Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 5138442, 13 pages
https://doi.org/10.1155/2022/5138442

Research Article

Collaborative Research on Mouth Shape and Lyrics in Singing Practice Based on Image Processing

Lujia Xu[1] **and Chen Chen**[2]

[1]*Xiamen University Tan Kah Kee College, Zhangzhou 363123, China*
[2]*The First Affiliated Hospital of Xiamen University, Xiamen 361000, China*

Correspondence should be addressed to Chen Chen; g20041117@xmu.edu.cn

Received 22 November 2021; Revised 10 December 2021; Accepted 14 December 2021; Published 27 January 2022

Academic Editor: Baiyuan Ding

Image processing is a mainstream processing method. When people enjoy artists' singing videos, there will be a problem that the subtitles of the lyrics are out of sync with the singer's mouth shape. This problem needs to be solved using image processing technology, letting the computer realize lip-reading recognition function and correct the mouth shape and lyrics subtitles in the image according to the extracted lip-reading data, so that the mouth shape and lyrics in singing practice can be synchronized. Lip-reading information can effectively improve the accuracy of language cognition, save part of capital and manpower investment, and make viewers get a good audio-visual interactive experience. The results show the following: (1) After the UI test, the system user interface function design is reasonable and there is no bad BUG. We can find that the average processing time of each frame is 628 ms, the system performance evaluation is good, and the success rate can be as high as 98.80%. 0.36724 s is the average time for each step when the system processes the image. (2) The human image can basically identify the portrait area and lip area from various angles. (3) Compared with DCT and DWT, the recognition rate of the two cascade lip region feature extraction methods is improved by nearly 10%, and the feature vector dimension is reduced by nearly 65%. (4) Classify the mouth shape more finely and optimize the image of the tester's mouth shape to make the mouth shape closer to the standard mouth shape. (5) After systematic correction of mouth shape and subtitles, the success rate is higher than 90%. Finally, we can find that the running effect is good and the method has achieved high results, which can carry out the details of the next optimization work.

1. Introduction

With the spring breeze of science and technology sweeping the world, times have quietly changed dramatically. How to use the power of science and technology to make people's daily life more efficient and convenient is the most frequently considered problem nowadays. In recent years, people are using lip-reading more and more frequently, although they do not realize that they are using it. However, lip reading technology is no longer a "spring snow." It has gone out of the laboratory and gradually stabilized from little known to development and research. The relevant application technology is becoming more and more practical, and more and more people from outside are beginning to understand and learn its technology and characteristics. In this study, if we want to coordinate the mouth shape and lyrics in the video, the first step is to accurately locate people's lips in the image, and then a series of mouth shape recognition tasks are carried out. There are many literature and materials related to computer lip-reading recognition in existing journals, papers, and other places, and we select some of them for reference study here. Literature [1] introduces the reality and development level of lip reading, which arouses people's attention and interest. Literature [2] proposes algorithms for lip detection, feature extraction, and other technologies to realize a lip-reading prototype system. In [3], the face recognition technology is used in the attendance system. Literature [4] uses optical flow to analyze lip images, calculates two visual feature sets in each frame, and proposes a multimode speech recognition method. Literature [5] analyzes lip-reading technology according to typical deep learning and lists the existing lip-reading databases. Literature [6] uses a three-dimensional motion capture system to extract accurate parameters of facial motion features through lip reading. Literature [7] describes the public database of lip reading, target speech content, equipment, camera orientation, frame rate, and application. In [8], the coupled hidden Markov model is used to combine audio-visual signals with Polish speech recognition under the condition of highly interfered audio signals. Literature [9] studies the dual-mode fusion algorithm of video and audio, combining lip-reading speech recognition and information fusion technology. Literature [10] tests the fusion technology of different modes and analyzes different methods of audio-visual speech recognition. Literature [11] proposed the application of a lip-reading method based on a convolution neural network to tandem three-sequence key frame images. Literature [12] developed the deep convolution neural network model of HLR-Net for lip reading. Literature [13] proposes a new VSR method, multiangle lip reading, for audio-visual speech recognition. Literature [14] classifies those binoculars, enzymes, and muscle contractions in language when lip reading using multichannel SEMG signals. Literature [15] designs a graphic structure and lip partial cut network, using a local adjacent function extractor and lip reading with multilevel function fusion.

2. Theoretical Basis

2.1. Video and Audio Bimodal Corpus. Video and audio bimodal corpus [16] serves as the research basis. It will process the video according to the speech signal and visually include various face location images and mouth images related to the speech pronunciation. Because corpora involve many factors that need to be considered, such as format, coding, environment, and storage, and because there are few bimodal corpora that can be shared in the market, there are Tulips, M2VTS, ViaVoice, and other corpus databases in the world. There are the relevant parameters of the corpus, as shown in Table 1.

2.2. Lip Detection and Positioning Method. Different lip feature extraction methods obtain different lip areas [17]. If the first step of this positioning is not successful, all the functions based on lips will not be realized smoothly. The detection of lip area requires the system to quickly and accurately find the approximate range of the face in each frame of an image. After face recognition, it is necessary to accurately locate the lips according to their characteristics. In this way, we can ensure that after the previous work, we have laid a good foundation for lip feature extraction and make the subsequent image processing smoother and easier. The following is an introduction to typical lip detection methods.

2.2.1. Face Detection. Observing the facial features of human beings, there are three main human organs: eyes, nose, and mouth. Unless there are man-made or unexpected factors, their positions will not change much for a long time, and they have relative stability. Based on the stable features of face organs, we can use computer programs to detect the specific location of the face, and then use the Canny operator and projection method to quickly locate the lip area. W_{f} denotes the face width and H_{f} denotes the face height and the lip area of interest is as follows:

$$\begin{aligned} &\frac{1}{4}W_f \leq x \leq \frac{3}{4}W_f, \\ &\frac{2}{3}H_f \leq y \leq \frac{1}{15}H_f. \end{aligned} \tag{1}$$

However, due to the limitations of technology, time, and space, this detection method cannot accurately obtain and recognize the position of the face and specific lip area in some cases. This is because, in the process of singing, different singers have different faces, hairstyles, and postures. If there is a certain movement range (even sometimes there will be intense dance movements) to perform, it is impossible for the face to keep a positive and stable posture for detection at any time, and the detection efficiency is greatly reduced. Moreover, the relative position and size of the lip area will change with factors such as speech and expression. Therefore, this method is only suitable for preliminary and rough determination of lip area, which reduces some calculation work, but further accurate detection and confirmation are needed.

2.2.2. Color-Based Detection. According to the color distribution of the human face, we can find that the color of lips of most people is quite different from that of other areas of their face. Therefore, we can analyze the color intensity and color interval of lips according to different colors so as to correctly detect the parts of lips.

Because the red part of the face and lips account for different proportions. We usually choose to exclude the red gamut in the face by the red exclusion method, which is very classic and can effectively promote the solution of the problem. After reducing the influence of the red part on color detection, the situation represented by the original red area is replaced by blue and green. Here, we use the concept and meaning of three primary colors (RGB). This method can save the steps of color space conversion and tedious workload. There is also a method called the pseudocolor method, which is used to enhance the contrast between skin color and lip color. Finally, the image is converted to YIQ space, excluding the Y component, and good color discrimination can be seen. The formula is as follows:

$$\log\left(\frac{G}{B}\right) < \beta, \tag{2}$$

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.528 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{3}$$

Formula (2) is a discriminant, where G and B represent green and blue component values, respectively, which are

Table 1: Main parameters of corpus.

Bimodal corpus database	
Recognition primitive	Phonemes, words, sentences
Scale	Several to dozens, even hundreds of thousands
Number of people collecting corpus	From one to dozens
Visual channel information	Color or grayscale image, storage format, frame rate
Audio channel information	Background complexity, illumination, etc.

thresholds calculated by statistics. If the discriminant is satisfied, the pixel becomes a lip color pixel, otherwise it becomes a skin color pixel. Formula (3) is the formula of color intermediate conversion, which converts the image from RGB space to YIQ space, where Y is the gray component and I and Q are the chrominance components.

However, it should be noted that people have different skin tanning degrees, and each person's lips and skin color are different, so the specific colors need to be classified and discussed differently, which is not universal and brings more troubles to the specific recognition work. In addition, because singers will wear certain makeup because of social etiquette, work, and other factors, all kinds of lip gloss and cosmetics will change the color of their face, which greatly reduces the accuracy based on color detection.

2.2.3. Model-Based Detection. This method mainly uses the key points above the lip shape in the image to confirm the contour of the lip. Using the ACM algorithm, we can quickly build a matching lip model. The calculation degree of this method is very complex. If it is applied in practice, the burden of image processing will be very large, and it is difficult to correct mistakes in time once they go wrong. The relevant formula is as follows:

$$\begin{aligned} E_{\text{Snake}}(V) &= \sum_{i=1}^{N} E_{\text{Snake}}(i) = \sum_{i=1}^{N}\left(E_{\text{int}}(i) + E_{\text{ext}}(i)\right), \\ E_{\text{int}}(i) &= \alpha_i \left\| V_i - V_{i-1} \right\|^2 + \beta_i \left\| V_{i-1} - 2V_i + V_{i+1} \right\|^2, \\ E_{\text{ext}}(i) &= E_{\text{image}}(i) + E_{\text{con}}(i). \end{aligned} \tag{4}$$

A Viola–Jones method for lip region detection was established [18]. In this paper, we choose to fuse the abovementioned lip detection methods, and combine the face detection method with Viola–Jones method to learn from each other's strengths. A flow chart related to lip area detection is shown in Figure 1.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'), \tag{5}$$

$$s(x, y) = s(x, y-1) + i(x, y), \tag{6}$$

$$ii(x, y) = ii(x-1, y) + s(s, y). \tag{7}$$

In Figure 1, the Haar feature extraction module is mainly used. The value of integral image *ii* of any image *i* at any pixel *(x, y)* is defined as formula (8), and formula (9) and formula (10) can be obtained by calculation.

The training for AdaBoost is as follows:

$$h_i(x) = \left\{ 1, \quad \sum_{t=1}^{T} a_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} a_t, 0, \text{otherwise.} \right. \tag{8}$$

2.3. Image Processing Techniques. The digital image is represented by a two-dimensional array [19, 20]. Image processing technology has made more and more applications in recent years [21, 22]. Each pixel is unique, with unique plane position coordinates and values, and is the basic storage unit of digital images. In order to improve the clarity of the image, we can sharpen and enhance the image.

$$d(x, y) = f(x, y) - g(x, y). \tag{9}$$

The image is preprocessed. The color image is grayed [23]. Doing so allows you to have fewer pixels, less memory for your files, and less burden on your computer when it comes to processing images. The part formula of image grayscale is as follows:

$$\begin{aligned} \text{gray} &= R * 0.3 + G * 0.59 + B * 0.11, \\ \text{gray} &= G, \\ \text{gray} &= (R * 76 + G * 151 + B * 28) \gg 8, \\ \text{gray} &= \frac{(R * 30 + G * 0.59 + B * 11)}{100}. \end{aligned} \tag{10}$$

It means far greater than that. Four gray value formulas represent four different gray methods of images.

Binarization [24] makes the image black or white as follows:

$$g(x, y) = \begin{cases} 0\,(\text{gray value is less than threshold}\,(T)), \\ 255\,(\text{gray value is greater than threshold}\,(T)). \end{cases} \tag{11}$$

In the image binarization operation, T represents the gray value of pixels on the image. The confirmation of T is mainly as follows: when the gray level is higher than the threshold pixel, it is determined to be represented by the gray level value 255; otherwise, the gray value is 0, indicating the background or exceptional object area.

Take the Canny operator as an example, as shown in Figure 2:

Common convolution kernel templates are as follows:

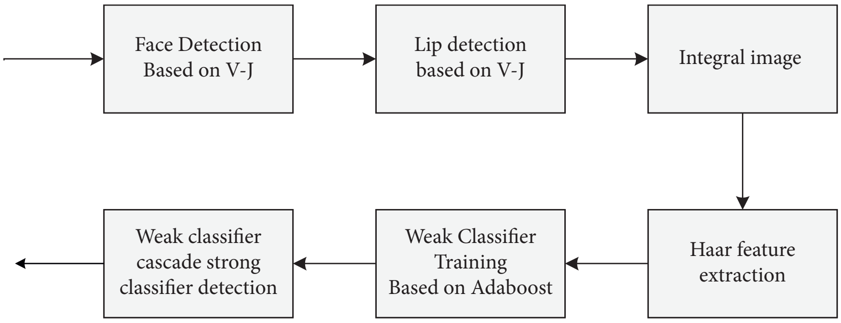
Face Detection Based on V-J
Lip detection based on V-J
Integral image
Haar feature extraction
Weak Classifier Training Based on Adaboost
Weak classifier cascade strong classifier detection

Figure 1: Viola–Jones lip detection flow chart. In the Haar rectangle feature.

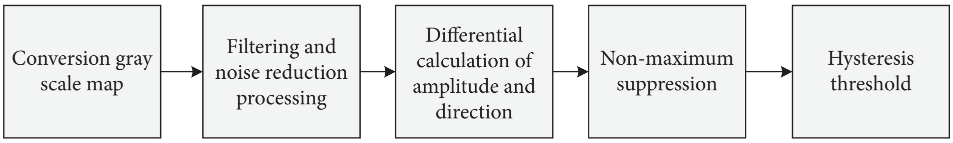
Conversion gray scale map
Filtering and noise reduction processing
Differential calculation of amplitude and direction
Non-maximum suppression
Hysteresis threshold

Figure 2: Edge detection flow of the Canny operator.

$$\begin{bmatrix} \frac{1}{16} & \frac{2}{16} & \frac{1}{16} \\ \frac{2}{16} & \frac{4}{16} & \frac{2}{16} \\ \frac{1}{16} & \frac{2}{16} & \frac{1}{16} \end{bmatrix}. \tag{12}$$

Calculate the magnitude and direction of the gradient, and the following are the x direction and y direction:

$$\begin{aligned} S_x &= \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \\ S_y &= \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \end{aligned} \tag{13}$$

Let the images H(i, j) and C be the gradient to be calculated.

$$H(i, j) = \begin{bmatrix} A_0 & A_1 & A_2 \\ A_3 & C & A_5 \\ A_6 & A_7 & A_8 \end{bmatrix}. \tag{14}$$

We can get gradients in the x and y directions, respectively, by using the following equation:

$$\begin{aligned} G_x &= 2 \times A_5 + A_2 + A_8 - (2 \times A_3 + A_0 + A_6), \\ G_y &= 2 \times A_7 + A_6 + A_8 - (2 \times A_1 + A_0 + A_2). \end{aligned} \tag{15}$$

At this point, the gradient amplitude and direction at point C are as follows:

$$\begin{aligned} G_{C(i,j)} &= \sqrt{G_x^2 - G_y^2}, \\ \theta &= \arctan\left(\frac{G_y}{G_x}\right). \end{aligned} \tag{16}$$

2.4. Lip Feature Extraction Method. The lip feature extraction method [25] is shown in Table 2.

In this paper, we design a lip feature extraction flow chart, as shown in Figure 3, to get the final feature vector.

3. Design of a Collaborative System of Mouth Shape and Lyrics

3.1. System Development Environment. In view of the realization of the function of coordinating the subtitles in the video with the singer's mouth shape, we designed a lip-reading recognition system. The system functions are mainly divided into four functions: lip detection, feature extraction, lip-reading recognition, and automatic subtitle correction. Adjusting the appearance time of lyrics in the video makes the mouth shape coordinate with the lyrics in singing practice. The specific development environment is shown in Table 3.

3.2. Lip Reading Based on the HMM Model. The process of the hidden Markov model is double stochastic. When people speak and sing, the staying time of each lip movement is different and it is difficult to determine. This model is consistent with the process of human lip movement, can describe the pronunciation state, and is widely used in lip reading applications. In this paper, the discrete hidden Markov model (DHMM) is adopted. As shown in Figure 4, it is the composition diagram of the HMM.

Considering that pronunciation has strong continuity in time in the actual lip reading system, this paper chooses the

Table 2: Comparison of three extraction methods.

Method	Typical algorithm	Advantages	Disadvantages
Shape-based method	Geometric feature method, snake model method, and active contour model method	Low feature dimension; it is not easy to change	The lip movement information reflected is not comprehensive; the image has clear edges
Pixel-based method	Method based on image transformation	All pixels represent visual information together, and the loss of information is relatively small	High feature dimension; image geometric transformation sensitivity
Hybrid methods	Method based on motion analysis	Combine shape-based and pixel-based features	The algorithm is complex; it is difficult to extract the ideal contour in processing

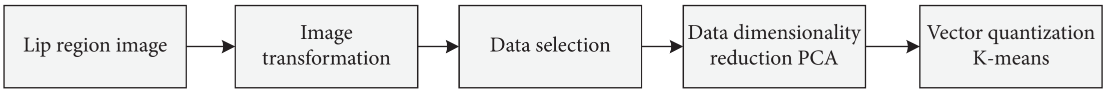
Lip region image
Image transformation
Data selection
Data dimensionality reduction PCA
Vector quantization K-means

Figure 3: Flow chart of cascade lip feature extraction.

Table 3: Lip reading system development environment.

Design interface	MFC class library
Computer	Intel CORE i7 8th gen
Development tools	Visual studio 2020 integrated development environment
Operating system	Windows 10
Function library	OpenCV, MFC

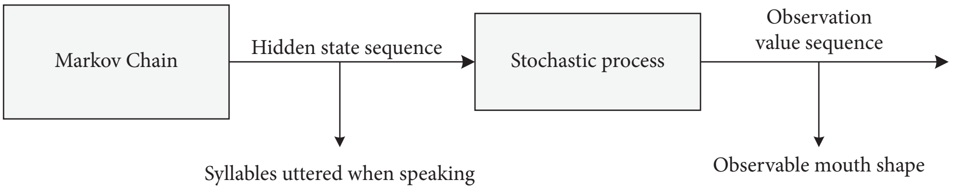

Figure 4: Schematic diagram of the HMM composition.

topological structure of a typical Markov chain from left to right without spanning, as shown in Figure 5.

The HMM model mainly has five parameters, which can determine the model. Reasonable selection of initialization parameters can increase the recognition rate of lip reading. Where B needs to be initialized, and N chooses a coefficient of 11.

$$\lambda = (M, N, \pi, A, B). \tag{17}$$

The values of π and A need to meet the following conditions:

$$\begin{gathered} 0 \leq \pi_i \leq 1, \\ \sum_{i=1}^{N} \pi_i = 1, \\ 0 \leq a_{\mathrm{ij}} \leq 1, \\ \sum_{j=1}^{N} a_{ij} = 1. \end{gathered} \tag{18}$$

$M = 4$, and the base mouth shape is shown in Figure 6.

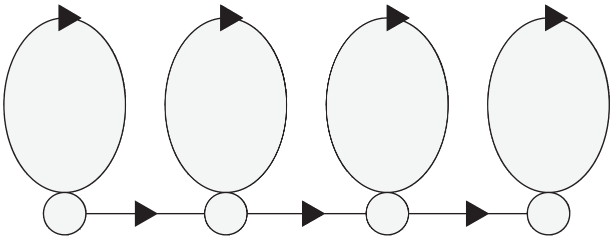

Figure 5: Markov chain structure.

3.3. Lip Reading Recognition and Correction. This research designs the interface module of the lip-reading system in which mouth shape and lyrics cooperate in singing practice. The whole interface strives to be concise and clear, and the functional distinction is clear. The first interface is designed as the placement area of each frame image of video interception, and all the specific processes of processing singing images are realized in this interface. The second interface button is designed as a lip area detection function. The third interface is the realization of the feature extraction function for the lip area. The fourth interface is the area where the image is finally recognized. After the recognition is successful, the data is automatically transmitted to the fifth

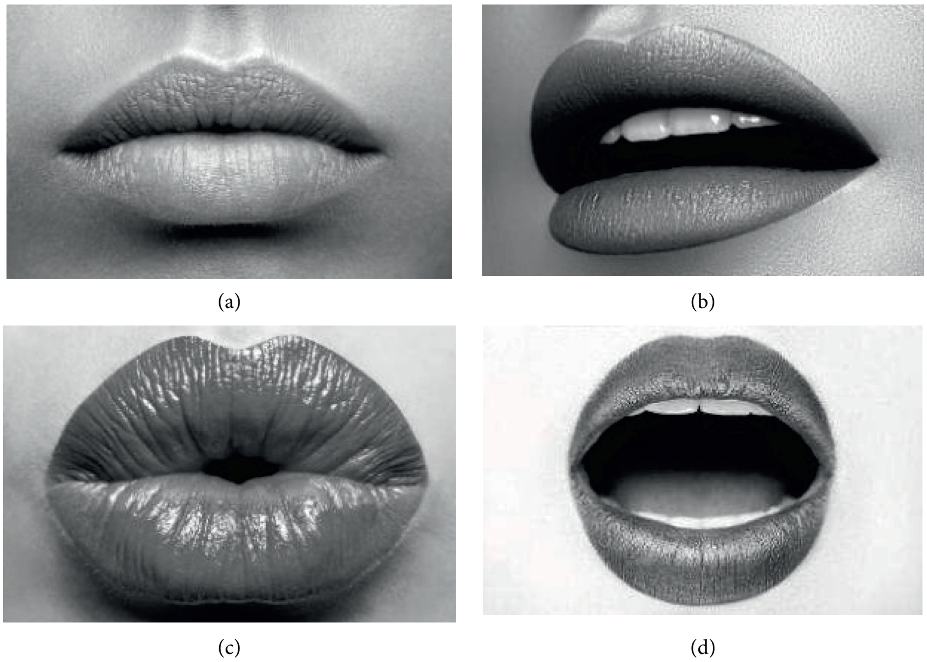

Figure 6: Four basic lip types of M: (a) shut up, (b) micro-opening, (c) pouting, and (d) open one's mouth.

interface to start the realization of automatic subtitle correction. At this time, the mouth shapes of the characters in the video will correspond to the subtitles, which successfully meets our experimental needs. The interface of the system is shown in Figure 7.

4. Experimental Data and Analysis

4.1. System Testing

4.1.1. UI Testing. The purpose of UI testing is to ensure that the user interface will provide users with appropriate access or browsing functions through the functions of the test objects. The content of the test is basically perfect. Check the rationality of the system operation interface designed in this paper, and the test results are shown in Table 4.

4.1.2. Performance Testing. For the experiment, the response within 2s shows that the system has the best performance; 5–10s shows that the system performance is average; if it exceeds 10s, the system performance is not good, and the system needs to be improved. The system response time is calculated as follows:

$$\text{time} = N_1 + N_2 + N_3 + N_4 + A_1 + A_3 + A_2. \tag{19}$$

The network transmission time is $N_1 + N_2 + N_3 + N_4$, the application server processing time is $A_1 + A_3$, and the database server processing time is A_2.

We select a video whose mouth shape and lyrics subtitles are not synchronized, intercept all the images for subtitle proofreading, and test the response time of the system to deal with a large number of mouth shape images at the same time. We tested the response time for 100, 200, 300, 400, and 500 frames, respectively, as shown in Figures 8 and 9.

From the specific case of system response time in Figure 8, we can find that the more image frames are requested to be processed, the longer the system response time is. The average processing time of each frame is 628 ms, and the system has good performance. From Figure 9, we can know that the success rate of system response is 98.80%, and the test results are satisfactory.

4.1.3. Running Time of Each Step. The running time of each system step designed in this paper is counted, as shown in Figure 10. After calculation, we can know that the image processing speed of each step is 0.36724 s, and the most time-consuming part is lip feature extraction.

4.2. Lip Area Inspection Test. The first step of the lip reading system is tested, and the images of characters from various angles are processed frame by frame for face recognition and lip region positioning. Figure 11 is a flowchart of lip region detection.

The part of the image is selected for space limitation, as shown in Figure 12. It shows face and lip recognition and detection renderings. Whether it is front or side, close-range or long-range, the system can basically accurately identify the portrait area and lip area.

4.3. Feature Extraction of the Lip Region

4.3.1. Two Cascade Methods. Using DCT-PCA and DWT-PCA, the dimension of the feature vector can be reduced by nearly 65%, and the effectiveness is greatly improved. According to the test, the recognition rate of these two

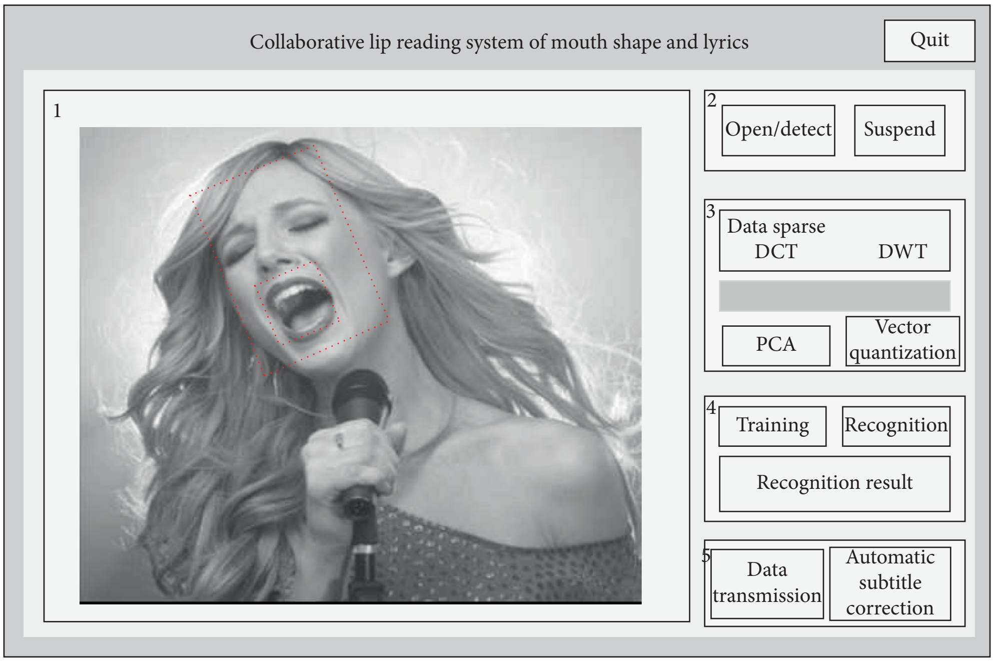

Figure 7: Schematic diagram of the system interface.

Table 4: UI testing.

Serial number	Test content	Test results
1	Interface layout	Normal
2	Text display	Normal
3	Font size	Normal
4	Garbled code	None
5	Hyperlink	Normal
6	Color style	Accord with
7	Shortcut key	Normal
8	Options button	Normal
9	Text box, dialog box	Normal
10	Clarity	Clear
11	Phenomenon of jamming or flashback	None
12	Keep a certain scale	Pass

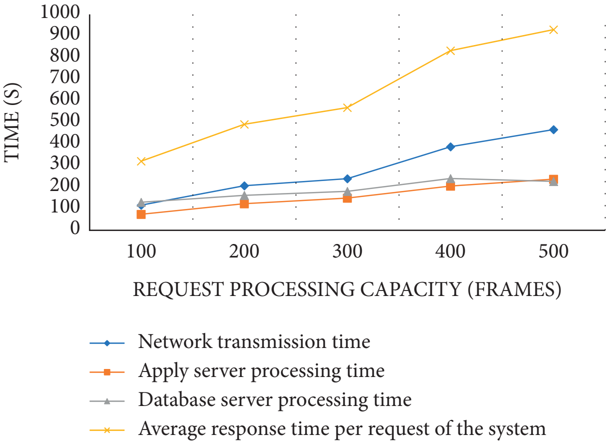

Figure 8: System response time.

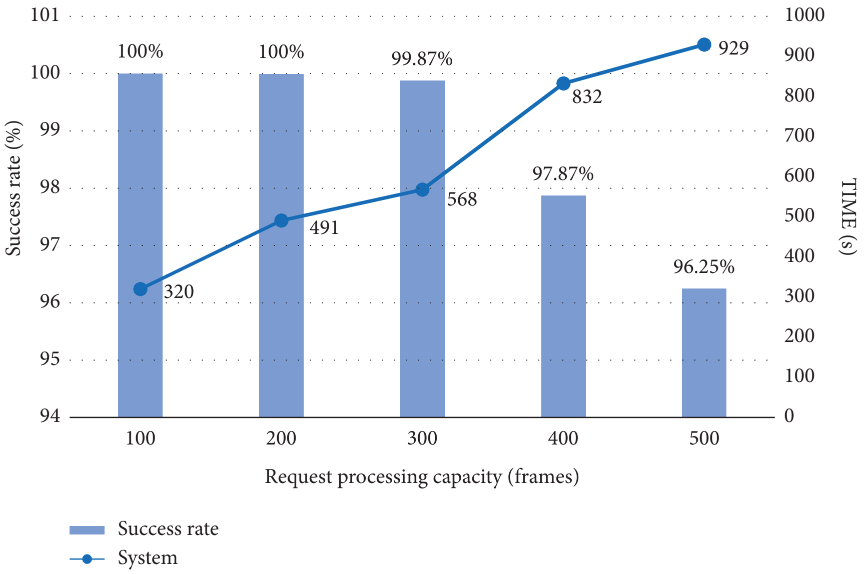

Figure 9: Success rate of the system response.

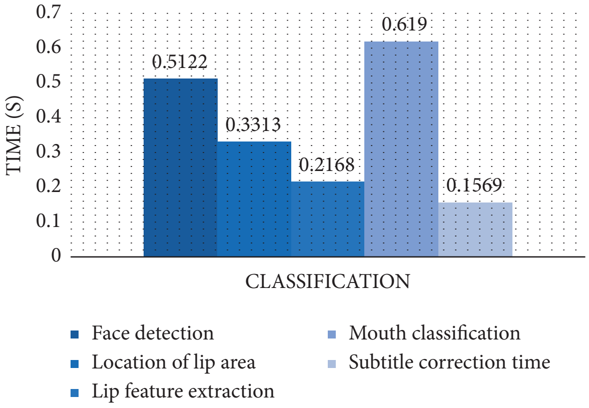

Figure 10: Running time statistics of each step.

methods is nearly 10% higher than that of DCT and DWT alone, as shown in Figure 13.

As shown in Figure 14, it is the change of recognition rate of two cascade methods under different feature dimensions. We can find from the figure that the higher the feature dimension, the higher the lip reading recognition rate will gradually increase within a certain range, and when it reaches a certain fixed-point peak, the recognition rate will gradually decrease slowly. This is because the number of samples in real life is limited, so when the feature dimension exceeds a certain number, it will lead to the decline of the lip reading recognition rate.

4.3.2. Labial Partial Experiment. As shown in Figure 15, the mouth shapes are classified more finely.

According to the image of human lips, lip templates are matched. Because everyone's lip shape looks different from the standard lip shape, it needs to be optimized when identifying. As shown in Figure 16, we take the O-mouth shape of an experimenter as an example. We can see that after optimization, the O-mouth shape is obviously improved, and it is easier to classify the lip shape.

4.4. Subtitle Correction. After automatic subtitle correction, the system allows the audience to experience scoring. The specific situation is shown in Figure 17. Only the correction of 10 frames of images is shown here. It can be seen that the success rate of subtitle correction is higher than 90%, with little fluctuation and good condition. However, the audience's scoring is more subjective and fluctuates more violently.

The success rate of subtitle correction judgment is higher than 90%, which is determined by objective factors. For the audience, no matter what the subtitle correction is, the

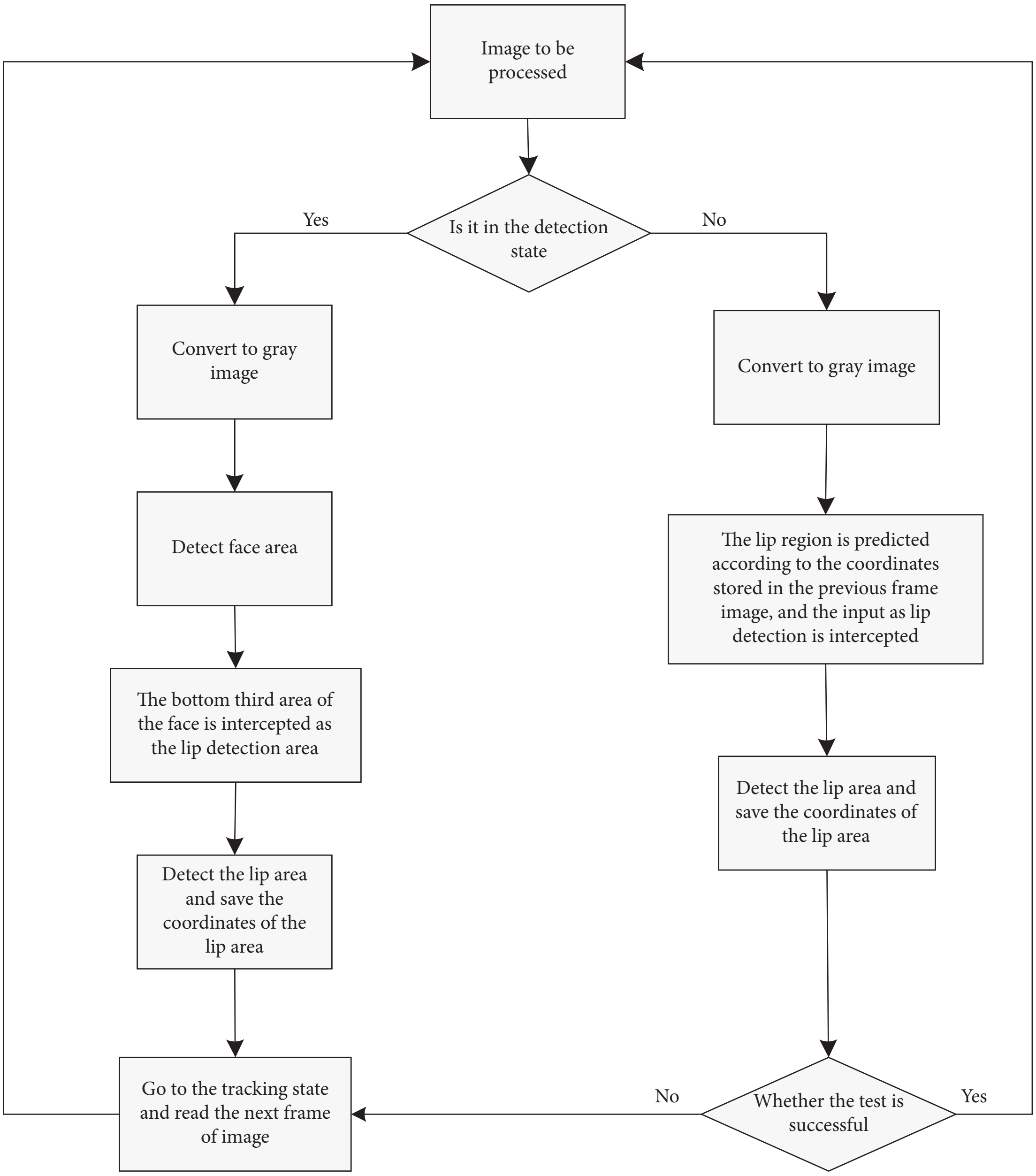
Image to be processed
Is it in the detection state
Yes
No
Convert to gray image
Detect face area
The bottom third area of the face is intercepted as the lip detection area
Detect the lip area and save the coordinates of the lip area
Go to the tracking state and read the next frame of image
Convert to gray image
The lip region is predicted according to the coordinates stored in the previous frame image, and the input as lip detection is intercepted
Detect the lip area and save the coordinates of the lip area
Whether the test is successful
No
Yes

Figure 11: Flow chart of lip area detection.

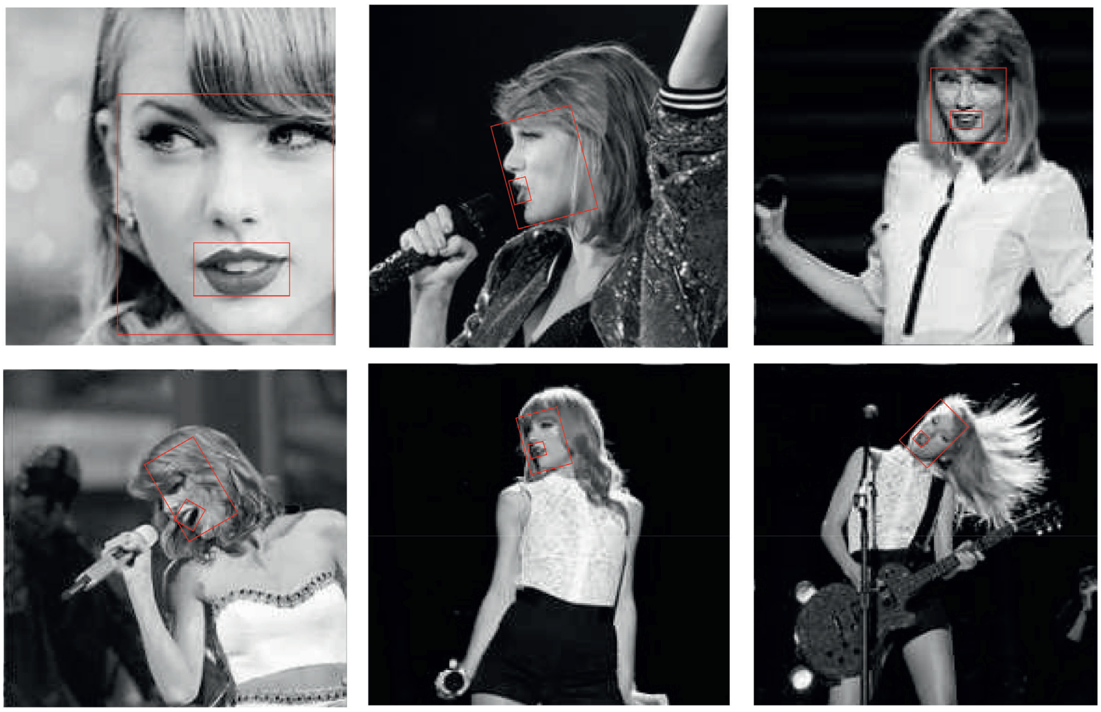

Figure 12: Recognition and detection renderings.

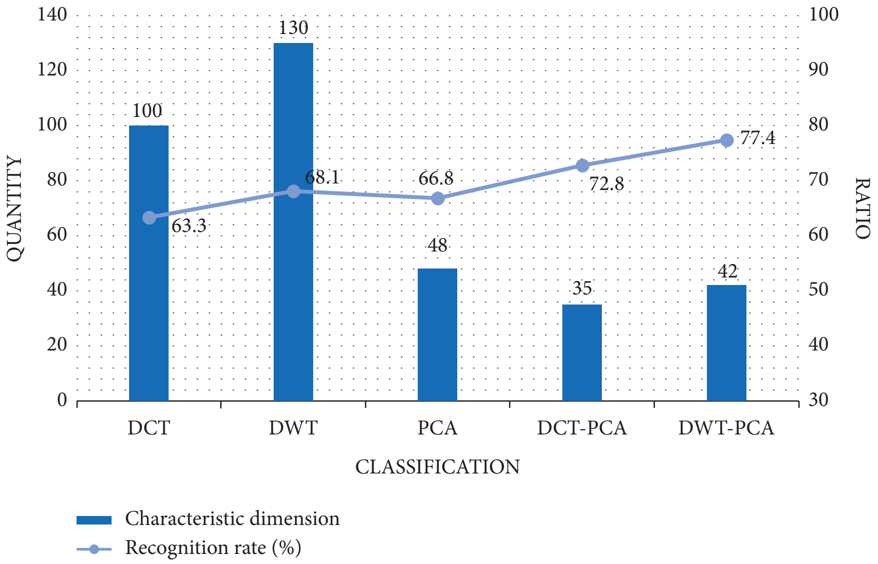

Figure 13: Optimal recognition rate of different features.

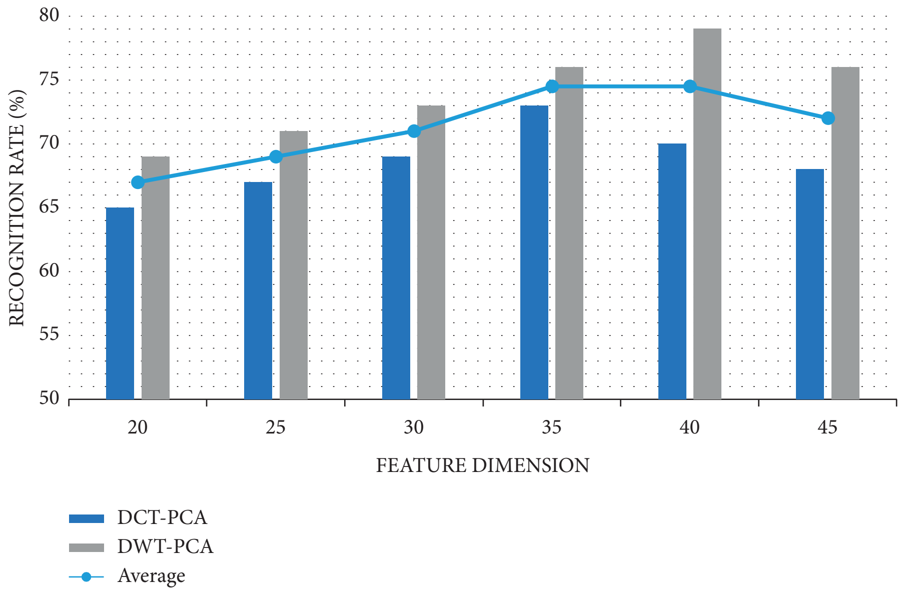

Figure 14: Comparison of recognition rates under different dimensions.

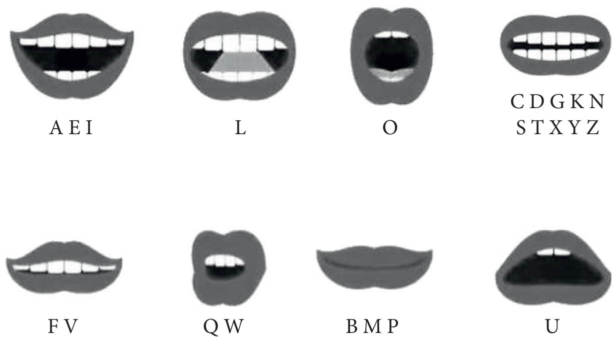

Figure 15: Classification according to the mouth shape of letters.

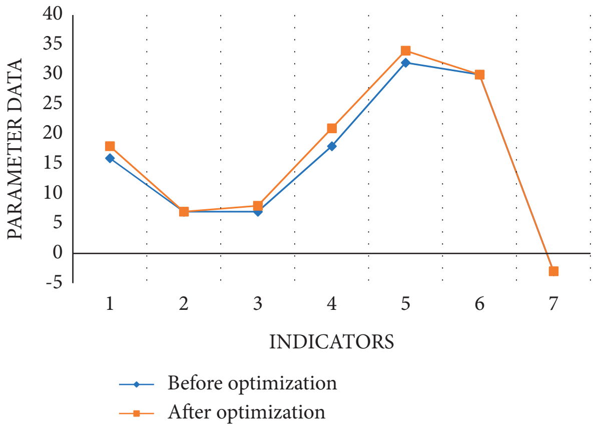

Figure 16: Optimization.

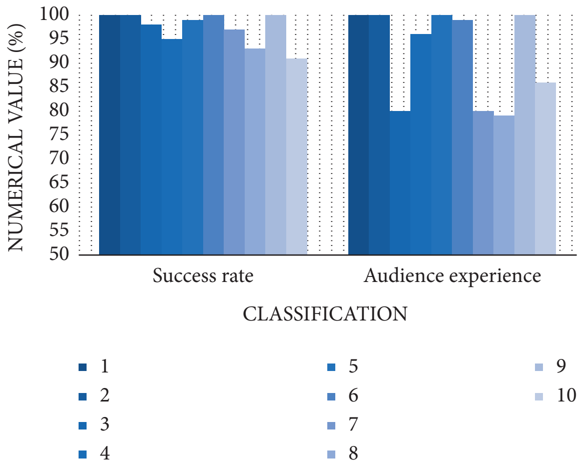

Figure 17: Subtitle correction diagram.

audience will subjectively judge their satisfaction with the experience. This results in a large difference between the two, which is inevitable.

5. Conclusion

Lip-reading technology is gradually moving out of the laboratory and into real life-applications, gradually integrating into people's lives without being perceived by them as its existence and role, which is its greatest success and advantage. In this paper, image processing technology is used to accurately locate the lip area for feature extraction. According to the data of the video and audio bimodal corpus, a lip reading model based on HMM is established, so that we can recognize the mouth shape accurately and adjust the lyrics synchronously with the singers in the video, thus effectively avoiding the bad experience of the audience's unsynchronized pronunciation and character in the process of watching and singing.

The research results show the following: (1) After a UI test, the functions of various operation interfaces are reasonable. The average processing time of each image frame is 628 ms. The system has good performance, and the success rate is as high as 98.80%. The test results are satisfactory. The image processing speed of each step is 0.36724 s, and the lip feature extraction part is the most time-consuming. (2) No matter the front or side, whether the character image is close-range or long-range, the system can basically accurately identify the portrait area and lip area. (3) The recognition rate of DCT-PCA and DWT-PCA is about 10% higher than that of DCT and DWT alone, and the dimension of the feature vector is about 65% lower than that of DCT and DWT alone. When the feature dimension exceeds a certain number, the recognition rate of lip reading will decrease. (4) In order to make the mouth shape of the tester closer to the standard mouth shape in recognition, some image optimization is carried out. (5) The success rate of subtitle correction is higher than 90%, and the audience's scoring is more subjective and fluctuating.

To sum up, the whole experimental results of this study ran well. The lip-reading work has achieved good results, but due to the limitations of funds, time, and technology, there are still more details to be solved in this study. For example, there are a few different types of mouth shapes, and more pronunciation and mouth shapes need to be matched in practical application; the detection of the lip area can also separate skin color and lip color more accurately; the speed of image processing by the model can be more efficient and fast [26–28].

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this study.

References

[1] H. X. Yao, W. Gao, and R. Wang, "A survey of lipreading——one of visual languages," *Acta Electronica Sinica*, vol. 29, no. 2, pp. 239–246, 2001.

[2] *Research on Lip reading Recognition Technology Based on Video Image by Tao Hong [D]*, Jiangsu University, Zhenjing, China, 2005.

[3] *Pan Finland's Research on Intelligent Attendance System Based on Face Recognition Technology*, pp. 31–40, Zhejiang University, Zhejiang, China, 2014.

[4] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 36, no. 2/3, pp. 117–124, 2004.

[5] M. Hao, M. Mamut, and N. Yadikar, "A survey of research on lipreading technology," *IEEE Access*, vol. 8, pp. 204518–204544, 2020.

[6] D. W. Kang, J. W. Seo, and J. S. Choi, "Monosyllable speech recognition through facial movement analysis," *Transactions of the Korean Institute of Electrical Engineers, A*, vol. 63, no. 6, pp. 813–819, 2014.

[7] T. Saitoh, "Lip reading Technical research report of electronic information and communication society," *Emm, High Technology Enrichment of Multimedia Information*, vol. 114, pp. 29–34, 2014.

[8] M. Kubanek, J. Bobulski, and L. Adrjanowicz, "Characteristics of the use of coupled hidden Markov models for audio-visual Polish speech recognition," *Bulletin of the Polish Academy of Sciences, Technical Sciences*, vol. 60, no. 2, pp. 307–316, 2012.

[9] L. J. Shi, P. Feng, J. Zhao, L. R. Wang, and N. Che, "Study on dual mode fusion method of video and audio," *Applied Mechanics and Materials*, vol. 734, pp. 412–415, 2015.

[10] I. Denis, R. Dmitry, and K. Alexey, "An experimental analysis of different approaches to audio–visual speech recognition and lip-reading," in *Proceedings of 15th International Conference on Electromechanics and Robotics "Zavalishin's Readings"*, vol. 32, no. 8, pp. 167–173, Kursk, Russia, September 2020.

[11] L. Poomhiran, P. Meesad, and S. Nuanmeesri, "Improving the recognition performance of lip reading using the concatenated three sequence keyframe image technique," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6986–6992, 2021.

[12] D. Hussein, D. M. Ibrahim, A. M. Sarhan, and N. M. Elshennawy, "HLR-net: a hybrid lip-reading model based on deep convolutional neural networks," *Cmc -Tech Science Press-*, vol. 68, no. 2, pp. 1531–1549, 2021.

[13] S. Isobe, S. Tamura, S. Hayamizu, M. Nose, and Y. Gotoh, "Multi-angle lipreading with angle classification-based feature extraction and its application to audio-visual speech recognition," *Future Internet*, vol. 13, no. 7, p. 182, 2021.

[14] J. Deny, R. R. Sudharsan, and E. M. Kumaran, "An orbicularis oris, buccinator, zygomaticus, and risorius muscle contraction classification for lip-reading during speech using sEMG signals on multi-channels," *International Journal of Speech Technology*, vol. 24, no. 3, 2021.

[15] C. Zhang and H. Zhao, "Lip reading using local-adjacent feature extractor and multi-level feature fusion," *Journal of Physics: Conference Series*, vol. 1883, no. 1, Article ID 012083, 2021.

[16] Y. Lu and J. Yan, "Automatic lip reading using convolution neural network and bidirectional long short-term memory,"

International Journal of Pattern Recognition and Artificial Intelligence, vol. 34, no. 1, pp. 2054003.1–2054003.14, 2020.

[17] J. A. Dargham, A. Chekima, and S. Omatu, "Lip detection by the use of neural networks," *Artificial Life and Robotics*, vol. 12, no. 1-2, pp. 301–306, 2008.

[18] S. K. Bandyopadhyay, "Lip contour detection techniques based on front view of face," *Journal of Global Research in Computer Science*, vol. 2, no. 5, 2011.

[19] R. C. Gonzalez and R. E. Woods, "Digital image processing," *Prentice Hall International*, vol. 28, no. 4, pp. 484–486, 2008.

[20] X. Ning, K. Gong, W. Li, and L. Zhang, "JWSAA: joint weak saliency and attention aware for person re-identification," *Neurocomputing*, vol. 453, pp. 801–811, 2021.

[21] L. Mcquillan, "Is lip-reading the secret to security," *Biometric Technology Today*, vol. 2019, no. 6, pp. 5–7, 2019.

[22] L. Zhang, W. Li, L. Yu, L Sun, and X Dong, "GmFace: An explicit function for face image representation," *Displays*, vol. 68, no. 1, Article ID 102022, 2021.

[23] J. C. Russ, J. R. Matey, and A. J. Mallinckrodt, "The image processing handbook," *Computers in Physics*, vol. 8, no. 2, p. 177, 1994.

[24] F. A. Kruse, A. B. Lefkoff, J. W. Boardman et al., "The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data," *American Institute of Physics*, vol. 283, pp. 192–201, 1993.

[25] Li Liu, G. Feng, and D. Beautemps, "Inner lips feature extraction based on CLNF with hybrid dynamic template for Cued Speech," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, 2017.

[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, p. 2421, 2006.

[27] I. A. Karpukhin and A. S. Konushin, "Constructing a speech audio–video corpus by aligning long segments of speech and text," *Moscow University Computational Mathematics and Cybernetics*, vol. 41, no. 2, pp. 97–103, 2017.

[28] N. Harte and E. Gillen, "TCD-TIMIT: an audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.