and suggestions for future research directions.

2. Proposed Model

In this section, we propose the research framework and main steps. The framework we designed is a video recognition system based on mobile devices. Considering the performance limitations of mobile devices, we propose a framework as shown in Figure 1. First, we need to handle the dynamic video. We design an efficient method to extract the fixed frame. Second, we implement face location

and face correction. Then we segment the mouth image region using MobileNets to extract features. Finally, we learn the temporal features and predict recognition results from LSTM.

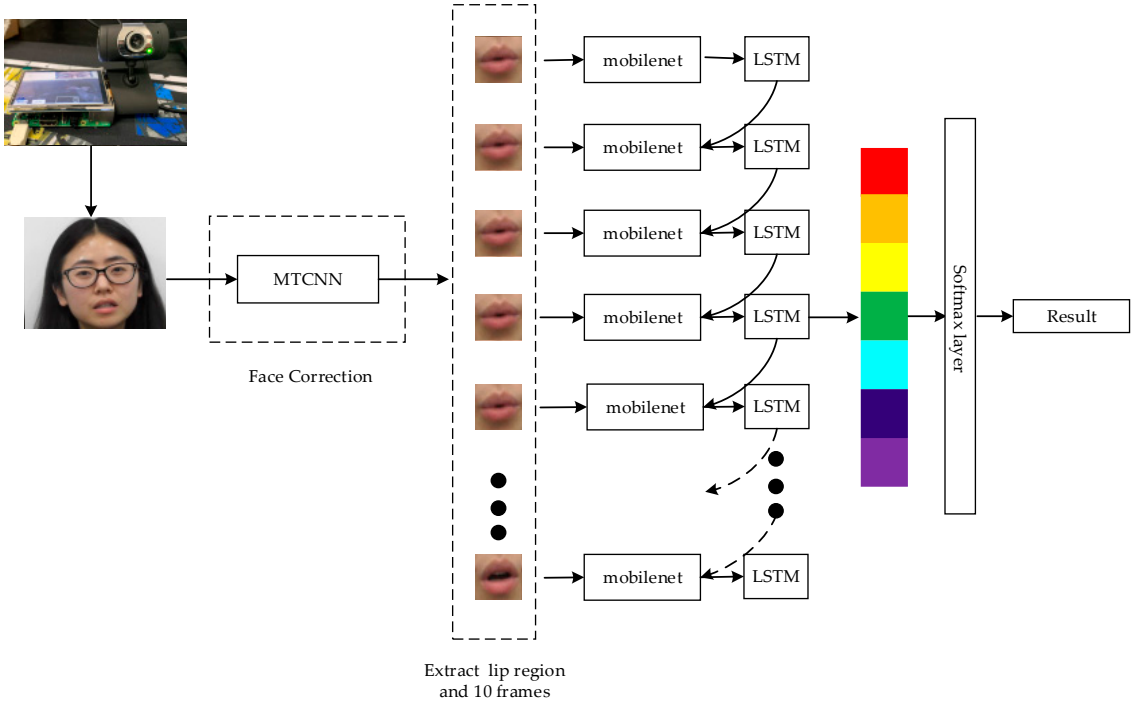


Figure 1. The architecture of our proposed Lip Reading System.

2.1. Extraction of Fixed Frames, MTCNN Detection and Correction of Lips

Lip detection is an essential part of lip reading recognition. However, previous studies were based on Dlib to locate the lips of a face [18]. Then the lip is segmented for feature extraction. In this paper, we independently design a frame extraction rule and propose a multi-task cascade convolution network (MTCNN) to extract the lip region as training data and locate key lip points to correct lip areas.

The quality of extracting fixed frames directly determines the quality of the recognition results. Therefore, we design a frame extraction scheme for lip recognition. In order to increase the robustness of the model, we design and implement a partition-based random selection method. If the total number of frames in a video segment is V, we first divide the video V into x blocks (x = 1, 2, 3, 4, 5 ... n). F represents the sequence number of each frame, because there may be situations where it cannot be divisible, we reduce the total number of frames. As shown in Formula (1).

$$x = v - \frac{v}{n} * n \tag{1}$$

Among them, $\lfloor \ \rfloor$ is the downward integer operator, the first X blocks the increase of the number of frames by, for each block, two frames are extracted as fixed frames. As shown in Formula (2).

$$F = A^{i}_{block_n} \tag{2}$$

Among them, $A^{i}_{block_n}$ represents selecting i frames in $block_n$ orderly.

MTCNN has an absolute advantage in the performance and recognition speed of face detection [19]. It is based on a cascade framework and can be divided into three layers: P-Net, R-Net, and O- Net [20]. The specific network structure is as follows:

- Proposal Network (P-Net): The network structure mainly obtains the regression vectors of the candidate windows, and the boundary areas of the face. The boundary areas are used for

regression analysis to calibrate the candidate windows, and then merge the highly overlapping candidate windows by Non-maximum Suppression (NMS). (See Figure 2a).

- Refine Network (R-Net): The network structure also removes the false regions by boundary area regression analysis and NMS. However, due to the difference between the network structure and the P-Net network structure, there is an additional full-connection layer, so it can achieve a better effect of restraining the misjudgment rate. (See Figure 2b).
- Output Network (O-Net): This layer has one more convolution layer than the R-Net layer, so the processing results will be better. It works the same as the R-Net layer, but the layer monitors more of the face area and outputs five landmarks. (See Figure 2c).

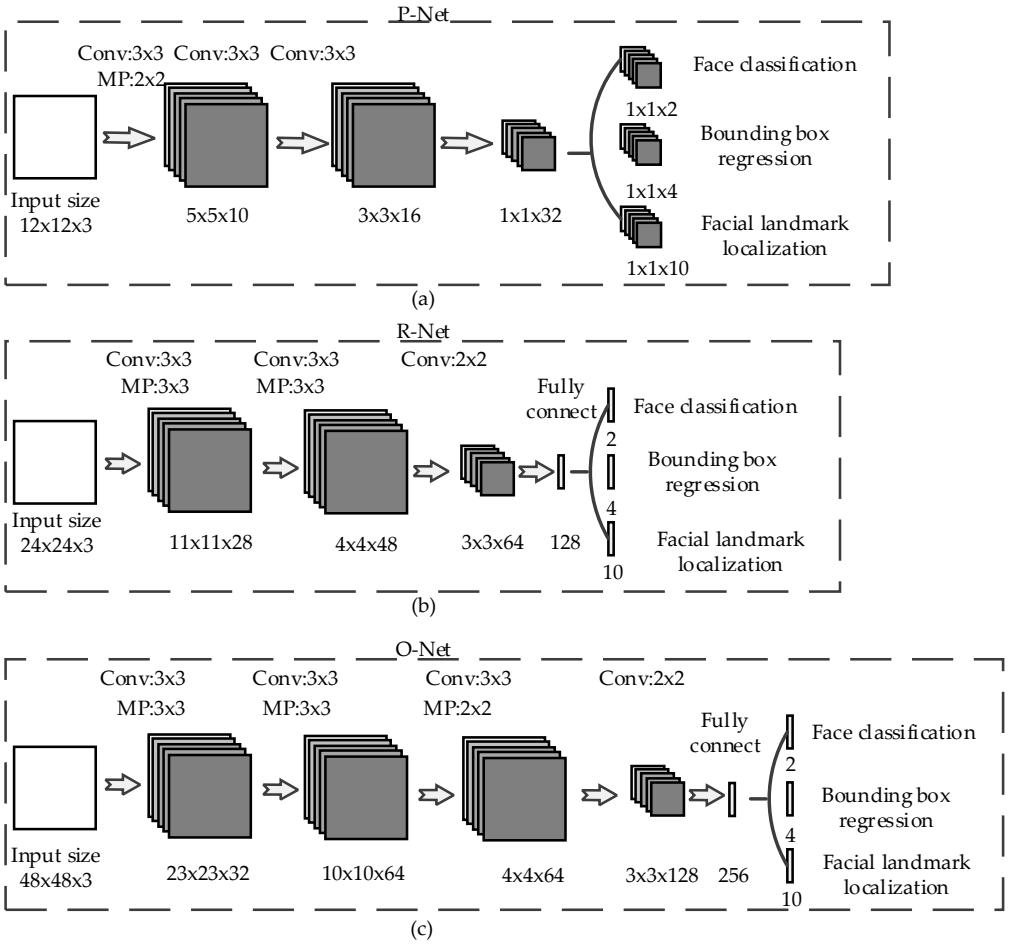


Figure 2. Multi-task cascade convolution network (MTCNN) architecture.

MTCNN model can detect the face area and face landmarks concurrently, and realize the calibration of feature landmarks. In this process, the model uses the method of Non-maximum Suppression. Based on this, we can achieve the goal of correcting the face. We achieve an effect by using MTCNN as shown in Figure 3 to improve the accuracy of the following recognition.

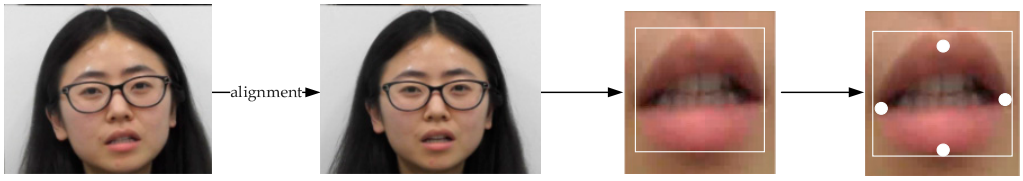
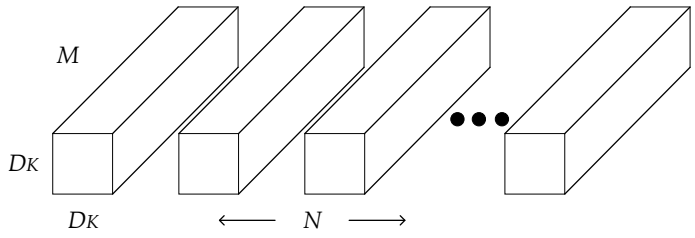


Figure 3. MTCNN face correction and lip extraction.

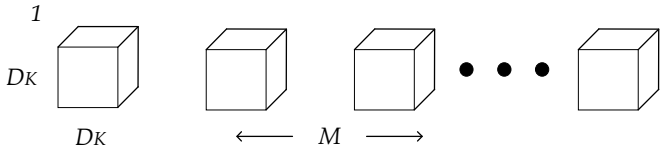
2.2. MobileNets Architecture

MobileNets based on a Streamlined Architecture uses Depthwise Separable Convolutions to construct a lightweight deep neural network. We introduce two simple global hyper-parameters. These hyper-parameters allow the model generator to select the appropriate size model for its application according to the constraints of the problem, thus reducing the complexity of the model [21].

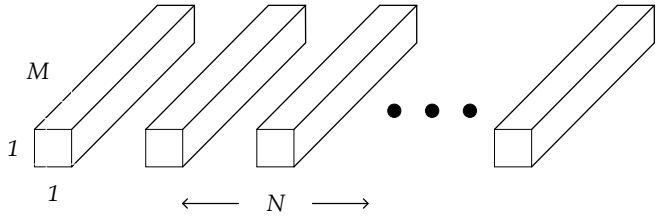
The main work of MobileNets is using Depthwise Separable Convolutions instead of Standard Convolutions to solve the problems of computing efficiency and the parameters of the convolutional network [22–24]. The Standard Convolutions are shown in Figure 4a. It decomposes the standard convolution into Depthwise Convolutions and Pointwise Convolution. It is a key component of many effective neural network structures. The basic idea is to use a decomposition version instead of a complete convolution operator to decompose the convolution into two separate layers. The first layer is shown in Figure 4b, called Depthwise Convolution, which performs lightweight filtering by applying a convolution filter to each input channel. The second layer is Figure 4c, which is a 1×1 convolution called Pointwise Convolution. It is responsible for building new features by calculating the linear combination of the input channels.



(a) Standard Convolutions Filters



(b) Depthwise Convolutions Filters



(c) 1x1 Convolutions Filters called Pointwise Convolution in the context of Depthwise Convolutions Filters

Figure 4. MoblieNets model structure.

In addition to the Depthwise Separable Convolutions, which is the basic component of MobileNets, the ReLU activation function is used in the model. Therefore the basic structure of Depthwise Separable Convolutions is shown in Figure 5. BN and ReLU are used to speed up the training speed and improve the recognition precision of the model [25].

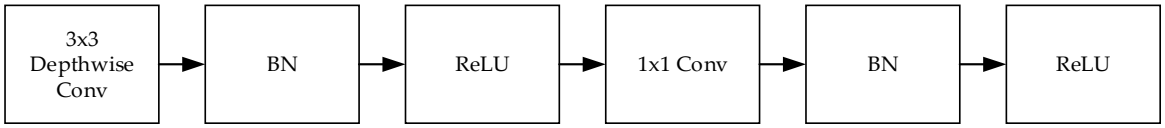


Figure 5. Depthwise separable convolutions basic structure.

2.3. LSTM (Long Short-Term Memory)

In order to solve the problem of gradient disappearance and gradient explosion when RNN processes long-sequence data, Hochreiter [26] proposed an improved form of RNN, called long short-term memory (LSTM), which is specially used to deal with information missing in long-term dependent sequences [27]. LSTM stores historical information by introducing memory units. By introducing three control gate structures, including the input gate, forget gate, and output gate, LSTM controls the increase and removal of information flow in the network. To better discover and utilize long-term dependencies from sequence data (such as video, audio, and text), memory cell remembers the associated information that needs to be remembered in a long sequence and forgets some of the useless information. Figure 6 shows the operations performed within a single LSTM cell. Among them, x_t represents the input vector of the network node at t time, h_t represents the output vector of the network node at t time, i_t, f_t, o_t, and c_t represent the input gate, forget gate, output gate and memory unit at t time respectively.

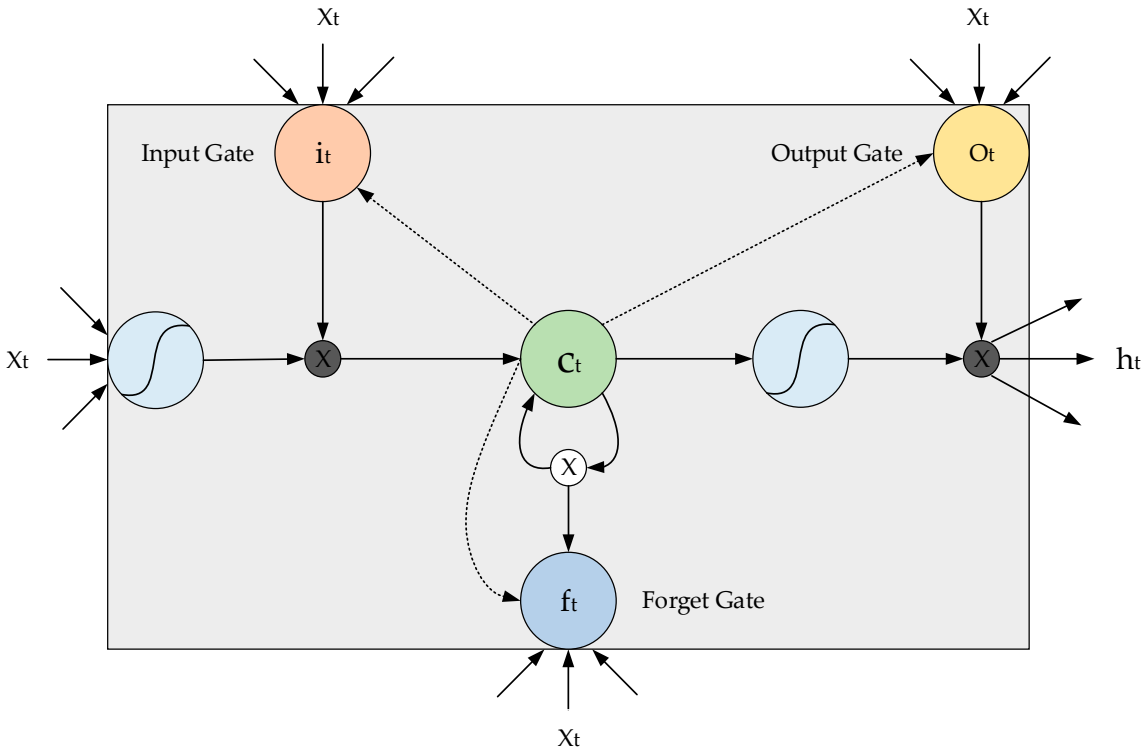


Figure 6. Long Short-Term Memory basic unit diagram.

The calculation steps of the input gate, forget gate, memory unit, and output gate in the LSTM unit are as follows:

1. Input gate: This gate is used to control the input node information. The mathematical expressions of the input gate output and candidate information are as follows:

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \quad (3)$$

$$g_t = \tan h\left(U_g x_t + W_g h_{t-1} + b_g\right) \tag{4}$$

Among them, U_i, W_i, and b_i represent the weights and biases of input gates, U_g, W_g, and b_g represent the weights and biases of candidate states, σ represents the sigmoid activation function, and $\tan h$ is the activation function.

2. Forget gate: This gate is used to control which information is discarded by the current LSTM unit. The mathematical expression of the forget gate is as follows:

$$f_t = \sigma\left(U_f x_t + W_f h_{t-1} + b_f\right) \tag{5}$$

Among them, U_f, W_f, and b_f denote the weights and biases of the forget gates respectively, and σ represents the sigmoid activation function.

3. The memory unit (memory cell): is used to save the state information and update the state. The mathematical expression of the memory unit c is as follows:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{6}$$

Among them, $\odot$ represents the Hadamar product.

4. Output gate: The gate is used to output the control of node information. The mathematical expression of the initial output value and the output of the LSTM unit is:

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{7}$$

$$h_t = o_t \odot \tan h(c_t) \tag{8}$$

Among them, U_o W_o, and b_o represent the weight and bias of the output gate, respectively.

We input the pre-processed images into Mobilenets, extract the high-dimensional features of the images in fully connected layers, and input the features into LSTM model for learning the past and future information of the sequence features. In the memory unit of LSTM, putting in all the data passes through only one cell unit in different timing states. Also, it can reduce the number of parameters by keeping updating the weights. (See Figure 7) Among them, W(f), W(i), W(j), W(o) are weight parameters in the cell unit of LSTM. We aim to train these four weight parameters to optimize the LSTM network and reduce the input parameters.

The Dropout technique is used to mitigate the over-fit problems that have occurred during the training process. The Dropout technique reduces the complexity of the model by randomly dropping part of the neurons during each training process, thus improving the generalization ability of the model. In particular, it is assumed that a neural network with n nodes, in each training procedure, randomly discards the neurons in the network hidden layer at a probability p, and the probability of the retention of the neurons is 1–p. In general, this probability value p is set to 0.5 (referred to as the Dropout rate), since the randomly generated network structure is the most, that is, a set corresponding to 2n models. In addition, the joint action between the various neurons can be reduced, so that the appearance of a certain feature does not depend on the characteristic of the fixed relation, and can be used for weakening the interaction between the various features, so that the model is not too dependent on some local characteristics. Thus, the generalization ability of the model is enhanced.

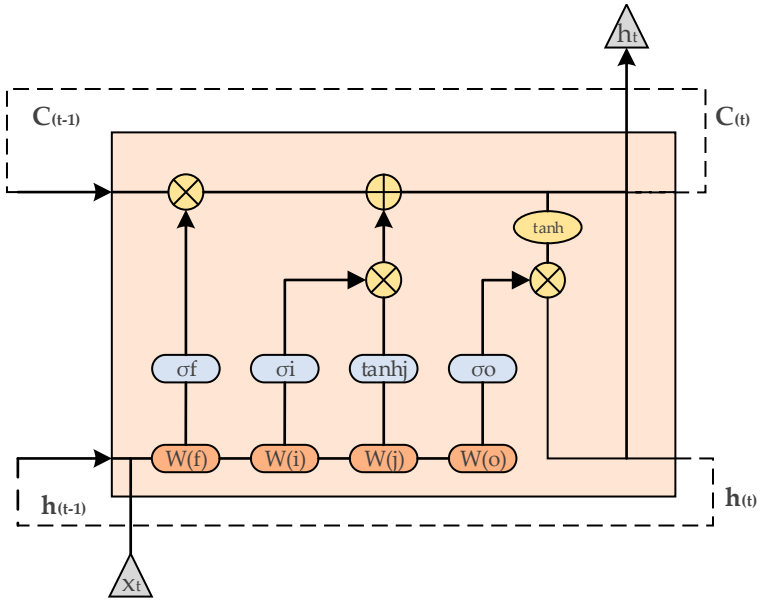


Figure 7. Neural network model based on weight analysis.

3. Experimental Dataset and Results

3.1. Dataset

The dataset in this paper is our self-made lip-language video database, recorded by six different producers (three men and three women) in a single constant environment. At this stage, it is worth emphasizing that privacy restrictions on datasets cause most of the data that we get to be Asian. The system has a good performance with Asian features. During recording, the head and the camera remain relatively static. The recorded content is the independent pronunciation of ten English words 0–9. Each person makes 100 sounds and divides them into different video clips. Then the sample size of the database is 6000. At the same time as we made the data enhancement to the original data, the dataset was expanded to 12,000 samples by increasing the light-and-dark, the image, the rotation, the Gaussian noise, the pepper and salt noise, etc. The original image has a resolution of 1920 × 1020, approximately 25 frames per second.

3.2. Results and Discussion

In this section, we evaluate the designed mobile lip reading recognition system, and analyze and compare the results on our dataset. We randomly disrupt the dataset and divide the training set and the test set according to 90% and 10%. We built MTCNN and LSTM networks with PyTorch. The random gradient drop method is used to train the network. The training model is inputted in 64 units. The learning rate of the first 100 iterations is 0.1, and then changed to 0.001 (in order to speed up the convergence rate).

We choose Raspberry Pi (Raspberry Pi 4, 4GB of LPDDR4 SDRAM, Dual monitor support, at resolutions up to 4K) based on the Linux system to realize dynamic lip reading recognition on the mobile end, as shown in Figure 8. Compared with the general PC computer platform, Raspberry Pi has the advantages of small size, low power consumption, and low cost. It can complete some tasks and applications that a PC platform can normally realize.

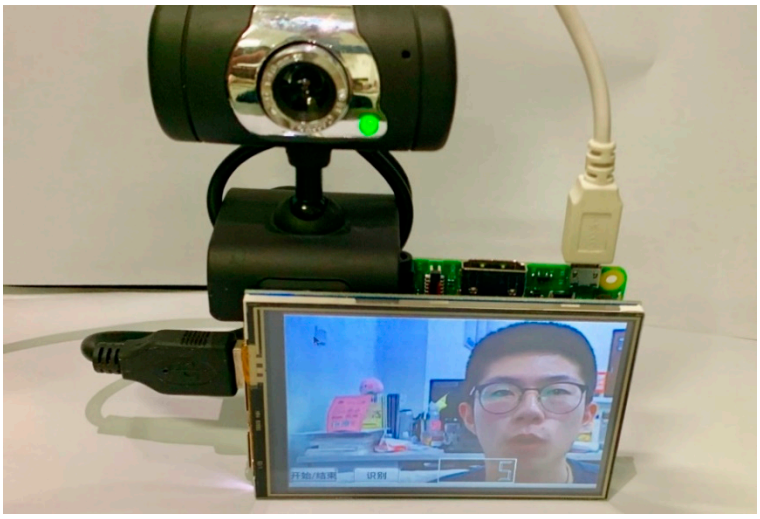


Figure 8. Physical photo of Lip Reading System on Raspberry Pi.

In order to evaluate the performance of the mobile lip reading system, we compared the mainstream research methods [28,29] through a large number of experiments, and the results are shown in Table 1. The proposed method can reduce a large number of parameters and reduce the complexity of the model and does not significantly degrade the performance of the model. We propose that the recognition time of the model is the time of video recognition, including decision- making. It can be seen that the recognition accuracy of the lightweight model proposed by us is smaller than that of the deep convolution hybrid network, and the recognition speed is greatly improved, which can meet the deployment and application of the mobile terminal.

Table 1. Performance comparison of the mainstream research methods.

Network	Accuracy	Time	Model Parameter (Million)
BiLSTM + AlexNet (No data expanded)	85.7%	10.0 s	61
AttentionLSTM+VGG16 (No data expanded)	88.2%	16.3 s	139
LSTM + MobileNets (Data expanded)	86.5%	7.3 s	5.2

The proposed system can be adapted well to the real environment without excessive degradation of model performance.

The training dataset and the test dataset are input into two MobileNets respectively, and then the sequence features of 4096 × 10 are extracted with the same LSTM model. Loss, accuracy, and recall of each period are shown in Figures 9 and 10.

In Figure 9, when the period (epochs) is about 19, the loss tends to be stable, indicating that the optimal solution has been reached at this time. The accuracy of the proposed network model in the test dataset is 86.5%. The test set performs very well, and the accuracy and loss eventually tend to balance, which shows that the spatial and temporal characteristics have been learnt.

As we want to identify the results as accurately as possible and reduce the situation of confusion recognition, we therefore pay more attention to the recall evaluation index of the model. Figure 10 below is the recall of our proposed model and the recall of the comparative model. It can be seen that our model performs well in pronunciation 2, 7, and 8, which is of great significance compared with previous studies. Considering the above experimental considerations, our research can be deployed well in the mobile terminal, and the efficiency is high.

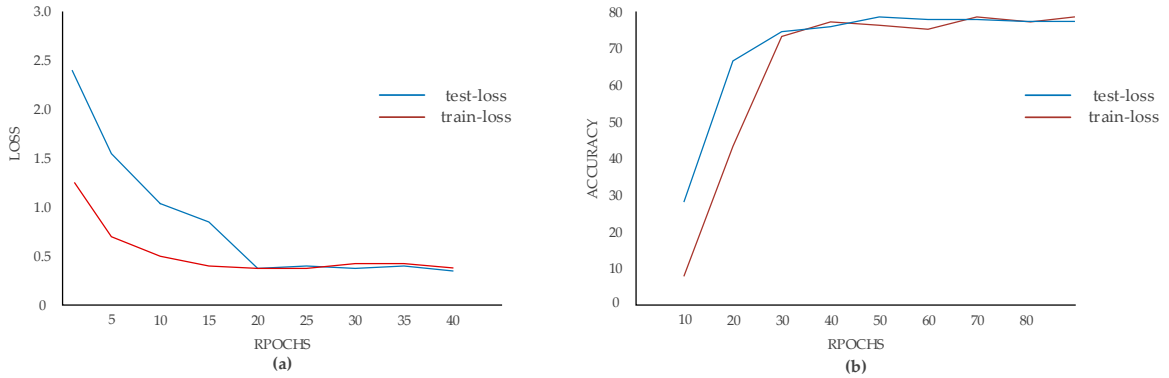


Figure 9. Comparison of our proposed model (**a**) Losses of each period in two networks. (**b**) Accuracy of each period in two networks.

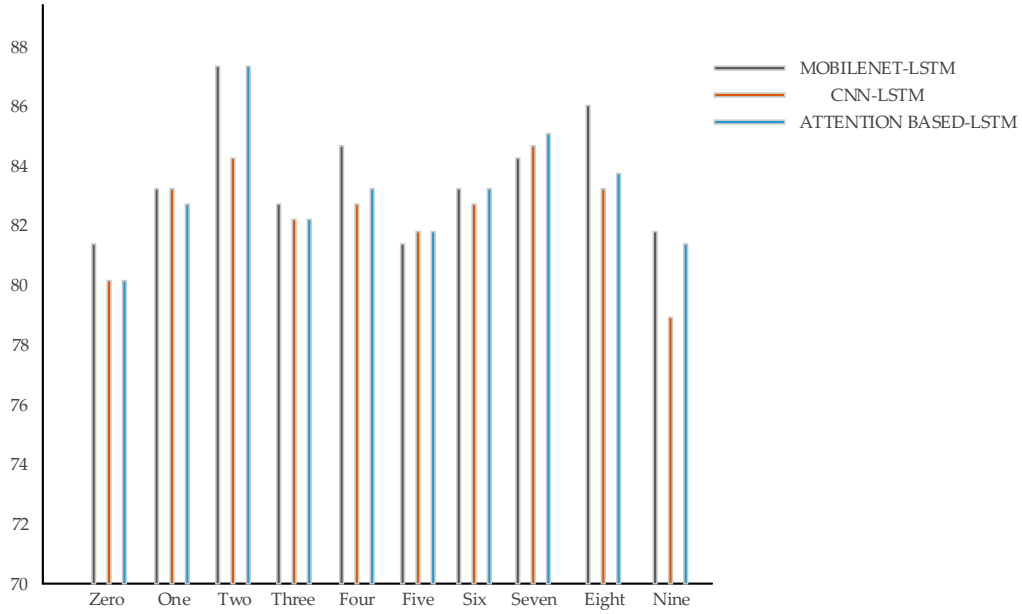


Figure 10. Recall of model and comparison model.

4. Conclusions

This paper concerns a lip language video obtained from Raspberry Pi. In order to optimize the recognition results and reduce redundant information, first, we extract the fixed-length frame sequence with our efficient and concise method, use MTCNN to correct the lip, and then use the lightweight MobileNets structure to train the model. Then, LSTM network is used to learn the sequence weights and sequence information between frame-level features. Finally, two full connection layers and one softmax layer are used to implement the classification. We independently established a dataset consisting of three men and three women. We recorded the pronunciation of English from 0 to 9. Each digital pronunciation was divided into independent video clips. We expanded the original dataset. Experimental results show that the mobile lip reading recognition system can effectively recognize words from the video, the complexity of the model is low, the amount of parameters has been reduced by 20 times, and the speed increased by 50%. This is the first mobile lip reading recognition system that uses a lightweight network in lip language research. It has reached the highest level of our research. We have also expanded the data to make it more versatile. However, our research of lip reading recognition is chronological, not aiming at a particular type of lip movement at a certain time. Therefore, the real-time performance is not good. In future research, we will focus on how to improve the speed of the recognition system based on time series and a train lip reading model on

news video datasets, including news video samples from different environments to test our designed recognition system. According to the extended research of VR in the future, we will be more proficient in deploying and naming algorithms of mobile devices, so as to add multi-dimensional input to the VR scenes. For saving space of mobile devices and speeding up the operation and data-sharing, we will try to transfer data from raspberry pi by 5G (5th generation mobile networks) and utilize a server for algorithm identification and then return to the mobile devices, adding interactive virtual sensing technology to enable a wide range of facial recognition applications.

Author Contributions: Data curation, Y.L. and J.W.; Formal analysis, Y.L. and J.W.; Methodology, Y.L. and J.W.; Project administration, Y.L.; Resources, Y.L.; Supervision, Y.L.; Validation, J.W.; Visualization, J.W.; Writing—original draft, J.W.; Writing—review and editing, Y.L. and J.W.

Funding: This research was supported by the National Natural Science Foundation of China (61571013), by the Beijing Natural Science Foundation of China (4143061), by the Science and Technology Development Program of Beijing Municipal Education Commission (KM201710009003) and by the Great Wall Scholar Reserved Talent Program of North China University of Technology (NCUT2017XN018013).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
2. Loomis, J.M.; Blascovich, J.J.; Beall, A.C. Immersive virtual environment technology as a basic research tool in psychology. *Behav. Res. Methods Instrum. Comput.* **1999**, *31*, 557–564. [CrossRef] [PubMed]
3. Hassanat, A.B. Visual passwords using automatic lip reading. *arXiv* **2014**, arXiv:1409.0924.
4. Thanda, A.; Venkatesan, S.M. Multi-task learning of deep neural networks for audio visual automatic speech recognition. *arXiv* **2017**, arXiv:1701.02477.
5. Biswas, A.; Sahu, P.K.; Chandra, M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *Int. J. Speech Technol.* **2016**, *19*, 159–171. [CrossRef]
6. Scanlon, P.; Reilly, R. Feature analysis for automatic speech reading. In Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing, Cannes, France, 3–5 October 2001; pp. 625–630.
7. Matthews, I.; Potamianos, G.; Neti, C.; Luettin, J. A comparison of model and transform-based visual features for audio-visual LVCSR. In Proceedings of the IEEE International Conference on Multimedia and Expo, Tokyo, Japan, 22–25 August 2001; pp. 825–828.
8. Aleksic, P.S.; Katsaggelos, A.K. Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 5, pp. 917–920.
9. Stork, D.G.; Wolff, G.; Levine, E. Neural network lip reading system for improved speech recognition. In Proceedings of the International Joint Conference on Neural Networks, Baltimore, MD, USA, 7–11 June 1992; pp. 285–295.
10. Shaikh, A.A.; Kumar, D.K.; Yau, W.C.; Azemin, M.C.; Gubbi, J. Lip reading using optical flow and support vector machines. In Proceedings of the 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; pp. 327–330.
11. Puviarasan, N.; Palanivel, S. Lip reading of hearing impaired persons using HMM. *Expert Syst. Appl.* **2011**, *38*, 4477–4481. [CrossRef]
12. Lu, Y.; Li, H. Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory. *Appl. Sci.* **2019**, *9*, 1599. [CrossRef]
13. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.0486.
14. Kamal, K.C.; Yin, Z.D.; Wu, M.Y.; Wu, Z.L. Depthwise separable convolution architectures for plant disease classification. *Comput. Electron. Agric.* **2019**, 8. [CrossRef]
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556v6.
16. Ma, M.; Wang, J. Multi-view Face Detection and Landmark Localization Based on MTCNN. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018.

17. Martins, A.; Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1614–1623.
18. Lu, Y.; Yan, J. Automatic Lip Reading Using Convolution Neural Network and Bi-directional Long Short-term Memory. *Int. J. Pattern Recognit. Artif. Intell.* **2019**. [CrossRef]
19. Edwin, J.; Greeshma, M.; Mithun Haridas, T.P.; Supriya, M.H. Face Recognition based Surveillance System Using FaceNet and MTCNN on Jetson TX2. In Proceedings of the International Conference on Advanced Computing & Communication Systems, Coimbatore, India, 15–16 March 2019.
20. Jia, X.; Gengming, Z. Joint Face detection and Facial Expression Recognition with MTCNN. In Proceedings of the 2017 4th International Conference on Information Science and Control Engineering, Changsha, China, 21–23 July 2017.
21. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
22. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
24. Chen, H.Y.; Su, C.Y. An Enhanced Hybrid MobileNet. In Proceedings of the IEEE International Conference on Awareness Science and Technology, Fukuoka, Japan, 19–21 September 2018.
25. Michele, A.; Colin, V.; Santika, D.D. Santika MobileNet Convolutional Neural Networks and Support Vector Machines for Palmprint Recognition. *Procedia Comput. Sci.* **2019**, *157*, 110–117. [CrossRef]
26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. In Proceedings of the 9th International Conference on Artificial Neural Networks: ICANN'99, Edinburgh, UK, 7–10 September 1999.
28. Weilin, Z.; Huilin, X.; Zhen, Y.; Tao, Z. Bi-directionallong short-term memory architecture for person re-identification with modified triplet embedding. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017.
29. Cho, K.; Courville, A.; Bengio, Y. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. Multimed.* **2015**, *17*, 1875–1886. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).