Hindawi
Mobile Information Systems
Volume 2022, Article ID 6224608, 8 pages
https://doi.org/10.1155/2022/6224608

Research Article

Design of the Oral English Teaching Method Based on Multimodal Feature Fusion

Xiaolei He

Department of Foreign Languages, Zhaoqing Medical College, Zhaoqing 526020, Guangdong, China

Correspondence should be addressed to Xiaolei He; hexiaolei@zqmc.edu.cn

Received 31 May 2022; Accepted 15 July 2022; Published 8 August 2022

Academic Editor: Le Sun

In order to solve the problems of too complex speech extraction algorithm and insufficient representation ability in oral English teaching, this paper proposes a speech scoring mechanism based on multimodal fusion. Firstly, feature extraction of multimodal audio and video is carried out, and a multimodal speech error detection model of LSTM-CTC is proposed; Then, the distance of MCFF, volume intensity, and pitch track are calculated by the DTW algorithm, and the speech scoring model is established. The experimental results show that under the condition of no noise and strong noise, multimodal speech detection can achieve a better error detection effect, and its system score is close to the actual situation, which can provide new ideas for oral English teaching methods.

1. Introduction

Oral pronunciation is an important embodiment of English ability. The accuracy of pronunciation determines the smooth degree of communication, and there are great differences in pronunciation habits among Chinese regions. There are 129 kinds of dialects recognized by the state. Although our country attaches great importance to English learning, we are also encouraged to learn English, but the pronunciation habits of Putonghua and dialect affect the learning of correct English pronunciation [1, 2]. Oral English teachers are prone to make mistakes in their teaching. Since the level of science and technology has made remarkable progress, it has led to the upsurge of online language teaching, but most of online teaching does not check the students' pronunciation and point out their errors. The realization of automatic error correction can detect the errors of learners' pronunciation and provide feedback for them, which is convenient for learners' oral learning and has practical significance.

The evaluation of oral English teaching is a huge multidimensional work. Most of the existing evaluation methods can only form quantitative evaluation on component values [3, 4]. With the rapid development of machine learning, speech recognition and oral teaching evaluation are gradually integrated. From the perspective of machine learning, phoneme pronunciation error detection (PED) can be regarded as a dichotomy problem, that is, to determine whether the phoneme pronunciation is correct. So, many researchers design and improve the PED system from the perspective of a classifier. Qian [5] and others studied the acoustic modeling in English PED. Compared with knowledge-based and data-driven speech rules, this method can capture better speech errors, but has a higher computational cost; Lee [6] and others used DBN instead of the Gaussian mixture model to detect the word level pronunciation errors, aligned a non-native language sample with at least one native language sample, and extracted features describing the degree of dislocation from the alignment path and distance matrix. The performance of the system is improved by replacing the input of unsupervised MFCC or the Gaussian posterior map with the DBN posterior map. In addition, research on multimodal speech recognition based on deep learning is also gradually carried out. Noda K et al. [7]combined denoising autoencoder with CNN to extract video features and audio features, respectively, which improves the reliability and robustness of the recognition system. Hu [8] proposed a new time multimodal deep

learning architecture called cyclic time multimodal RBM, which models multimodal sequences by transforming the connected MRBN sequences into probability series models, which is easier and efficient in learning fusion features.

At present, most of the automatic PED only rely on the language signal, ignoring the role of video signal in PED and error correction. Therefore, it is meaningful to integrate voice signal and video signal to realize end-to-end multimodal PED.

In view of the above research background, this paper makes an in-depth study on multimodality, pronunciation correction, and teaching quality evaluation. A multimodal English PED model based on audio and video is proposed and evaluates the oral English teaching of teachers.

2. Multimodal Feature Extraction

2.1. Audio Feature Extraction. Many researchers refer to the human auditory system and put forward the Mel frequency closely related to sound, which is nonlinear corresponding to the common sound frequency (Hz). MFCC feature extraction method uses this relation to calculate the cepstrum coefficients in the Mayer frequency domain. MFCC feature extraction process includes preweighting, framing, windowing, fast Fourier transform, Meyer filter bank, discrete cosine transform, dynamic feature, etc. [9]. Among them, fast Fourier transform and filter bank are the two most important processes, which mainly play the role of feature dimension reduction. Figure 1 shows the MFCC extraction flow.

Some audio in the audio data set is not clear due to recording or saving problems. It is necessary to manually remove the unqualified audio one by one to avoid adverse effects on the detection results. The format of audio data set is unified as the WAV file, the sampling rate is set to 50 kHz, and the channel is a dual channel. The most important step in the preprocessing stage is to highlight the characteristics of the audio signal and remove the redundant parts to facilitate the subsequent audio feature extraction. Therefore, pre-emphasis, framing, and windowing are indispensable. Compared with the traditional acoustic model, the complex forced alignment process can be abandoned in the preprocessing stage.

After pre-emphasis, the voice signal needs to be divided into frames. In this paper, the frame length is 25 ms and the frameshift is 10 ms. In fact, the sound signal is segmented by a fixed length window where there is a 15 ms overlap between two adjacent audio frames, which is the dynamic process of analog sound. There are still drastic changes between frames, so it is necessary to add Windows to the signals after frame splitting.

2.2. Video Feature Extraction. The purpose of video feature extraction is to obtain video features which are conducive to pronunciation detection and error correction. Video is a dynamic information, so it is difficult to extract video directly. The mainstream method is to convert video into continuous video frames, and the dynamic information of video is represented by continuous image frame information. The essence of video feature extraction is to extract features of the image frame. When the speaker pronounces, the lip part provides the greatest help to speech detection, so the most important part of the image frame feature extraction is to extract the lip feature.

In order to remove the influence of noise in video frames on the recognition results, it is necessary to denoise the segmented video frames to obtain the video frame set containing face information.

2.2.1. Video Framing. Video data frame processing is divided into two parts, the first part is to determine the interval of video segmentation and the second part is to determine the video segmentation interval. The process is shown in Figure 2.

The video duration is T (seconds), and the segmentation speed is v (frames/s), thus, the number of video frames N can be obtained. The number and order of video segmentation frames are very important in lip reading prediction, because the video frames in the set are arranged according to the sequence, and the number or sequence error will affect the performance of dynamic information. After segmentation, each frame has a digital number, where the number of the video frame data set must be arranged in the order from small to large, and there should be no wrong order or less order, otherwise, the corresponding relationship between audio and video information in time sequence will be affected.

2.2.2. Normalization. Face detection is carried out on these video frames to eliminate unqualified data, and then the lip features are extracted. Finally, the extracted lip key points are normalized to save the video features. It is necessary to get a lower visual representation of the vowel sound by using the pronunciation of the vowel. Because the dimensions of pixel-based video features and hybrid video features are relatively large, and the calculation process is relatively cumbersome, this paper selects the model-based video feature extraction method, where a cascaded residual regression tree is constructed to obtain the real face shape. Each leaf node of each level of the residual regression tree has a residual regression quantity. Therefore, face alignment can be achieved by stacking all the residuals [10].

In order to avoid the adverse effects caused by the inclination of the speaker's lip, the left and right key points of the lip are detilted, so that the lines connected by the two are parallel to the horizontal line to ensure that the lip of each frame is vertical. As shown in Figure 3, when the lip is tilted, the line between the two points and the horizontal line forms an included angle, and all key points are rotated in the opposite direction, which can be considered that the lip image has been straightened. The next step is to normalize the lip size. The point on the left lip is A, the point on the right lip is B, the point on the upper lip is C, and the point on the lower lip is D.

The normalized coordinate representation of point C is shown in the following equations.

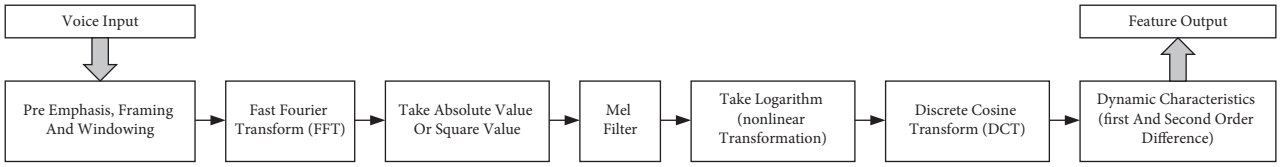

Figure 1: MFCC extraction process.

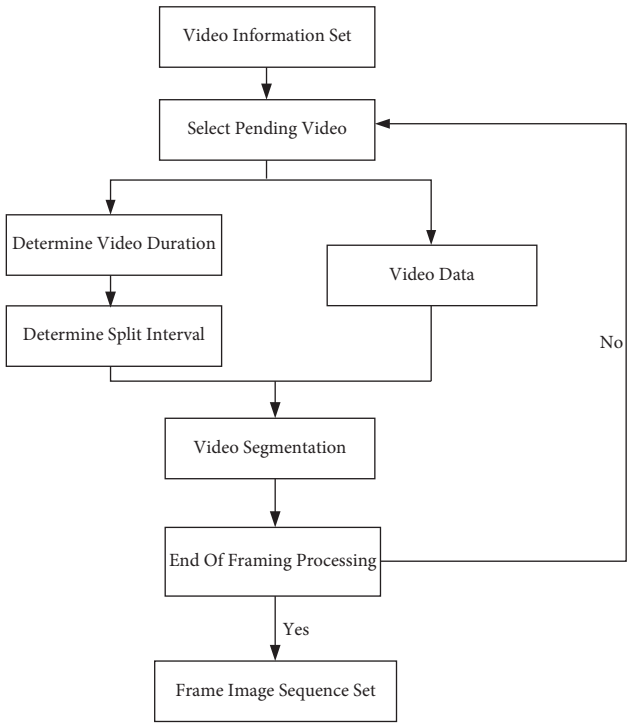

Figure 2: Video feature extraction process.

$$X_C' = \frac{X_C}{X_B - X_A}, \tag{1}$$

$$Y_c' = \frac{Y_C}{Y_C - Y_D}. \tag{2}$$

Here, (X_c', Y_c') represents the normalized coordinates of point C, and the remaining points are the information at the end of the normalization. The 20 key points of the lip were normalized in this way, and finally the visual feature vector containing the normalized information of the 20 key points was obtained.

2.3. Feature Fusion. Multimodal fusion can take advantage of the complementarity of different modes to obtain better fusion features than single mode features. The use of fusion features can improve the generalization ability of the deep learning model and has stronger robustness. The feature fusion used in this paper adopts a joint framework, as shown in Figure 4. Fusion features of audio and video are obtained by cascading feature vectors.

Multimodal joint architecture can be divided into additive join and multiplicative join. The addition connection is shown in the following equation:

$$z = f\left(w_1^T v_1 + \cdots + w_n^T v_n\right). \tag{3}$$

Here, v represents different kinds of single-mode information input, w. In this way, the semantics of different modes can be transformed into space, and the cascaded features have different modal information.

The multiplication connection is shown in the following equation:

$$z = \begin{bmatrix} v^1 \\ 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} v^n \\ 1 \end{bmatrix}. \tag{4}$$

Here, $\otimes$ represents the outer product operator.

3. Oral Scoring Mechanism Based on Multimodal Fusion

3.1. PED Model. This paper presents a multimodal speech error detection model of LSTM-CTC based on the lip angle. After the audio and video features are extracted by the previous method, they are input into the network model. Because the recurrent neural network can predict the state of the next moment according to the state of the previous time, which is necessary in multimodal speech recognition, the recognition of each phoneme is closely related to the context, it is necessary to predict the current phoneme through the phoneme of the previous moment and the next moment. However, because RNN is prone to gradient explosion and gradient disappearance, temporal information is captured by a long-term and short-term memory network, and the state of context can be captured at the same time.

3.1.1. BiLSTM Model. First, a bidirectional LSTM network model is constructed in this paper to learn audio and video features and output the posterior probability of phonemes. Among them, the activation function adopted by the LSTM unit is tanh function, which expressed as

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{5}$$

The structure of LSTM is mainly composed of forgetting gate, input gate, and output gate, where the forgetting gate is shown in the following formula:

$$f_t = \sigma\left(W_t\left[h_{t-1}, x_t\right] + b_f\right), \tag{6}$$

where σ represents the function value of Sigmoid. W_t represents the weight matrix, h_{t-1} represents the output of the LSTM neural network of the previous layer, x_t represents the input, and b_f represents the offset quantity. The value of

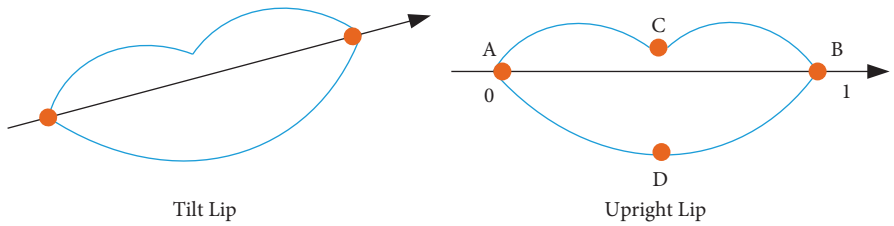

Figure 3: Normalization of lip feature.

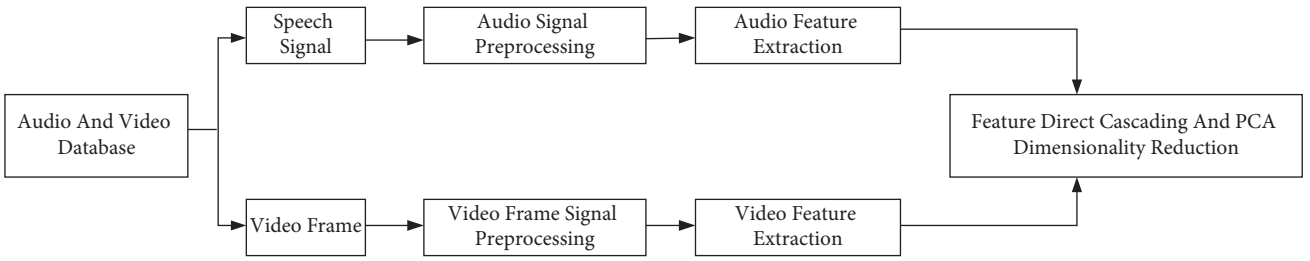

Figure 4: Multimodal feature fusion process.

the f_t element ranges from 0 to 1, indicating the degree of forgetting, with 0 indicating all forgetting, and 1 indicating all remembering.

The input gate is shown in (7) and (8), where i_t represents the amount of input state and c_t represents the selection of i_t. The forgetting gate and input gate determine the state information of the current neural network layer, namely, (9).

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right), \tag{7}$$

$$c_t = \tanh\left(W_t\left[h_{t-1}, x_t\right] + b_c\right), \tag{8}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c_t. \tag{9}$$

The output gate is shown in (10) and (11).

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_i\right), \tag{10}$$

$$h_t = o_t \cdot \tanh\left(c_t\right). \tag{11}$$

BiLSTM structure refers to that in LSTM, the current time is not only connected to the next time, but also connected to the previous time, so as to better obtain the connection between sequences through the bidirectional transmission of information.

For audio and video features, the prediction of each phoneme will be affected by the pronunciation state of the previous moment and the next moment. Bidirectional LSTM can better obtain the context relationship of audio and video fusion features, so bidirectional LSTM is selected for modeling. The number of phonemes involved in this data set is 40, so that CTC can recognize the continuous same phoneme. Therefore, the number of nodes in the Softmax output layer at the top layer of bidirectional LSTM is 41, and the posteriori probability of the output phoneme sequence is obtained.

3.1.2. CTC Training Model. Connectionist temporal classification (CTC) algorithm is an optimization method for neural network loss function. In order to make sure that each frame of speech recognition needs to be trained repeatedly, it is necessary to carry out repeated training for each frame. CTC uses end-to-end mode that it only needs to input and output sequences, then the probability of sequence prediction will be output.

The input length of CTC is greater than or equal to the output length. Assuming that the length of input sequence x is T and y_t is the output vector normalized by the Softmax layer, the probability of network output label k at time T is shown in the following formula [11]:

$$P\left(k, t|x\right) = \frac{\exp\left(y_t^k\right)}{\Sigma_{k'} \exp\left(y_t^{k'}\right)}, \tag{12}$$

where y_t^k is the kth element of y_t. So the probability $P(a|x)$ of an output path of a can be represented by

$$P\left(a|x\right) = \prod_{t=1}^{T} P\left(a, t|x\right). \tag{13}$$

Given the target sequence y, $P(y|x)$ can be expressed as (14) because a has many-to-one relation with y.

$$P\left(y|x\right) = \sum_{\alpha \in \beta^{-1}} P\left(a|x\right). \tag{14}$$

Here, β is the mapping from a to y, and β^{-1} is the inverse of β. The mapping function first merges the adjacent repeating classes and then removes the empty classes, that is, the given label sequence y. Objective function of CTC is defined as

$$CTC\left(X\right) = -\ln P\left(y|x\right). \tag{15}$$

CTC decoding is to find the sequence with the highest probability and output it under the given input sequence, as shown in the following formula:

$$a' = \text{argmaxp}(a|x). \tag{16}$$

3.2. Speech Scoring Model. MCFF, volume intensity, and pitch track of the speech are obtained to score, and the three feature parameters of the standard speech and the test speech are obtained, respectively. DTW algorithm is used to calculate the distance of three feature data in two speech. The larger the distance is, the smaller the speech similarity is, and the smaller the distance is, the higher the similarity is. According to this distance, two thresholds can be set to obtain the similarity score. After repeated experiments, the final scoring formula is as follows:

$$\text{sim} = \frac{1}{1 + a(\text{dis})^b} * 100\%. \tag{17}$$

Among them, a, b can be obtained by many test data. We use a pronunciation that is very similar to the standard pronunciation (if 90%). As a test speech, the distance between them is about 2.5 and with a voice that is very similar to the standard speech (if 20%), the distance between them is about 11, so we can get the value of a, b

dis is the distance calculated by the DTW algorithm, and sim is the similarity of two speech sounds. Since three feature parameters are used to obtain the final similarity score, after speech preprocessing, the three features of MFCC, volume intensity, and pitch track are, respectively, calculated, and then the distance of the three feature parameters dis1, dis2, and dis3 are calculated. The results of final score can be obtained by the following formula:

$$\text{sim}_{\text{all}} = w_1 \cdot \frac{1}{1 + a(\text{dis}_1)^b} + w_2 \cdot \frac{1}{1 + a(\text{dis}_2)^b} + w_3 \cdot \frac{1}{1 + a(\text{dis}_3)^b}. \tag{18}$$

Among them, $w1$, $w2$, and $w3$ represent the weights of MFCC, volume intensity, and pitch track, respectively. The weight of these three characteristic parameters is as follows: MFCC weight $w1$ is 70.15%, volume intensity weight $w2$ is 7.45%, and pitch track weight $w3$ is 22.40%.

4. Experiment and Results

4.1. Analysis of Training Model. Under the condition of no noise, the speech recognition training process of speech mode, multimode based on multimodal fusion is shown in Figures 5 and 6.

From the above results, we can see that with the increase of the number of iterations of the model, the deficiency of the training set is diminished, and the training exactness is likewise moved along. Under the condition of no noise, the model converges at 180 iterations approximately. However, in the presence of noise, the multimode fusion gradually converges in the 350th round, as the number of iterations of the model increases. However, in the single-mode voice, there is a jump, and the convergence is unstable. The convergence

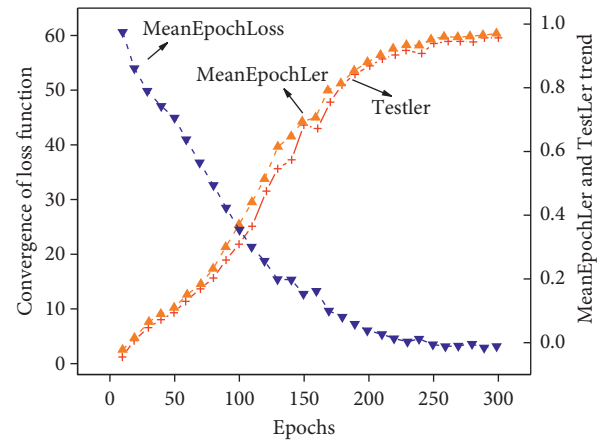

Figure 5: Model training process (no noise).

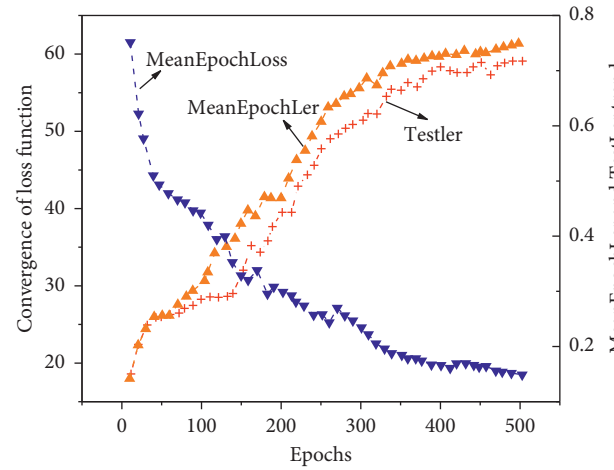

Figure 6: Model training process (with noise).

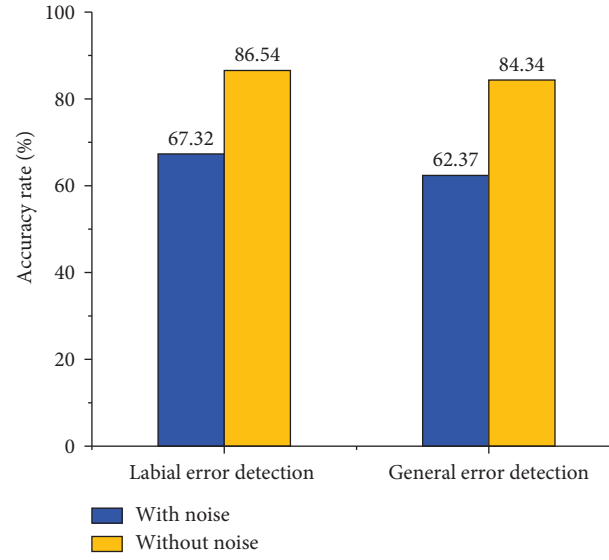

Figure 7: Accuracy rate of PED.

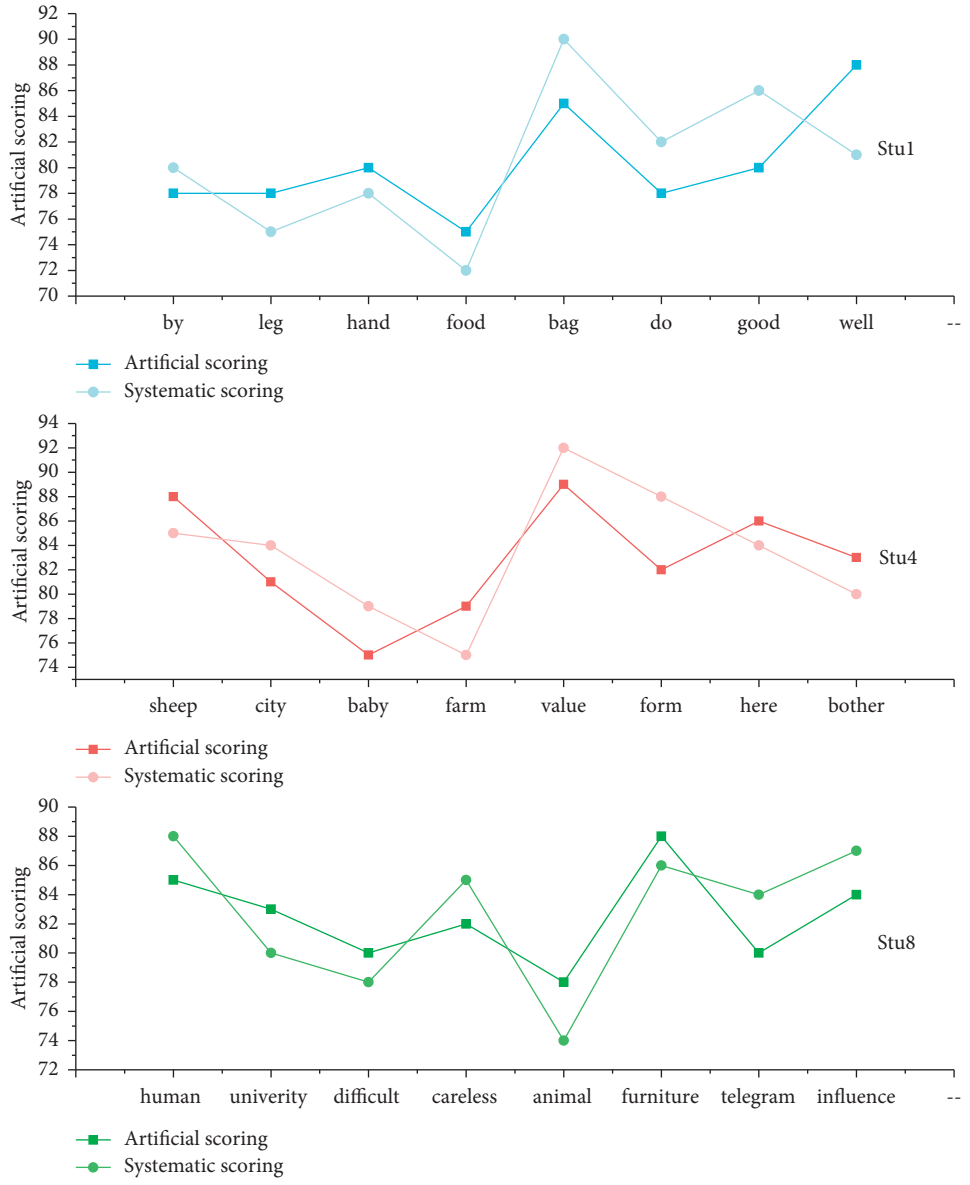

Figure 8: Speech score results.

of the two kinds of multimode fusion has little difference, which shows that the multimodal model has a certain anti-noise performance. In the noisy environment, the extracted audio features are complex and contain too much noise, which is difficult to accurately identify. Therefore, the proposed multimodal feature recognition rate will not be affected by the noise in the video feature recognition.

After the model training was experienced, the sound and video corpus of the error was brought into the above prepared model to additional action the precision of PED. The results are shown in Figure 7.

In the case of no noise, speech mode plays a more important role in recognition. Therefore, in the case of no noise, a single-mode recognition rate and an angle feature fusion with a low dimension ratio have a better recognition rate. Based on the multimodal feature fusion method, it has a higher error detection rate for the round spread lip errors, which can be seen that the introduction of angle features improves the accuracy of pronunciation detection.

4.2. Speech Scoring Effect. Partial data from 40 students (male and female) were tested on monosyllabic, two-syllabic, and polysyllabic English words. Stul-Stu3 refers to the data of three randomly selected students to test monosyllabic words, and the tested words are randomly selected from 30 monosyllabic word lists. Stu4-Stu7 refers to the data of three randomly selected students to test disyllabic words, where each student's imitative words are also randomly selected from 30 disyllabic word lists. Stu8-Stul0 refers to a random selection of 4 students to conduct a

Table 1: Sample test.

	Test of variance equation		*T* test of mean value equation						
								95% confidence interval	
	F	Sig.	*T*	Df	Sig	Mean difference	Standard error value	Lower limit	Upper limit
The variance is equal	1.232	0.269	0.262	158.000	0.793	0.337	1.286	−2.201	2.891
Variance inequality			0.262	156.016	0.793	0.33750	1.286	−2.206	2.876

polysyllabic word test, and each student imitates words from a random selection of 30 polysyllabic word lists. The artificial scoring represents the average score of 10 English majors who score the students' imitative pronunciation, and the system score represents the similarity score obtained by the system after the students' imitation. Stul, Stu2, and Stu8 are selected for analysis, and the results are shown in Figure 8.

In order to analyze whether there is any difference between system scoring and manual scoring, an independent sample t-test was conducted with scoring form (system s and artificial a) as factor. Firstly, the hypothesis is tested and the test level is defined. H0:$\mu 1 = \mu 2$, that is, there is no difference between different scoring forms on the students' score level. H1:$\mu 1 \neq \mu 2$ means that there is no difference in the scores of students. The results of the independent sample t-test are shown in Table 1.

It can be seen from Table 1 that the mean value of sample A is basically equal to that of sample S. The independent sample t-test shows that sig of F is equal to 0.269 and greater than 0.05, so the variance is homogeneous, and the corresponding bilateral test $p = 0.739$ is greater than 0.05, indicating that there is no significant difference between the scores of sample A and sample S at the level of 0.05. Therefore, the hypothesis H1 is rejected, that is, there is no difference between sample A and sample S.

Through the above data analysis of manual scoring and system scoring, it can be seen that no matter for monosyllabic, disyllabic, and polysyllabic words, the similarity results between the system scoring and the master's scoring of English major are not different, which basically conforms to people's subjective feelings. Generally speaking, after the corresponding test, the system is basically qualified, which can meet the requirements of the comparison of middle school English words and sounds, which is conducive to students' better learning of English word pronunciation after class.

5. Conclusion

The features of multimodal audio and video was extracted, and the LSTM-CTC multimodal oral PED model was proposed. The model realizes feature learning and classification through BiLSTM, and CTC is used to detect the pronunciation error. Experimental results show that in multimodal speech error detection, it can achieve a better error detection effect under the condition of no noise and strong noise, which has a positive impact on the improvement of error detection accuracy, and there was no significant difference between the score of the system and that of the master of English.

In future research, we will focus on the interactive nature of this teaching method, such as applying VR technology to multimodal oral English teaching.

Data Availability

The dataset used to support this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by "Youth Innovative Talent Program (Humanities and Social Sciences) of the Guangdong Provincial Department of Education," the project number is 2018GWQNCX161. It was supported by the "Education and Teaching Reform Program of Guangdong Medical and Health Teaching Steering Committee of Higher Vocational Colleges," the project number is 2021LX082.

References

[1] zipei Liu and W Zhao, "Factors affecting Chinese college students' English pronunciation learning and how to use mobile app to practice English pronunciation," *Overseas English*, no. 07, pp. 52-53, 2018.

[2] J. Li, *A Study of Foreign Accent of Chinese English Learners [D] Shanghai Foreign*Studies University, beijing, 2014.

[3] G. Liang and J. Zhang, "Classification evaluation of teaching quality based on SOM network," *Journal of Hainan University (NATURAL SCIENCE EDITION)*, vol. 28, no. 03, pp. 282–284, 2010.

[4] J. Zhu, "Research on the evaluation method of classroom teaching quality in Colleges and Universities Based on fuzzy neural network," *Contemporary educational practice and teaching research*, no. 10, pp. 85–87, 2015.

[5] X. Qian, H. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *Proceedings of the IEEE International Conference on Acoustics*, May 2013.

[7] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[8] D. Hu, X. Li, and X. Lu, "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3574–3582, IEEE Computer Society, Washington, DC, USA, June 2016.

[9] Z. Tan, *End to End English PED Based on Multimodality*Donghua University, Shanghai, China, 2021.

[10] Xi Nan, *Research and Implementation of Video Face Replacement Technology Based on 3D Reconstruction*Beijing Jiaotong University, Beijing Municipality, China, 2020.

[11] J. Dong and G. Liu, "Research on speech recognition method based on gru-ctc hybrid model," *Modern computer*, no. 26, pp. 13–16, 2019.