Hindawi
Journal of Sensors
Volume 2022, Article ID 2208653, 10 pages
https://doi.org/10.1155/2022/2208653

Research Article

English Pronunciation Calibration Model Based on Multimodal Acoustic Sensor

Yurui Zhou[1] **and Guolong Zhao**[2]

[1]*School of Foreign Languages, Xinyang University, Xinyang 464000, China*
[2]*College of Teacher Education, Xinyang Normal University, Xinyang 464000, China*

Correspondence should be addressed to Guolong Zhao; zgl5127@xynu.edu.cn

Received 18 January 2022; Revised 26 February 2022; Accepted 4 March 2022; Published 5 April 2022

Academic Editor: Wen Zeng

In recent years, with the increasing frequency of international exchanges, people have gradually realized that language is a tool of communication and communication, and language learning should attach importance to oral teaching. However, in traditional classrooms, one of the problems faced by oral teaching is the mismatch of the teacher-student ratio: a teacher has to deal with dozens of students, one-on-one oral teaching and pronunciation guidance is impossible, and it is also affected by the teachers and the environment constraints. Therefore, the research on how to efficiently automate pronunciation training is becoming more and more popular. Many phonemes in English have different facial visual features, especially vowels. Almost all of them can be distinguished by the roundness and tightness of the lips in appearance. In order to give full play to the role of lip features in oral pronunciation error detection, this paper proposes a multimodal feature fusion model based on lip angle features. The model interpolates the lip features constructed based on the opening and closing angles and combines audio and video in time series. Feature alignment and fusion and feature learning and classification are realized through the two-way LSTM SOFTMAX layer, and finally, end-to-end pronunciation error detection is realized through CTC. It is verified on the GRID audio and video corpus after phoneme conversion and the self-built multimodal test set. The experimental results show that the model has a higher false pronunciation recognition rate than the traditional single-modal acoustic error detection model. The increase in error detection rate is more obvious. Verification by the audio and video corpus with white noise was added, and the proposed model has better noise immunity than the traditional acoustic model.

1. Introduction

The ultimate goal of English learning is communication. The method of communication is mainly spoken language, and spoken language is realized through voice. As one of the three major elements of language, speech, is the foundation and necessity of learners and it plays a vital role in second language acquisition. Therefore, English teaching should also be based on English phonetics teaching. However, in most colleges and universities, the English phonetics course is only a "semi-independent" course. In addition, traditional English phonetics teaching is based on the monomodal teaching of students' hearing, which makes students lose their interest in phonetic learning. Secondly, restricted by the Chinese examination system, most students tend to "dumb English", because of emotional attitude, learning motivation, individual differences, and other factors, and most people speak a strong Chinese English. With the development of advanced science and technology, English phonetic teaching is no longer "speaking and ear learning" or traditional single-modal teaching, but gradually becoming a multimodal teaching combining multimedia technology and visual speech software. Teachers can use multimodality. The synergistic effect of attitude enables students to understand the characteristics of English pronunciation from hearing, vision, and touch and improve English pronunciation. Figure 1 shows the multimodel [1–10].

In traditional English learning, teachers pay more attention to writing and grammar teaching, and oral training has

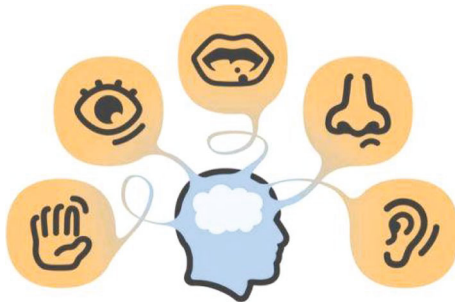

Figure 1: Multimodal model.

always been neglected. Therefore, some people ridicule that the students taught by Chinese English teaching are "dumb English," that is, most Chinese students can proficiently master English written test skills in test-oriented education, but few students are proficient in daily oral communication in English. In recent years, with the increasing frequency of international exchanges, people have gradually realized that language is a tool of communication, and language learning should pay attention to oral teaching. However, in traditional classrooms, one of the problems faced by oral teaching is the mismatch of the teacher-student ratio: a teacher has to deal with dozens of students, one-on-one oral teaching and pronunciation guidance is impossible, and it is also affected by the teachers and the environment constraints. Therefore, the research on how to efficiently automate pronunciation training is becoming more and more popular. Since the second half of the 20th century, educational technology has been one of the fastest growing fields. The use of computers as a communication medium and the emergence of the Internet have reshaped the role of computers in language learning. The computer is no longer just a tool for information processing and display, it has been given the function of communication. As a result, the Computer Assisted Language Learning (CALL) system came into being. Qian et al. divides the development of CALL into three stages, namely, active, interactive, and comprehensive [11–16].

In college oral English teaching, educators generally believe that the main task of oral teaching is to help students convey existing ideas in new languages and more refined and authentic expressions. Therefore, teachers place great emphasis on language imitation and neglect content creation when arranging oral teaching tasks, which causes language learning to break away from nonlinguistic factors, such as thought, culture, and context on which language depends, and even lack the endogenous expressive power of language learning. Although there are more and more researches on oral English in the domestic and foreign language circles, how to improve the oral level of the larger group of non-English majors and how to improve the efficiency of output training in oral English classes, so as to counteract the initiative of students in oral learning, there is too little research on independence and creativity [17–21]. Therefore, how to make full use of the limited class time to improve the status quo of college students' English pronunciation is a question worthy of consideration by college teachers.

Multimodal research emerged in the West in the 1990s. The New London Group put forward "multiple literacy," which was the first to apply multimodality to language teaching. Representatives of Western studies of multimodality teaching include Stein and Royce. In China, foreign language teaching based on multimodality has also made some progress. Multimodal theory is based on Halliday's system functional linguistic theory. It encourages teachers to include two or more modal symbols in their instructional design and appropriately uses images, sounds, text, and other interactive methods to stimulate students' learning in language. Multiple sensory experience is a teaching mode that mainly includes the training of multimodal teaching design (instructional design) and multiple reading and writing (multiliteracy), which can simulate the real context to the greatest extent and enrich the communication occasions, and it can also allow students to imitate language and create to the greatest extent and express the content so as to meet the requirements of oral teaching. Since the rise of this theory in the 1990s, although there have been a few case studies suggesting that it can effectively improve the teaching efficiency of oral English classrooms, it has been seldom used in oral English teaching, and there is still a lack of scientifically designed empirical research. In addition, with the development of information technology, more and more speech analysis software has emerged, and multimodal teaching research based on speech technology is imperative. This article is mainly based on phonetic technology, combined with linguistics, phonetics, and acoustics and explores the advantages of multimodal English phonetic teaching through the visualization of English phonetic characteristics [22–25]. In view of this, this paper proposes a multimodal end-to-end English pronunciation error detection and correction model based on audio and video. It does not require forced phoneme alignment of the pronunciation video signal to be processed and uses rich audio and video features for pronunciation error detection.

2. Multimodal Theory

Modality is a form of information transmission and communication. Regardless of spoken language mode or written language mode, it needs to rely on the language medium of sound signs or written signs or nonverbal media such as images, actions, and technical equipment. There is an interactive relationship of complement, reinforcement, synergy, and overlap between them. Multimodality refers to the inclusion of different symbolic modalities in a communication product or communication activity. It also refers to various ways of mobilizing different symbolic resources in a specific text to construct meaning. Multimodal discourse, as a communicative phenomenon, is mainly based on Halliday's system-functional linguistic theory. It is believed that other sign systems outside language, such as images and sounds, are also sources of meaning and have conceptual, interpersonal, and language functions. Article function, in the teaching design, the teacher integrates the modal symbols of two or more symbols into the teaching design and presents the teaching content of the teaching mode, which

is multimodal teaching. In the field of multimodal teaching research, the New London Group has pioneered the application of multimodality to language teaching. They believe that cultivating students' multiple literacy and multimodal meanings is the main task of language teaching. Stein clearly proposed the multimodal teaching method (multimodal pedagogies), pointing out that the multimodal teaching method highlights the indivisibility of the body and the brain to participate in communication through multimodal and multisensory collaboration. Therefore, teachers should design multimodal teaching tasks. Students should also use multiple modalities to complete tasks. The most fruitful research on multimodal discourse analysis is by Kress and van Leuwen, who proposed a design plan and application principles for the cultivation of multiple literacy skills in a multimodal environment. Royce then analyzes the complementary relationship between images and text in multimodal texts and the coordination relationship between multiple symbolic modalities in language teaching, provides an understanding of how teachers should use the visual and auditory modalities presented on the computer in the classroom, to help students develop a multimodal discourse communicative competence for research, and specifically pointed out that the reading and writing activities integrated into multimodal teaching methods can also be introduced into listening and speaking classes to cultivate students' listening and speaking skills. With the rise of multimedia teaching, Jewitt explored the relationship between teaching and modern media technology by observing the resource allocation of rhythm, multimodality, and interactivity when teachers use new technologies, it is done by multiple modalities, and points out that students should transform multiple modal signals in the learning process and practice teaching design together with teachers [26–30]. To a large extent, the accuracy of error detection is improved, especially in a noisy environment. Aiming at the shortcomings that the current lip feature extraction algorithm is too complicated, and the characterization ability is insufficient; a feature extraction scheme based on the opening and closing angle of the lips is proposed.

It can be seen that relevant researches at home and abroad agree with this multimodal collaborative and multimedia teaching model, which also laid a solid theoretical foundation for the application of multimodality in oral teaching. However, the current domestic and foreign researches generally have the following two problems: (1) most of the researches are based on case studies, and there is a lack of rigorous randomized controlled empirical research, so the credibility of the results needs to be improved; and (2) the current multimodal research are mostly concentrated in the areas of listening, reading, and writing and less involved in oral teaching.

The characteristics of English pronunciation include two aspects: segment and super segment. Segments mainly refer to vowels and consonants; super segments include intonation, stress, and rhythm. Therefore, the focus of English phonetic multimodal teaching is how to enable students to accurately grasp the characteristics of pronunciation through multimodal sensory stimulation. As far as English speech segment teaching is concerned, three-dimensional animation can be used to intuitively and vividly present the dynamic process of tongue position and lip shape in the pronunciation process, coupled with sensory stimulation such as hearing and touch and corresponding text modalities, and students can master the essentials of pronunciation quickly and comprehensively and get twice the result with half the effort. Take a program based on the English course of the University of Iowa in the United States to help Chinese English learners learn American pronunciation as an example. For example, the monophonic image shows the dynamic process of the pronunciation organs (tongue-jaw-lips-vocal cords) during the pronunciation of the vowel. In addition, the real-life three-dimensional animation can also truly present the mouth shape during pronunciation (see Figure 2(a)). A little fingertip can be placed between the upper and lower teeth when the sound is pronounced, which is convenient for students to feel the sense of touch. For English diphthongs, it is also possible to combine real-person facial profiles to enable students to master the essentials of pronunciation through video teaching. For example, diphthongs are a process of sliding from to, when pronounced, the lips are rounded to the corners of the mouth and the corners of the mouth are slightly grinning backwards, and the tongue is raised from the back of the tongue to the front of the tongue and approaching the upper palate forward, with the tip of the tongue touching the gums (see Figure 2(b)). Therefore, this research adopts a rigorous scientific research design and observes the application effect of multimodal theory in college oral English teaching through randomized controlled research, in order to provide new ideas for college oral English teaching practice.

3. Working Principle of Multimodal Pronunciation Calibration

Many students will have the problem of substandard pronunciation in the process of learning English, but it is very difficult to solve the problem of misreading only by themselves. Therefore, the research of automatic pronunciation error detection has practical significance. Most spoken pronunciation errors can be divided into four types: phoneme mispronunciation, missed pronunciation, pronouncing more pronouncing, and pronouncing time error. Phoneme is the smallest unit in the audio field. Any English word or sentence can be composed of phoneme.

When the vowel sounds are pronounced, the lips continue to remain open, and the various organs in the oral cavity are not in direct contact and will not hinder the passage of the pronunciation airflow. For vowels, the appearance can be distinguished by the roundness of the lips, the position of the tongue, and the tightness of the lips. In the frequency domain, the angle can be distinguished by the formant. The formant is the frequency band where the sound energy is concentrated. In fact, there are many correlations between the formant and the position of the tongue. There are three formants (F1, F2, and F3) for each vowel. Generally, F1 and F2 can be used to distinguish vowels. In the corresponding relationship between American vowels and formants, the horizontal line represents F2 and the

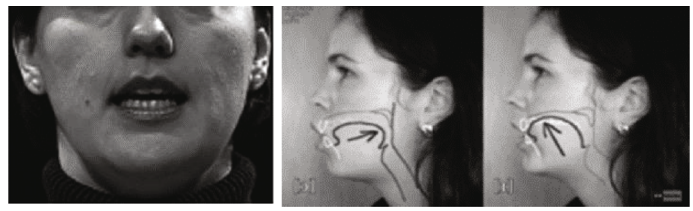

Figure 2: Mouth shape during pronunciation.

vertical line represents F1. Therefore, the pronunciation accuracy of each phoneme is the key information to measure the correctness of pronunciation.

In the field of acoustics, F1 relates to the height of the tongue, and F2 relates to the front and back of the tongue. For example, in the two phonemes of /I/ and /a:/, /I/ has a lower F1 and a higher F2, and /a:/ has a higher F1 and a lower F2. As can be seen, /I/ in front has a higher tongue position than /a:/. Therefore, the tongue position information during pronunciation can be judged based on the difference of formants, and corrective opinions are given for this information. It can help students learn English pronunciation better.

3.1. Visual Field. In the visual field, the roundness of the lips can be used to judge the pronunciation of different vowels: (1) the lips are obviously opened and rounded when pronounced, and the lips are obviously not as round when pronounced and (2) the roundness of the lips is closely related to the pronunciation of vowels, and suggestions for correcting errors can be given based on the roundness of the lips. The pronunciation of consonants is different from vowels. There is a blockage of sound airflow in the oral cavity. There is little difference in the appearance of the lips, so it is difficult to judge them with the visual information of the lips, which is mainly judged by the acoustic characteristics. The pronunciation of consonants can be judged from the pronunciation position and the way of pronunciation, according to the position of the consonant when it is pronounced. /p/, /b/, and /m/ use the lips to pronounce. /s/, /z/, /ts/, and /dz/ use the tongue and front jaw to pronounce. According to the way of pronunciation, /p/, /b/ ,/t/, /d/ first block the airflow during the pronunciation and then release it. There are also other consonant pronunciation positions and ways in the picture. In the audio field, the main difference between unvoiced and voiced sounds is whether the vocal cords vibrate or not. Vibration affects the frequency spectrum of the sound. This feature can be used to distinguish consonants. When the detection model detects the wrong pronunciation of consonants, it can give learners suggestions for correcting errors according to the way of pronunciation and the position of pronunciation. Pronunciation error detection is to detect the phoneme sequence of the pronunciation sentence to find the wrong part and error type of the phoneme pronunciation. This section will analyze the pronunciation principle of spoken language.

Audio feature extraction is an important step in improving the accuracy of detection in oral pronunciation detection. The obtained audio features are more suitable for deep learning models than the original audio. Common feature extraction methods in the field of speech detection include Linear Prediction Coefficient (LPC), FBank (Filter Bank), and Mel-Frequency Cepstral Coefficient (MFCC).

3.2. Preemphasis. The sound propagation is essentially the propagation of energy. The energy loss of high-frequency sound is more serious than that of lower-frequency sound. The domain is more stable, and the spectrum can be obtained with the same signal-to-noise ratio over the entire frequency band. The preemphasis is calculated as follows:

$$s'_m = s_m - 0.95 s_{m-1}, \tag{1}$$

where s_m represents the sampling point of the sound.

3.3. Framing. Sound framing is a fixed-duration segmentation process for sound in the time domain. In essence, a fixed number of sampling points are integrated into a unit, and the sampling value is generally 512. Another important aspect is to remove the effect between the vocal cords and the lips during vocalization. This can make the high-frequency formant more obvious. After framing, the audio signal is characterized by frame unit.

3.4. Windowing. After framing, the signal becomes smoother through Hamming window processing, reducing the size of sidelobes after fast Fourier transform processing and solving the problem of spectrum leakage. Compared with the ordinary rectangular window function, the Hamming window can obtain a higher quality spectrum. As shown in the following formula:

$$s''_n = \left\{0.54 - 0.46 \cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} s_n, \tag{2}$$

where "s_n" is the nth sampling point of preemphasis in a single frame.

3.5. Fast Fourier Transform. Compared with the time domain, it can reflect the characteristics of the sound signal in the frequency domain, so the sound signal is changed into the frequency domain. The energy distribution can be analyzed more intuitively, and the difference in energy distribution shows the difference in sound characteristics. Therefore, the energy distribution on the spectrum can be obtained through windowing and fast Fourier transform. The square of the spectrum can be calculated by the square of the modulus and the average spectrum of the output signal, as shown in the following formula:

$$S_k(i) = \sum_{n=1}^{N} s_n''(i) \cos\left(\frac{2\pi kn}{N}\right) - j\sum_{n=1}^{N} s_n''(i) \sin\left(\frac{2\pi kn}{N}\right), 1 \le k \le K,$$

$$P_k(i) = \frac{1}{N}|S_k(i)|^2. \tag{3}$$

Among them, K is the Fourier transform length, where i represents the number of frames, n represents the number of sampling points, and $s_n''(i)$ represents the value of the n sample point after windowing the i frame; $S_k(i)$ is i the kth value of the frame information spectrum. $P_k(i)$ represents the kth value of the power spectrum of the i frame.

3.6. Mel Filter Bank. After obtaining the frequency spectrum and power spectrum, there is still a lot of useless information in the frequency domain signal. Therefore, the amplitude of the frequency domain needs to be filtered through the Mel filter bank, and each single value represents a frequency band. Finally, the 26-dimensional Mel filter value is obtained:

$$M = 1125 \log\left(1 + \frac{x}{700}\right),$$

$$f_n = \frac{sfn}{K},$$

$$mfb_{nf} = 700\left(e^{fb_{nf}/1125} - 1\right),$$

$$Rf_{nf} = \sum_{k=1}^{K/2} P_k\left[\left(f_k - mfb_{nf}\right)/\left(mfb_{nf+1} - mfb_{nf}\right)\right], fb_{nf} \le Mf_k \le fb_{nf+1},$$

$$Rf_{nf} = \sum_{k=1}^{K/2} P_k\left[\left(mfb_{nf+2} - f_k\right)/\left(mfb_{nf+2} - mfb_{nf+1}\right)\right], fb_{nf+1} \le Mf_k \le fb_{nf+2}. \tag{4}$$

3.7. Logarithm. The human ears perception of sound signals is a nonlinear process, so nonlinear processing is required before cepstrum analysis can be performed. Nonlinear processing is the logarithmic operation of the value obtained by Mel filtering, as shown in the following formula. The prediction is shown in Figure 3.

$$LRf_{nf} = \log\left(Rf_{nf}\right). \tag{5}$$

3.8. Discrete Cosine Transform. In fact, each filter is partially repetitive in the filtering frequency band, so the energy value obtained also has a certain relevance. Discrete cosine transform can perform dimensionality reduction, compression, and abstract processing of data. After processing, the characteristic parameters have no imaginary part, which is more convenient in calculation. The discrete cosine transform dimension is 13, and the value of nc is between 1 and 13. The calculation is shown below.

$$D_{nc} = \sqrt{\frac{2}{NF}} \sum_{nf=0}^{NF-1} LRf_{nf} \cos\left(\frac{\pi nc}{NF}(nf + 0.5)\right) a_{nf}. \tag{6}$$

3.9. Dynamic Characteristics. Sound is a continuous signal in the time domain. The continuous signal is a dynamic process, but a single frame only reflects the characteristics of a single moment and cannot reflect the continuity of the signal. Therefore, the feature dimension is increased, and the dimension of the frame before and after it is added, which is the common first-order difference and second-order difference. The first-order difference calculation is as follows:

$$d_t = \frac{\sum_{st=1}^{ST} st(c_{t+st} - c_{t-st})}{2\sum_{st=1}^{ST} st^2}, \tag{7}$$

where the d_t indicates that the first-order difference is added to the data with the number of frames t, and $ct + s$ t is the feature of $t + st$ frame. In calculating the second-order difference, $ct + st$ indicates the first-order difference result of the corresponding frame, and d_t corresponds to the second-order difference value. The predicted value is shown in Figure 4.

Multimodal features can fuse and combine the feature information of multiple modals to provide more comprehensive information for the spoken pronunciation detection model. Multimodal fusion can be divided into feature-level fusion, decision-level fusion, and hybrid fusion based on the fusion relationship. Feature-level fusion feature fusion is also called front-end fusion. This method refers to the fusion of the input data that enters the model before the model learning, that is, the feature of each mode is fused through a certain method before entering the training model. We can understand this process as the process by which humans recognize the surrounding things. People recognize an object not only by its shape but also by combining its taste, touch, and other aspects to make judgments. These features are combined and transmitted to the brain for judgment. In practical applications, feature fusion needs to cascade the features of multiple modes after time synchronization and then uses a classifier to model this fusion feature. The current feature fusion methods mainly include feature direct connection, feature weighting, feature projection and mapping, and auditory feature enhancement. Decision fusion is also called back-end fusion, which uses the prediction results obtained after different modal information is trained separately for further fusion. This fusion method does not require the feature alignment of the two modalities in the previous period, and separate training of different modalities to avoid a huge impact on the results when a certain modal information is missing or an error occurs. Common decision fusion methods include maximum fusion, average fusion, Bayesian rule fusion, and ensemble learning. In order to balance the advantages of feature-level fusion and decision-level fusion in different aspects, some researchers have proposed a hybrid fusion model.

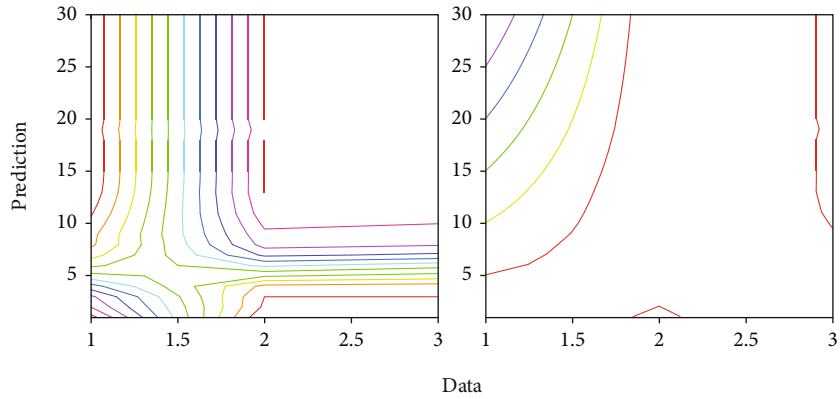

Figure 3: The prediction.

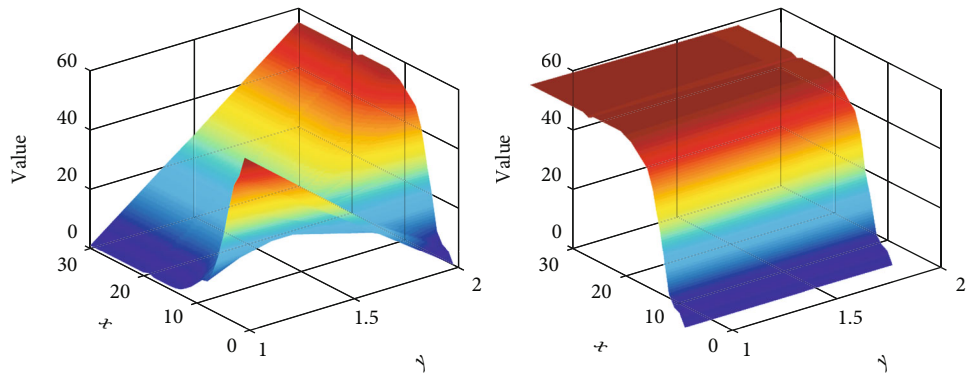

Figure 4: Predicted value.

4. System Construction

This section introduces the design process of an end-to-end multimodal pronunciation detection system, including corpus construction, audio and video data preprocessing, audio and video feature extraction, audio and video information fusion, and end-to-end pronunciation detection model construction. The most important thing in the framework is the fusion of audio and video information and the construction of end-to-end pronunciation detection models. The overall framework design of the system has five parts. The first part establishes an audio and video corpus, obtains audio and video files and annotation files suitable for multimodal detection, and records a multimodal pronunciation test set. The second part preprocesses audio information and video information separately. The third part extracts feature of audio information and video information, respectively. The fourth part establishes a pronunciation detection model based on audio and video feature level fusion, and the fifth part realizes pronunciation detection and error correction. The specific framework is shown in Figure 5. Compared with monomodal acoustic corpus, audio and video corpora are more scarce, and most audio and video corpora are not open to the outside world. GRID corpus is a sentence-level audio and video corpus that is rarely public at present, and it is widely used in the field of lip recognition.

To achieve a good recognition effect for a multimodal pronunciation detection model, a suitable audio and video data set must be selected. The quality of the audio and video data set has a decisive effect on the recognition accuracy. Common audio and video data sets include the AVLetters data set based on letter words, the BANCA data set based on number sequences, the GRID data set based on phrases, and the OuluVS data set based on everyday sentences.

The system architecture of the multimodal BiLSTM-CTC acoustic model based on audio and video fusion is mainly composed of the following parts. The first part preprocesses the audio to extract the acoustic features and preprocesses the video to get the key points of the lips. Information, normalization, and feature enhancement are performed to obtain video features. The second part interpolates the video information to ensure that the audio and video information rate is the same, and the audio and video are aligned and cascaded to obtain the audio and video fusion characteristics. The third part is the BiLSTM network, which uses the LSTM network to learn timing features, and through the Softmax classification layer, the probability of the output sequence is obtained. The fourth part is the CTC output layer, which is used to generate prediction output sequences. The error variation is shown in Figure 6. This model performs feature fusion on modalities with data synchronization and low correlation and performs decision

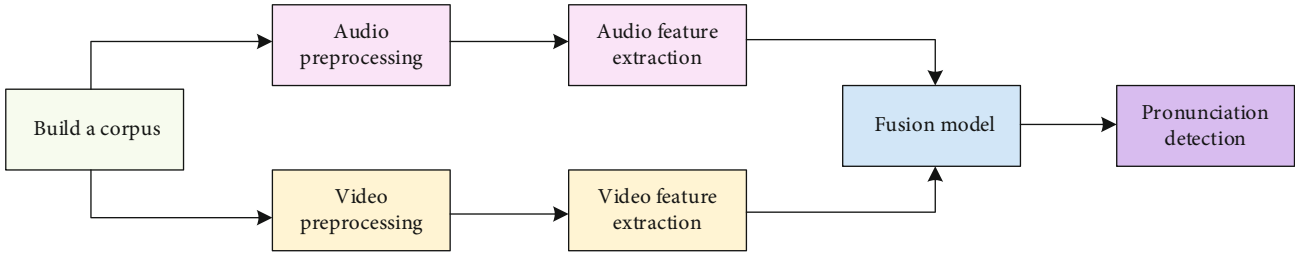

Figure 5: The specific framework.

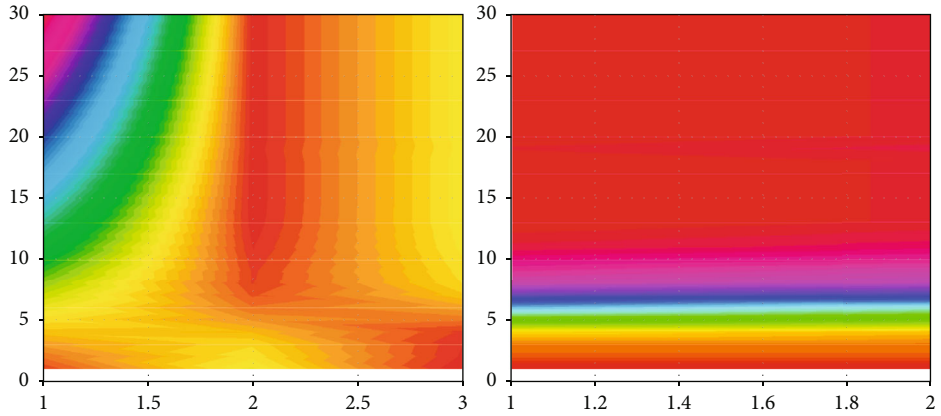

Figure 6: Error variation.

fusion on modalities with different data and updates and strong correlation.

The experimental data set in this chapter comes from the GRID phoneme annotation corpus constructed in previous content and the self-constructed multimodal pronunciation test set. The GRID corpus contains 34 speakers (18 males and 16 females), each with 1,000 spoken pronunciation videos and audio, the sampling rate of audio is 50 kHz, and the resolution of video is 720×576 dpi. The original GRID data set annotation file has annotated the words in the sentence. In Section 3, all the word annotations have been converted into phoneme annotations with reference to the cmudict dictionary to make it consistent with the usage of the model. The video data set of 1 out of 34 people does not exist, so the audio and video set of 33 people is selected, and each person selects the audio data set of 200 sentences, totaling 6600 video files. One-tenth of the audio and video files are selected as the model test set, and the remaining files are the training set. The self-built multimodal pronunciation test set is also used as the test set. In order to better reflect the robustness of the model, white noise is added to the original audio data set. The noise comes from the NoiseX-92 noise library. The audio data set after the noise is added and the audio stream of the original data set still need to be consistent. If this change will cause the audio information and the video information to be inconsistent in timing, it will affect the accuracy of multimodal detection. The signal-to-noise ratio of the added white noise is 10 dB audio signal, as shown in Figure 7(a). Compared with Figure 7(b), the audio after adding noise still maintains the synchronization relationship at the same sampling point. In actual engineering applications, there is a chance that it will have a better effect than pure feature fusion or decision fusion.

In the experiment, the feature fusion of audio and video information is carried out first, and the fusion feature is input into the long and short-term memory network. In the experiment, the structure of the bidirectional long and short-term memory network with 3 hidden layers is selected. The lip key point position information dimension is 40 dimensions, and the angle information dimension is 6 dimensions. The MFCC coefficient of the audio input is 39 dimensions. After feature fusion, the key point position fusion feature is 79 dimensions, and the angle fusion human feature is 45 dimensions. The number of iterations is set to 300, and the training batch size is set to 64. The experiment is based on the Windows 10 64-bit operating system and the Urbanu 18.4 operating system, the CPU is Intel I7, and the GPU is NVIDIA gtx1080.

Under the condition of no noise, the speech recognition training process of speech modal, multimodal based on key point position fusion, and multimodal based on angle feature fusion are shown in Figure 8. It can be seen from the above three figures that as the number of model iterations increases, the loss of the training set is continuously reduced, and the training accuracy is also continuously improved. Besides, the two subfigures are similar since variation is also similar. Among the three schemes, the angle feature fusion and the speech monomodal speech recognition converge faster and basically converge around the 150th round. The

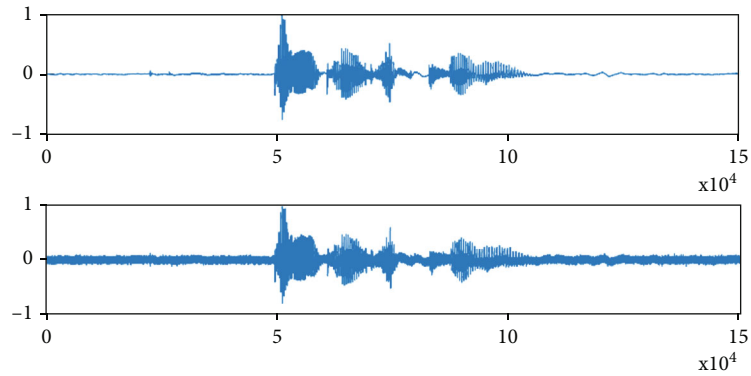

Figure 7: Audio signal with/without noise.

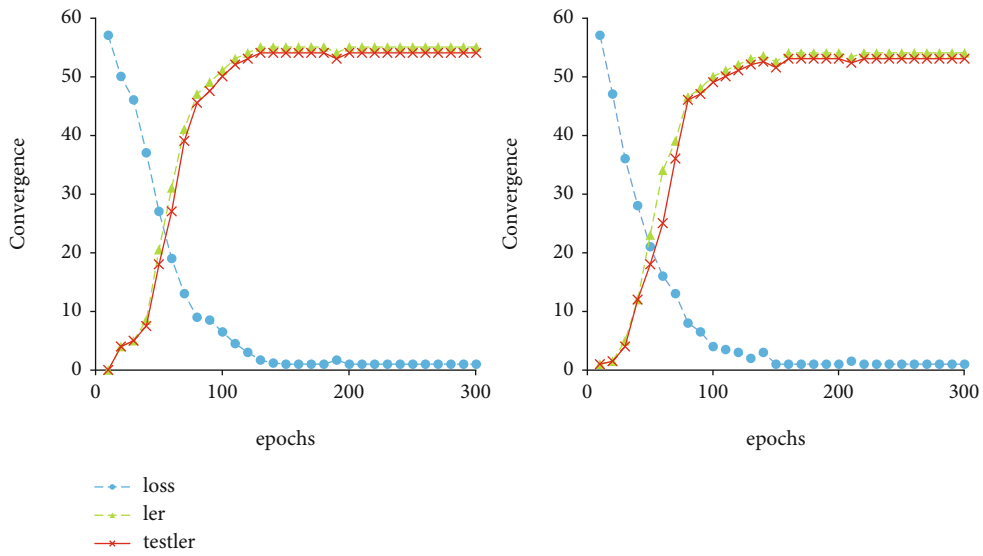

Figure 8: Training of the model.

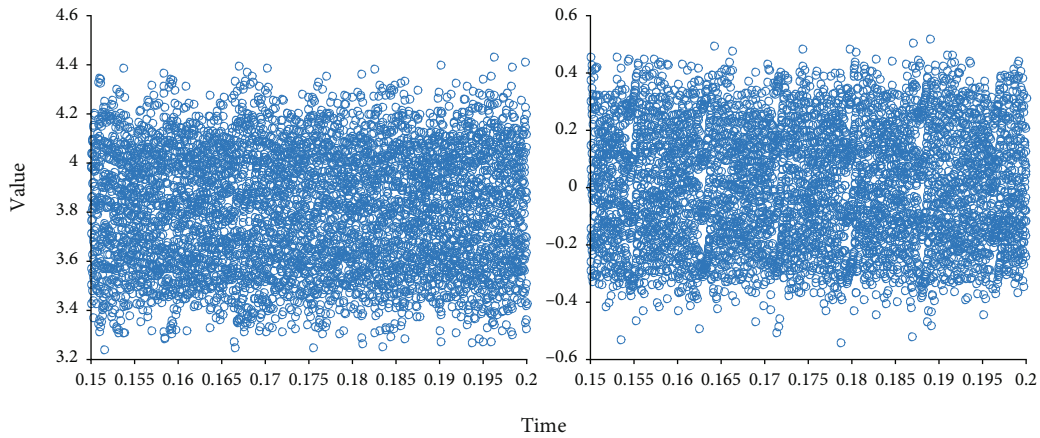

Figure 9: Predicted value vs time.

multimodal fusion based on key point features converges slowly, and it gradually converges in the 250th round. The predicted value vs time is compared in Figure 9. It can be seen from the figure that the value varies all the time. Though the waveform in these subfigures are similar, the average value is completely different.

5. Conclusion

From the perspective of multimodal discourse analysis, guided by the theory of systemic functional linguistics, combined with the characteristics of the oral English classroom, this study proposes a new mode of college English oral teaching based on the multimodal theory and conducts a randomized controlled demonstration research. The research found that compared with the traditional oral English teaching mode, the multimodal collaborative and student-centered multimodal output design in the multimodal oral English teaching can effectively improve the students' oral English level and also provide a good foundation for the college oral English classroom teaching.

Data Availability

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

Consent

The patient picture information involved in the manuscript has obtained my consent, and there is no violation of privacy and illegal use.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Xinyang University.

References

[1] L. Wang and W. Gou, "Influence of Negative Transfer of Dialects on English Pronunciation and Teaching Strategies," in *Proceedings of 4th International Workshop on Education Reform and Social Sciences (ERSS 2021)*, pp. 2–7, Chengdu, China, 2021.

[2] X. Wenqi and P. Moonyoung, "Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL Context," *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 11, no. 3, pp. 74–91, 2021.

[3] G. Zhang, P. Anand, K. S. Cheung Simon, H. C. Ching, and D. Sadia, "Quality evaluation of English pronunciation based on artificial emotion recognition and Gaussian mixture model," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 7085–7095, 2021.

[4] H. Jia, "Analysis on the path of English pronunciation teaching in colleges and universities," *Advances in Higher Education*, vol. 5, no. 1, pp. 1–9, 2021.

[5] H. Chao, Z. Feng, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *IEEE International Conference on Acoustics*, pp. 5077–5080, Las Vegas, NV, USA, 2018.

[6] I. Rehman, A. Silpachai, J. Levis, G. Zhao, and R. Gutierrez-Osuna, "The English pronunciation of Arabic speakers: a data-driven approach to segmental error identification," *Language Teaching Research*, vol. 1, no. 2, 2020.

[7] P. A. Dixon, "Book review: English pronunciation teaching and research: contemporary perspectives," *RELC Journal*, vol. 52, no. 1, 2021.

[8] G. Min, "Factors affecting Yi ethnic minority EFL learners' English pronunciation learning in Leshan Normal University, Sichuan, China," *English Language Teaching*, vol. 13, no. 6, pp. 104–118, 2020.

[9] H. C. Chen and J. X. Tian, "Developing and evaluating a flipped corpus-aided English pronunciation teaching approach for pre-service teachers in Hong Kong," *Interactive Learning Environments*, vol. 1, pp. 1–14, 2020.

[10] V. Aulia, "English pronunciation practices: from tongue twisters to YouTube Channel," *Script Journal Journal of Linguistic and English Teaching*, vol. 5, no. 1, pp. 44–54, 2020.

[11] X. Qian, H. Meng, and F. Soong, "The use of DBNHMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Thirteenth Annual Conference of the International Speech Communication Association*, shanghai ,china, 2021.

[12] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, 2017.

[13] R. I. Gusdian and R. Lestiono, "Incorporating Hijaiyah sounds in English pronunciation class: students' perception," *Journal of English Educators Society*, vol. 5, no. 1, pp. 83–88, 2020.

[14] A. P. Gilakjani and R. Rahimy, "Using computer-assisted pronunciation teaching (CAPT) in English pronunciation instruction: a study on the impact and the teacher's role," *Education and Information Technologies*, vol. 25, no. 2, pp. 1129–1159, 2020.

[15] K. N. Anastazija, M. C. Pennington, and P. Rogerson-Revell, "Martha C. Pennington and Pamela Rogerson-Revell.English pronunciation teaching and research: contemporary perspectives," *Journal of Second Language Pronunciation*, vol. 6, no. 2, pp. 265–269, 2020.

[16] K. Yeni, R. Amin, and C. Raqib, "Designing phonetic alphabets for Bahasa Indonesia (PABI) for the teaching of intelligible English pronunciation in Indonesia," *Indonesian Journal of Applied Linguistics*, vol. 9, no. 3, pp. 726–734, 2020.

[17] Y. Liu and K. W. Li, "A two-sided matching decision method for supply and demand of technological knowledge," *Journal of Knowledge Management*, vol. 21, no. 3, pp. 592–606, 2017.

[18] J. Byun and S. Jang, "Effective destination advertising: matching effect between advertising language and destination type," *Tourism Management*, vol. 50, no. 10, pp. 31–40, 2015.

[19] T. Aki and M. D. Kim, "Exploring Japanese EFL learners' attitudes toward English pronunciation and its relationship to perceived accentedness," *Language and Speech*, vol. 1, 2021.

[20] K. Igarashi, I. Wilson, and I. Wilson, "Improving Japanese English pronunciation with speech recognition and feedback system," *SHS Web of Conferences*, vol. 77, article 02003, 2020.

[21] Z. Qian, L. Yonghong, and L. Guo, "The facilitation of modern technics for English pronunciation class in foreign language learning in China," *Journal of Physics: Conference Series*, vol. 1437, no. 1, pp. 012027–012027, 2020.

[22] G. Celia, "Teaching L2 English pronunciation: research and course Design," *Studies*, vol. 41, no. 2, pp. 215–224, 2019.

[23] J. A. Brander and E. J. Egan, "The winner's curse in acquisitions of privately-held firms," *Review of Economics & Finance*, vol. 65, pp. 249–262, 2017.

[24] Z. Palmowski, "A note on var for the winner's curse," *Economics/Ekonomia.*, vol. 15, no. 3, pp. 124–134, 2017.

[25] I. Y. Pavlovskaya and L. Hao, "The influence of breathing function in speech on mastering English pronunciation by Chinese students," in *Proceedings of the 3rd International Conference on Social Sciences, Public Health and Education*, pp. 43–50, Huhhot, China, 2019.

[26] B. M. Celeste and C. C. Patricia, "The voice of novices on the teaching of English pronunciation," *Praxis Educativa*, vol. 23, no. 3, pp. 1–9, 2019.

[27] G. F. Smith and British Broadcasting Corporation, "Learning English: pronunciation," *Journal of Second Language Pronunciation*, vol. 5, no. 2, pp. 333–338, 2019.

[28] C. JYH, "The choice of English pronunciation goals: different views, experiences and concerns of students, teachers and professionals," *Asian Englishes*, vol. 21, no. 3, pp. 264–284, 2019.

[29] D. Ahn, S. Choi, D. Gale, and S. Kariv, "Estimating ambiguity aversion in a portfolio choice experiment," *Quantitative Economics*, vol. 5, no. 2, pp. 195–223, 2019.

[30] T. Hayashi and R. Wada, "Choice with imprecise information: an experimental approach," *Theory & Decision*, vol. 69, no. 3, pp. 355–373, 2010.