IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 4, August, 2021, pp. 1013-1021
Journal homepage: http://iieta.org/journals/ts

Extraction and Classification of Mouth Shape Features in Oral English Teaching Based on Image Processing

Qian Zhang[1], Liyan Xiao[2*], Yanfang Shi[3]

[1] Department of Public Courses Education, Hebei University of Chinese Medicine, Shijiazhuang 050200, China
[2] College of Foreign Language Education and International Business, Baoding University, Baoding 071000, China
[3] Department of Computer Application Engineering, Hebei Software Institute, Baoding 071000, China

Corresponding Author Email: xiaoliyan@bdu.edu.cn

https://doi.org/10.18280/ts.380411

Received: 5 April 2021
Accepted: 28 June 2021

Keywords:
oral English teaching, mouth shape feature extraction, mouth shape classification, image processing

ABSTRACT

Mouth shape identification helps oral English learners discover the features of their lip movements in English speaking, and correct their pronunciation more smoothly. So far, few scholars have applied image processing to identify mouth shape features of oral English learners. Most studies consider little about environmental factors, and ignore the changing mouth shape in pronunciation. Therefore, this paper explores the extraction and classification of mouth shape features in oral English teaching based on image processing. Firstly, an extraction and classification model were established for mouth shape features in oral English teaching. Then, the mouth shape images of oral English teaching were preprocessed. After that, the authors segmented the lips in oral English video frames based on neural network, extracted the lip boundaries from the said frames, and fitted them into curves. The proposed model was proved effective through experiments.

1. INTRODUCTION

Many groups of people have the need to improve the pronunciation of oral English, such as overseas students, emigrants to English-speaking countries, outbound tourists, and foreign company employees [1-4]. Anyone who wants to improve oral English must strengthen his/her imitation ability by practicing the mouth shapes of pronunciation [5-11]. With the aid of mouth shape identification technique, oral English learners can compare their mouth shapes of oral English with those of teachers, discover the features of their lip movements in English speaking, correct their pronunciation more smoothly, and better understand the relationship between mouth shapes and voices [12-20].

The existing studies are mostly about oral English teaching and pronunciation practice [21-24]. Yang [25] described and introduced how to apply speech matching analysis and identification to oral English learning on online language learning platforms. Their application scheme offers a new teaching model for local and online learners, and provides technical supports to cheaper and easier oral English learning, and largescale oral English evaluation. Ni [26] investigated Moodle-based online teaching model of oral English through human-computer interaction. Their research targets the students of two art majors under the international cooperation education framework of colleges. Chen [27] probed into the multimedia teaching methods and specific teaching designs for oral English based on computer technology, and introduced multimedia into oral English teaching to make classroom teaching more effective.

Fan [28] pointed out two critical problems in graduate oral English teaching: the teaching is inefficient; the oral English level of graduate students simply cannot meet the needs of international exchanges. Then, oral English teaching of graduate students was studied through questionnaire survey, interview, and statistical analysis. The main feature of language teaching was summarized as the demand for a good learning environment. To this end, a specific language environment should be created such that the students could scientifically and effectively improve the oral English teaching effect. Fang [29] detailed and discussed the significance of information technology (IT) to oral English teaching in colleges, and demonstrated the importance of IT to college oral English teaching. Taking college oral English teaching for example, Yu and Liao [30] explored the functions of FIF app in college English teaching, and verified that artificial intelligence (AI) helps to correct pronunciation, improve teaching effect and efficiency, boost the interest in online learning, and promote intelligent teaching of college oral English courses.

So far, few scholars have applied image processing to identify mouth shape features of oral English learners. The recognition effect of mouth shapes is mainly bottlenecked by uneven illumination, and the image acquisition hardware. These factors are rarely considered by scholars before. What is worse, the existing studies often ignore the changing mouth shape in pronunciation, and have limitation in the research of application scenarios of oral English teaching. To solve these problems, this paper explores the extraction and classification of mouth shape features in oral English teaching based on image processing. The main contents are as follows: (1) setting up an extraction and classification model for mouth shape features in oral English teaching; (2) preprocessing of mouth shape images of oral English teaching; (3) presenting the lip segmentation method in oral English video frames; (4) extracting the lip boundaries from oral English video frames, and fitting them into curves. The proposed model was proved effective through experiments.

2. EXTRACTION AND CLASSIFICATION MODEL AND IMAGE PREPROCESSING

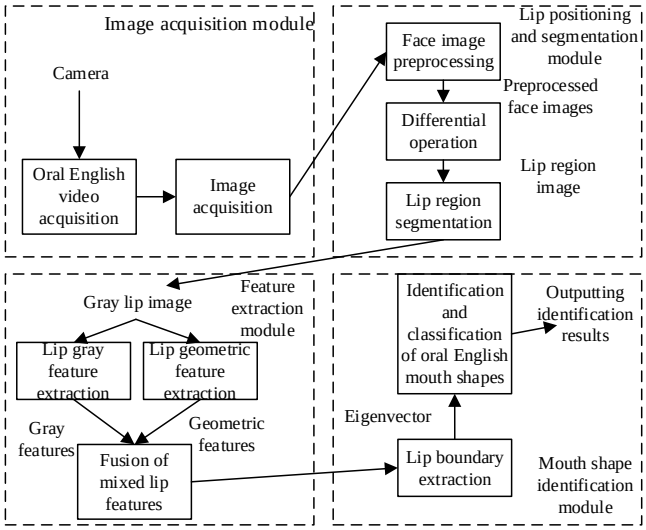


Figure 1. Framework of the extraction and classification model for mouth shape features in oral English teaching

Drawing on component programming, this paper divides the extraction and classification model for mouth shape features in oral English teaching into four modules: image acquisition, lip positioning and segmentation, feature extraction, and mouth shape identification (Figure 1). The preprocessing of learners' oral English video frames is an important step in extracting mouth shape feature of oral English teaching. This step mainly deals with two issues: detecting face image of each target, and position the lips in each face image.

The graying of face images is the starting point of image preprocessing. This paper adopts average graying: the red (R), green (G), and blue (B) channel components of each pixel in a color image equals the mean of the three components:

$$P_R = P_G = P_B = \left(P_R + P_G + P_B\right)/3 \tag{1}$$

The gray image needs to be binarized, i.e., be converted into a binary image characterized only by two preset values, which correspond to the target area and the background area, respectively. Suppose the gray values of the object in gray image $J(i, j)$ fall in the interval $[\psi_1, \psi_2]$. After binary thresholding, the binary image $E(i, j)$ can be expressed as:

$$E\left(i,j\right) = \begin{cases} 1 & \psi_1 \le J\left(i,j\right) \le \psi_2 \\ 0 & Otherwise \end{cases} \tag{2}$$

Energy normalization of images is necessary to facilitate the unified analysis of the energy of different mode samples, and reduce the influence of lighting on image feature extraction. In actual extraction of mouth shape features in oral English teaching, lighting intensity changes greatly affect the identification effect of learners' face and lips. For an $n \times m$ face image $FI(i, j)$ of the learner, the energy of the image can be defined as:

$$\left\|FI\right\| = \sqrt{\sum_{a=1}^{n}\sum_{b=1}^{m} FI^2\left(a,b\right)} \tag{3}$$

After energy normalization, image $FI^*(i, j)$ can be expressed as:

$$FI^*\left(a,b\right) = \frac{FI\left(a,b\right)}{\left\|FI\right\|} \tag{4}$$

In this paper, image smoothing is applied to small local areas of the image. For a local area centering on a pixel, the new value of the pixel can be obtained through two-dimensional (2D) discrete convolution of all pixel values in that area. Hence, the new value of the central pixel is closely correlated with the values of the other pixels in its neighborhood.

Let $FI(a, b)$ be the original face image of a learner; $G(a, b)$ be the smoothing operator; $n \times m$ be the size of the local area. Then, the 2D discrete convolution can be expressed as:

$$\begin{aligned} &FI\left(a,b\right) \otimes G\left(a,b\right) = \frac{1}{nm}\sum_{N=0}^{n-1}\sum_{M=0}^{m-1} FI\left(N,M\right)G\left(a-N,b-M\right) \\ &\begin{cases} a = 0,1,...,n-1 \\ b = 0,1,...,m-1 \end{cases} \end{aligned} \tag{5}$$

The averaging of all pixel values in a local area can adopt the weighted smoothing operator, which reflects the status of the central pixel. Let *Width* and *High* be the width and height of the original face image $FI(a, b)$ of a learner, respectively. For a pixel $h(N, M)$ in that image with $N<AL$ and $M<BR$, a neighborhood of the size $o \times o$ is taken with $h(N, M)$ as the center. Let $W(N, M)$ be the weighting coefficient of the neighborhood, and define $L=o/2$. Then, the weighted face image can be outputted as:

$$\hat{h}\left(N,M\right) = \frac{1}{o \times o}\sum_{a=N-L}^{N+L}\sum_{b=M-L}^{M=L} h\left(a,b\right)Q\left(N,M\right) \tag{6}$$

The smoothing eliminates most noises in learner's face image. The few remaining isolated noises can be removed with a median filter. If a pixel has a large gray value difference from its neighborhood pixels, the pixel value should be adjusted to a level closer to the gray values of its neighbors. Suppose the median value of a 3×3 neighborhood is *MF*. Then, the learner's face image after median filtering can be outputted as:

$$\tilde{h}\left(N,M\right) = MF\left(h\left(N-i,M-j\right), -1 \le i \le 1, -1 \le j \le 1\right) \tag{7}$$

During oral English teaching, the mouth shape continues to change. Therefore, this paper chooses to detect learner's lips through moving target detection. The moving direction of lips can be characterized by the trajectory detected by the accumulated differential image.

To identify the highly variable image region corresponding to the moving target, the differential image algorithm performs differential operation or subtraction between learner's face image and reference template.

Suppose the leaner's oral English video is a sequence of N_{IF} frames, where frame l is represented as $g_l(a, b)$. Let β_l be a weight coefficient that increases with l. Then, the accumulated differential image can be calculated by:

$$CDI\left(a,b\right) = \sum_{l=1}^{N_{IF}} \beta_l \left| g_1\left(a,b\right) = g_l\left(a,b\right) \right| \tag{8}$$

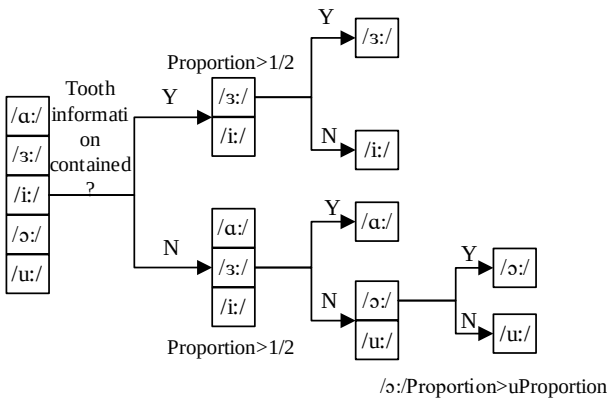


Figure 2. Flow of mouth shape identification of oral English teaching

To find a way to identify the lip shape features of different phonetic symbols, it is important to summarize the change law of lip shapes during the pronunciation of different long vowels. Figure 2 shows the flow of mouth shape identification of oral English teaching. After detecting lip boundaries, the lip shape of each long vowel can be identified accurately by judging whether the image contains tooth information.

Since the mouth shape features of oral English teaching depend greatly on lighting changes, this paper improves the above-mentioned differential detection algorithm. To effectively reduce the computing load of the model, dual differential operations were applied to the lower part of learner's face, in order to pinpoint the lip area. The lip motions can be described by parameters like the distance between left and right corners of the mouth, and the maximum distances from upper and lower lips to the centerline. To preserve all the lip information above, the height and width of the lip image are described as $Lips_H$=[2/5*High*, *High*], and $Lips_W$=[0, *Width*], respectively; the time interval is depicted as δ. Then, the specific flow of the proposed algorithm can be described by:

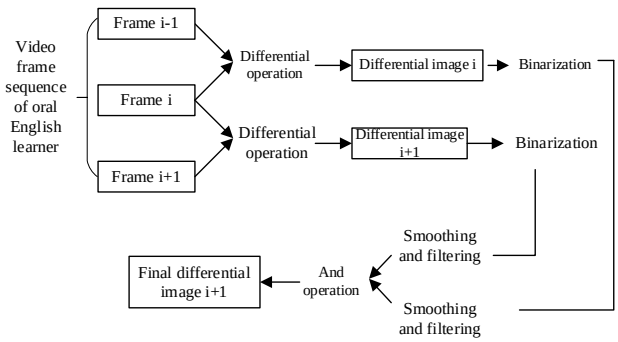


Figure 3. Algorithm flow

Figure 3 shows the flow of the improved differential algorithm. Firstly, differential operation is applied to frame $g_i(a, b)$ at time i in the video sequence of N_{IF} frames, producing the differential image h:

$$h_i(a,b) = \left| g_i(a,b) - g_{i-1}(a,b) \right| \tag{9}$$

Let Φ_i be the binarization threshold. The binary image h'_i after differential operation can be expressed as:

$$h'_i(a,b) = \begin{cases} 1 & if \left| g_{i+1}(a,b) - g_i(a,b) \right| \geq \Phi_i \\ 0 & else \end{cases} \tag{10}$$

Based on formulas (9) and (10), the mean AV_i and variance ε_i of differential image h_i can be derived. Let γ be the weight coefficient. The threshold $T\Phi_i$ can be calculated by:

$$\Phi_i = \gamma \cdot \varepsilon_i + AV_i \tag{11}$$

3. LIP SEGMENTATION

After positioning learner's lips in a frame, it is necessary to extract and identify the mouth shape features. Our research purpose is to recognize and classify the mouth shapes of oral English teaching based on the extracted characteristic parameters. The task of mouth shape extraction aims to recognize the few features, which best represent the pronunciation models of phonetic symbols, in high-dimensional data. To effectively segment lip color from skin color, this paper applies a neural network to realize the binary classification of learner's face.

The artificial neural network contains multiple nodes reflecting its internal state. Each node can process all input signals with an activation function. In general, a multi-input, single-output activation function involves two operations: weighted summation, and output processing.

Let a_i(i=1, 2, …, m) be the input signals applied to input layer nodes; θ_i be the connection weights of these nodes, which can be viewed as a proportionality factor that simulates the information transmission intensity of the nodes; ω be the weight of nodes; *EF* be the activation function of nodes. Then, we have:

$$\begin{cases} EF = \sum_{i=1}^{m} \theta_i a_i - \omega \\ b = \varepsilon(DF) \end{cases} \tag{12}$$

There are three kinds of activation functions for nodes: linear unit, threshold unit, and nonlinear unit. The response function of linear unit can be defined as an identity function $p(a)=a$. Then, the response function of threshold unit can be defined as a binary step function:

$$p(a) = \begin{cases} 1 & a \geq \omega \\ 0 & a < \omega \end{cases} \tag{13}$$

The response function of nonlinear unit can be defined as the following logarithmic tangent function and hyperbolic tangent function:

$$\begin{cases} p(a) = \dfrac{1}{1+e^{-\varepsilon a}} \\ p(a) = \dfrac{1-e^{-\varepsilon a}}{1+e^{-\varepsilon a}} \end{cases} \tag{14}$$

Neural network learning follows two kinds of rules: Hebb rule and Delta rule. Let $\Delta\theta_{ij}$ be the variation of connection weight; β be the learning coefficient; u_i be the activity of the current node. Then, the connection weight between two active nodes can be enhanced:

$$\Delta\theta_{ij} = \beta u_i u_j \tag{15}$$

Let *error* be the error between actual output and expected

output of the network. By adjusting the connection weight based on output error, it is possible to obtain a special set of weight coefficient and a special vector:

$$\Delta\theta_{ij} = \beta\xi_j u_i, \xi_j = K\left(b_i - \bar{b}_i\right) = K\left(error\right) \tag{16}$$

For a self-organizing competitive artificial neural network, the basic idea is that the nodes on network competition layer compete for the response to the classification of input samples. The winning node will output the final result of the network. To obtain the features of sample set and classification threshold, this paper needs to construct a neural network capable of classifying the skin color and lip color samples in learner face images, laying a good basis for the subsequent boundary extraction and curve fitting of binary lip images. Comparing various neural networks for image processing and target detection, it is learned that the competitive artificial neural network is the most suitable tool for segmenting the areas with distinctive features, such as lip color and skin color. Therefore, this network was selected to train the classification between skin color and lip color of learner face images.

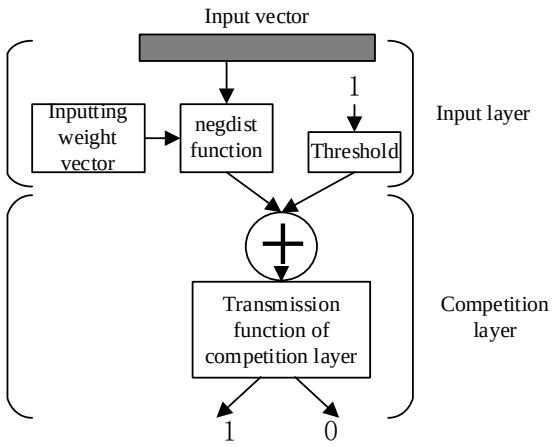


Figure 4. Structure of self-organizing competitive artificial neural network

Figure 4 shows the structure of the self-organizing competitive artificial neural network. Suppose the network contains m_{in} input layer nodes, and m_{out} output layer nodes. Since the output is a 2D plane, the output layer nodes should be as many as possible. Hence, m_{out} must be far greater than m_{in}. Then, an s function was defined to characterize the matching degree between input sample vector and connection weight θ_j(j=1, 2…, m_{out}). Let $\|A\text{-}\theta_j\|$ be the distance from the input vector A to the corresponding connection weight vector θ_j. Then, the function can be described by Euclidean distance:

$$s\left(A, \theta_j\right) = \left\|A - \theta_j\right\| \tag{17}$$

Suppose the d-th output layer node is the best match with A:

$$s\left(A - \theta_d\right) = MIND\left(A - \theta_j\right) \tag{18}$$

The output corresponding to θ_d can be described by b_d=MAX b_i. The weight adjustment is implemented on θ_d and θ_j in $b_i \in M_d$. The specific steps are described as follows:

Step 1. Weight initialization

Assign a small random value to the initial connection weight θ_j=θ_j(0)(j=0, …, m_{out}), when the number of iterations p equals 0.

Step 2. Matching degree calculation

From the frame sample set A_1, A, …, A_T, a random frame sample is selected as the input of the proposed network, and to compute the value of s function:

$$s_j = \left\|A^T - \theta_j\right\| = \sum_{i=1}^{m}\left(a_i^T - \theta_{ij}\right)^2 \tag{19}$$

where, j=1, 2, …, m_{in}. Among all s function values, the j-th value is the maximum s_{max}. Then, the corresponding b_j value is the maximum. The serial number d of the winning node also equals j.

Step 3. Weight adjustment

Let $M_d(p)$ be the modified area. Then, the connection weight θ_j can be adjusted by:

$$\begin{cases} \theta_j(p+1) = \theta_j(p) + \beta(p)\left(A - \theta_j(p)\right) & i \in M_d(p) \\ \theta_j(p+1) = \theta_j(p) & i \in M_d(p) \end{cases} \tag{20}$$

where,

$$\beta(p) = 0.9\left(1 - \frac{p}{\Phi}\right) \tag{21}$$

Suppose the square or hexagonal modification approximates the center point j of $M_d(p)$. The initial area is roughly half of the output 2D plane. Subsequently, the area of $M_d(p)$ will interactively reduce by:

$$M_d(p) = k_1 + k_2 e^{-\tau/d} \tag{22}$$

where, k_1, k_2 and τ are all constants.

Step 4. Add 1 to the number of iterations, i.e., p=p+1. Return to Step 2 to start another round of iteration, until the relationship between the input sample and the active nodes on the output 2D plane is relatively stable, or the number of iterations reaches the preset maximum number of iterations.

4. LIP BOUNDARY EXTRACTION AND CURVE FITTING

Red is the dominant color of lips and skin in the lower part of the learner's face image. Before classification training of frame set, this paper reduces the color space of the face image from three dimensions (R, G, and B) to two dimensions (G and B) by removing the red color.

Based on contour extraction, this paper extracts the lip boundaries from learner's face image through the following steps.

Step 1. Some of the lip contour pixels are identified by the detection rules. Based on the features of the identified contour pixels, the other pixels of lip contours are found following the tracking rules. Then, the boundaries are extracted from the binary image (Figure 5).

The detection rules are morphological gradient algorithm, which targets the binary image of learner's face. The boundaries are extracted mainly based on morphological gradient, due to the obvious boundaries between lips and

background in the binary image. Let FI_E be the binary image of learner's face to be processed; ST be the corresponding image structural element. Then, the most basic expression of morphological gradient can be given by:

$$MG_1 = (FI_E \circ ST) - (FI_E \bullet ST) \tag{23}$$

Fine boundaries can be further obtained based on the equivalent morphological gradient:

$$\begin{cases} MG_2 = (FI_E \circ ST) - FI_E \\ MG_3 = FI_E - (FI_E \bullet ST) \end{cases} \tag{24}$$

The binary image is detected based on the morphological gradient of formula (26). Corrosion can reduce the target area range in the image. In essence, this operation causes the boundary contraction of the image, thereby eliminating small and meaningless targets. This paper inverts the learner's face image, and then uses a 3×3 structural element to corrode the image, aiming to eliminate lip contours or boundaries. After that, the obtained corroded image was subtracted from the original image to obtain the desired lip boundaries (Figure 5).

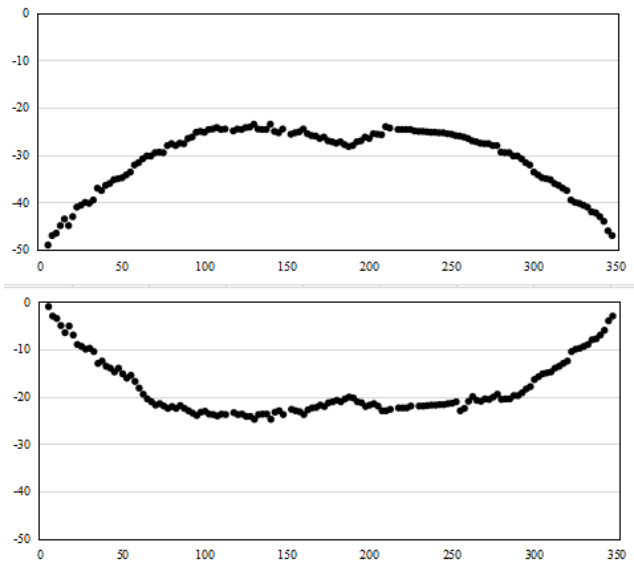


Figure 5. Lip boundary sample points

Finally, the lip boundary points are fitted based on a self-designed radial basis function (RBF) network. As a classic local approximation network, the RBF network only needs to adjust some connection weights and thresholds for each training sample. That is why it trains and converges faster than other neural networks. Figure 6 presents the structure of the RBF network. The hidden layer of the RBF network encompasses radial bases. Hence, the input vector is directly mapped to the hidden layer space, without needing connecting weights. Compared with other neural networks, the RBF network consume very little time, and require very few nodes to complete functional approximation. Let R be the activation function of the RBF network; a_q be the output of the q-th input layer node; z_i be the input of the i-th hidden layer node. Then, we have:

$$R(a_q - z_i) = e^{-\frac{\|a_q - z_i\|^2}{2error^2}} \tag{25}$$

Let υ_{ij} be the connection weight between hidden layer and output layer; N_{HL} be the number of hidden layer nodes. Then, the network output can be given by:

$$b_j = \sum_{i=1}^{N_{HL}} \upsilon_{ij} e^{-\frac{\|a_q - z_i\|^2}{2error^2}} \tag{26}$$

Let D_j be the desired output. The least squares loss function can be defined as:

$$error = \frac{1}{Q} \sum_{j=1}^{M} \|D_j - b_j z_i\|^2 \tag{27}$$

The RFB network, which includes an RBF hidden layer and a linear output layer, belongs to feedforward backpropagation network. Taking the coordinates of lip boundary points as network input vector, the RBF network goes through plotting training for less than 1.5s. In this way, the lip boundary curves can be fitted.

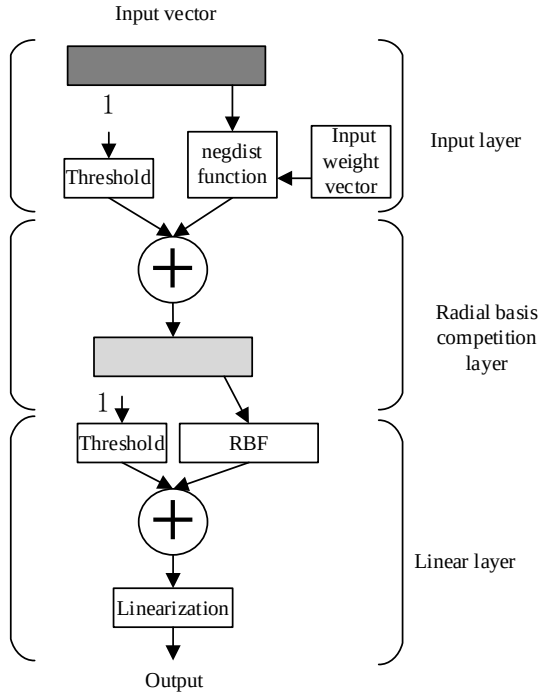


Figure 6. Model structure of RBF network

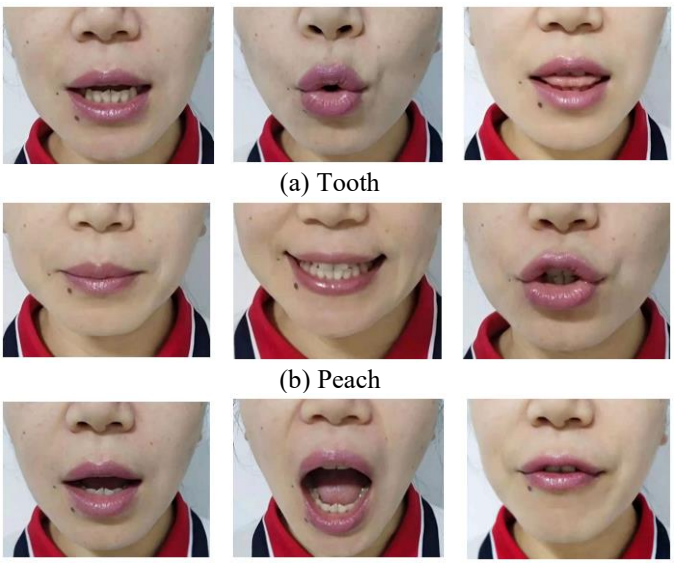
5. EXPERIMENTS AND RESULTS ANALYSIS

In our extraction and classification experiments of mouth shape features in oral English teaching, the mouth shape dataset includes the mouth shapes of 50 people, including 10 teachers and 40 learners, pronouncing words (tooth, peach, half, etc.) containing five long vowels, namely, /ɑ:/, /ɜ:/, /i:/, /ɔ:/ and /u:/ (Figure 7).

Firstly, eigenvectors were extracted from all mouth shape images, and used to establish a neural network. The network model was trained and applied to classify mouth shapes. The video frames of oral English teaching in the dataset were shot at different moments. These frames differ in lighting change, head deviation, and presence/absence of beards. For each word, 3/4 of the samples of each vowel were applied to train the

model, and 1/4 were adopted to test the model.

The mouth shape recognition effect and training speed of the proposed RBF network depend greatly on the number of hidden layer nodes. To determine the best number, different numbers of hidden layer nodes were compared through experiments. The results of the contrastive experiments are listed in Table 1.



(a) Tooth

(b) Peach

(c) Half

Figure 7. Mouth shape images of different words in in oral English teaching

As shown in Table 1, proper increase of hidden layer nodes could improve classification accuracy to a certain extent, but increases network training time and model overhead. Given fully consideration to training time and classification accuracy, it is suitable to design a hidden layer with 38 nodes. Figure 8 shows the training error curve of the neural network. Obviously, the network training lasted a short time, signifying the feasibility of our network.

Table 1. Experimental results under different number of hidden layer nodes

Number	Training time	Classification accuracy
26	6.85	78.5%
31	7.62	75.6%
33	7.53	81.7%
36	8.64	85.2%
38	7.39	92.9%
42	8.56	87.6%
43	8.84	85.4%
45	9.25	86.3%

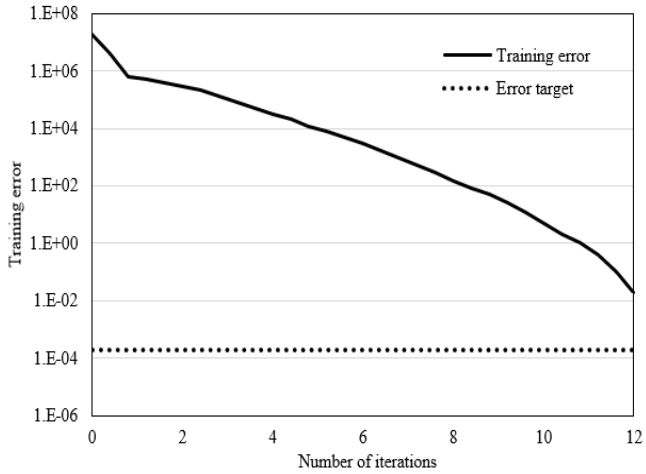


Figure 8. Training error curve of neural network

Table 2. Characteristic parameters of long vowel pronunciations

	Long vowel	Geometric features				Shape features
	/ɑ:/	1.2354	-0.3127	-1.8549	0.4551	0.7519
	/ɜ:/	1.2675	-0.3642	-2.0753	0.5236	0.8475
Mouth shapes in group 1	/i:/	0.6273	-0.1356	-0.1028	0.2984	0.4354
	/ɔ:/	1.3545	-0.7431	-2.1865	0.5286	0.822
	/u:/	0.7119	-0.1275	-1.1174	0.2483	0.4935
	Long vowel	1.1552	-0.2738	-1.7596	0.4351	0.7153
	/ɑ:/	1.3615	-0.3511	-2.0155	0.5176	0.8769
Mouth shapes in group 2	/ɜ:/	0.7589	-0.1825	-1.1954	0.2863	0.4725
	/i:/	1.2065	-0.3292	-2.1683	0.5656	0.8225
	/ɔ:/	0.6433	-0.1196	-0.1948	0.2364	0.4414

Further experiments were carried out following the steps of extracting mouth shape features in oral English teaching. Table 2 lists the geometric and shape features obtained for long vowel pronunciations. It can be observed that, when the same target pronounces the same long vowel, the eigenvalues of video frames should be similar, but differ greatly from the eigenvalues of the phonetic symbols of other words. The results confirm that our algorithm can extract the effective features from mouth shape images in oral English teaching, and correctly distinguish between the lip shapes of the phonetic symbols of other words, providing a reference for subsequent identification and classification of mouth shapes of oral English learners.

Table 3 presents a long vowel classification matrix with 38 hidden layer nodes. For each word, there are 65 frames. In this case, 92.3% of the mouth shapes are correctly classified. In the matrix, each row and column are the classification results of actual mouth shapes and recognized mouth shapes, respectively. Table 4 compares the experimental results on models trained with single-person mouth shape images and multi-people mouth shape images. Compared with template matching, our model achieved much better performance, especially when the training is performed on multi-people mouth shape images.

Based on the above experimental results, this paper further verifies the effectiveness of our model. A set of frames was randomly extracted from the test sets on five long vowels: /ɑ:/, /ɜ:/, /i:/, /ɔ:/ and /u:/. Then, the lip shape of each long vowel was compared with that of every other long vowel and 7 short vowels. The vowels /ɑ:/, /ɜ:/, /i:/, /ɔ:/, /u:/, /æ/, /ʌ/, /ɒ/, /ə/, /ɪ/,

/ʊ/ and /e/ were numbered 1-12 in turn. The experimental results are given in Figure 9.

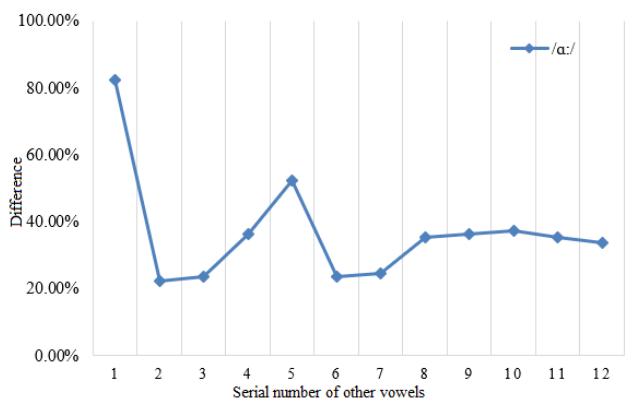
The mouth shape of each long vowel was treated as a good sample, and the mouth shape of any other long vowel or short vowel was regarded as a bad sample of that long vowel. As shown in Figure 9, the similarity between each long vowel and its own mouth shape was greater than 70%. The similarity of /ɑ:/ and /ɜ:/ was even greater than 80%. Meanwhile, the mouth shape similarity between each long vowel and any other long vowel or short vowel was maintained within 40%. Therefore, our mouth shape classification model has strong in-class similarity and between-class difference, providing a good identifier and classifier for mouth shapes in oral English teaching.

Table 3. Long vowel classification matrix with 38 hidden layer nodes

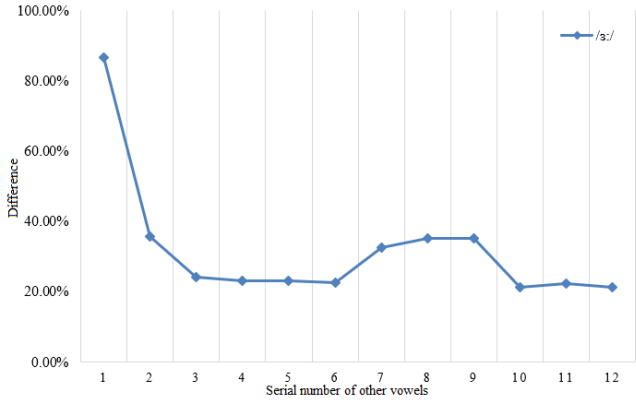
	/ɑ:/	/ɜ:/	/i:/	/ɔ:/	/u:/
/ɑ:/	55	0	0	0	0
/ɜ:/	0	49	0	0	3
/i:/	0	0	45	6	3
/ɔ:/	0	0	7	43	3
/u:/	0	2	3	3	46

Table 4. Classification results of different methods

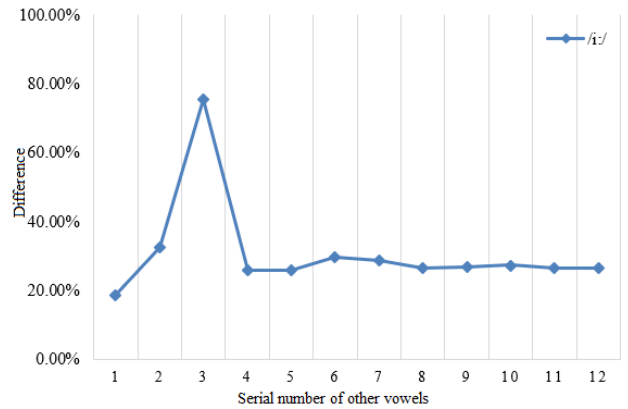
	Multi-people mouth shape images	Single-person mouth shape images
Template matching	73.5%	89.16%
Our model	91.8%	94.7%



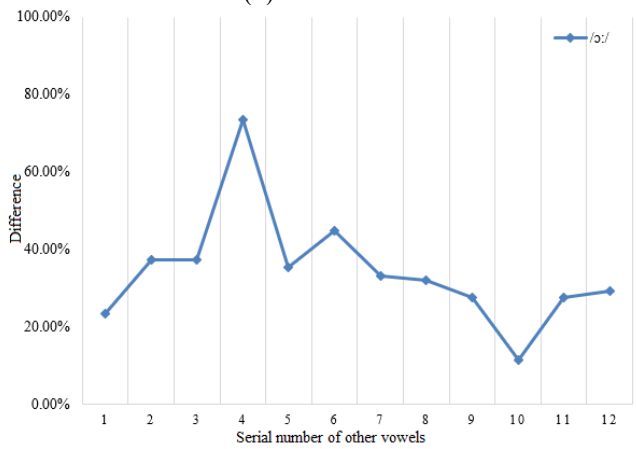
(a) /ɑ:/



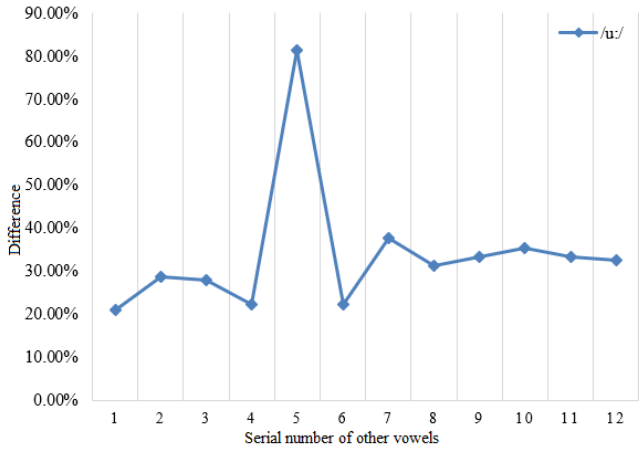
(b) /ɜ:/



(c) /i:/



(d) /ɔ:/



(e) /u:/

Figure 9. Lip shape difference between each long vowel and other vowels

6. CONCLUSIONS

This paper explores the extraction and classification of mouth shape features in oral English teaching based on image processing. Firstly, an extraction and classification model was established for mouth shape features in oral English teaching. After that, the authors explained the preprocessing of the mouth shape images in oral English teaching. Based on a competitive neural network, the lips were segmented from the oral English video frames. Next, the RBF network was adopted to extract lip boundaries and fit them into curves. Then, the proposed network was proved feasible through experiments on network performance under different number of hidden layer nodes and training error. Further, the characteristic parameters of the pronunciation of long vowels were compared to demonstrate the effectiveness of our algorithm, providing a reference for subsequent identification and classification of mouth shapes of oral English learners. Finally, the lip shape difference between each long vowel and other vowels was compared. The comparison confirms the strong in-class similarity and between-class difference of our classification model, which offers a good identifier and classifier for mouth shapes in oral English teaching.

REFERENCES

[1] Yan, J., Zhang, W., Yu, Y., Chang, J., Ding, G. (2015). Research and practice on college English oral test-a case study of Beijing institute of petrochemical technology. English Language Teaching, 8(3): 121-136.

[2] Ciekanski, M., Chanier, T. (2008). Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. ReCALL, 20(2): 162-182. https://doi.org/10.1017/S0958344008000426

[3] Dooly, M. (2011). Divergent perceptions of tellecollaborative language learning tasks: Task-as-workplan vs. task-as-process. Language Learning & Technology, 15(2): 69-91.

[4] Moreno, A.I., Vermeulen, A. (2015). Using VISP (Videos for Speaking), a mobile App based on Audio Description, to promote English Language Learning among Spanish Students: a case study. Procedia-Social and Behavioral Sciences, 178: 132-138. https://doi.org/10.1016/j.sbspro.2015.03.169

[5] Yan, J.J. (2015). Application of reciting methods on secondary school oral English learning-based on Krashen's input hypothesis. Foreign English, 20: 228-230.

[6] McCroskey, J.C. (1977). Oral communication apprehension: A summary of recent theory and research. Human Communication Research, 4(1): 78-96. https://doi.org/10.1111/j.1468-2958.1977.tb00599.x

[7] Solange Moras. (2001). Computer-Assisted Language Learning and the Internet. Karen's Linguistic Issues, 2001.

[8] Wang, X.F. (2014). The analysis of emotional factors in oral English teaching. Teaching and Management, 3: 21-24.

[9] Cai, J. (2010). A study of the reasons for and strategies of Post-CET reform. CAFL, 133(5): 3-12.

[10] Cai, J. (2002). On the evaluation of college students' English speaking ability. Foreign Language World, 87(1): 63-66.

[11] Allright, R. (1984). The importance of integration in ilassroom language learning. Applied languistics, 5: 156-171.

[12] Li, X.Y., Wang, Y.J. (2000). Interactive English Teaching Model Foreign Language and Foreign Language Teaching, 12: 22-24.

[13] Wang, J., Ma, G.M. (2005). Classroom interaction in the graduate students' oral English teaching. Chinese Higher Education Research, 6: 67-68. https://doi.org/10.3969/j.issn.1004-3667.2005.06.025

[14] He, W. (2002). On the Improvement of Graduates' English Learning. Journal of Chongqing University (Social Sciences Edition), 8(6): 189-191. https://doi.org/10.3969/j.issn.1008-5831.2002.06.068

[15] Chen, X., Liu, Y. (2005). Relocation of the teacher's role in oral English teaching in multimedia and the networking environment. Foreign Language Audiovisual Education, 6: 37-40.

[16] Sun, L.R. (2003). The application of virtual reality technology in college teaching. in journal of social sciences. Jiamusi University, 21(3): 90-91. https://doi.org/10.3969/j.issn.1007-9882.2003.03.042

[17] Wei, B. (2006). The application of virtual reality technology in college oral English teaching. Computer and Telecom, 12: 72-75. https://doi.org/10.3969/j.issn.1008-6609.2006.12.019

[18] Wang, L., Ling, L.X. (2010). College oral English teaching on the basis of virtual reality technology. Journal of Jilin Institute of Education, 26(12): 122-123. https://doi.org/CNKI:SUN:JJXK.0.2010-12-054

[19] Yang, G. Research on Three-Dimensional Textbook of College Oral English Supported by Virtual Reality. College English teaching and research, 3: 65-70. https://doi.org/CNKI:SUN:KSPA.0.2013-03-018

[20] Zhao, H. (2014). The Construction of College English Learning Environment Based on Virtual Reality Technology, 12.

[21] Wang, H.C., Ni, Y.J. (2015). Research on three-dimensional teaching materials of college English reading and writing based on virtual reality technology. Crazy English (Teacher Edition), 4: 19-23. https://doi.org/CNKI:SUN:FDJS.0.2015-04-007

[22] Zhang, J.W. (2017). Research on college oral English teaching based on virtual reality technology. Education, 2: 00070-00070.

[23] Molka-Danielsen, J. (2011). Exploring the role of virtual worlds in the evolution of a co-creation design culture. In Scandinavian Conference on Information Systems, 3-15. https://doi.org/10.1007/978-3-642-22766-0_3

[24] Zhang, W. (2014). On college oral English teaching in the base of virtual reality technology. In Applied Mechanics and Materials, 687: 2427-2430. https://doi.org/10.4028/www.scientific.net/AMM.687-691.2427

[25] Yang, J. (2019). Research on the teaching model of oral English training based on digital network. In 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 701-704. https://doi.org/10.1109/ICMTMA.2019.00160

[26] Ni, C. (2021, April). The human-computer interaction online oral English teaching mode based on Moodle platform. In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC),

633-635. https://doi.org/10.1109/IPEC51340.2021.9421327
[27] Chen, G. (2013). Computer-aided multimedia oral English teaching. In 2013 International Conference on Computational and Information Sciences, 1819-1822. https://doi.org/10.1109/ICCIS.2013.476
[28] Fan, X.Y., Wu, G. (2009). Utilization of the network and multimedia resources to conduct interactive oral English teaching of the graduate students. In 2009 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS), 3: 110-113. https://doi.org/10.1109/PEITS.2009.5406807
[29] Fang, N. (2019). An analysis of oral English teaching in college based on virtual reality technology. In 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 1-4. https://doi.org/10.1109/ICVRIS.2019.00008
[30] Yu, M., Liao, Y. (2021). The implementation strategies of smart teaching in college oral English court based on artificial intelligence. In 2021 International Conference on Internet, Education and Information Technology (IEIT), 607-612. https://doi.org/10.1109/IEIT53597.2021.00142