A Study of Correction Training for English Pronunciation Errors Through Deep Learning

Guannan Li
Department of Foreign Languages, Huihua College, Hebei Normal University, Hebei 050000, China
E-mail: ncz939@126.com

Keywords: deep learning, spoken English, pronunciation error correction, Mel-frequency cepstral coefficient, facial feature

Received: October 31, 2023

In the process of globalization, English has become an essential skill. This article provides a brief introduction to the recognition process of English pronunciation errors based on deep learning. In the recognition process, the audio features of pronunciation were combined with the video features of lip movements during pronunciation to improve error detection performance. Subsequently, simulation experiments were conducted on the error detection algorithm, and a case analysis was performed on 100 freshmen from Hui College at Hebei Normal University to verify the effectiveness of the algorithm in correcting pronunciation. The results showed that the long short-term memory (LSTM) algorithm based on audio and video converged the fastest during training and had the smallest loss function. Additionally, it achieved the highest accuracy in phoneme recognition and pronunciation error detection, while being less affected by noise interference. After using the pronunciation error detection algorithm proposed in this article for oral correction training, students' pronunciation was significantly improved.

Povzetek: Članek obravnava proces prepoznave napak v angleški izgovorjavi z globokim učenjem, kar izboljšuje izgovorjavo študentov.

1 Related works

Some studies related to English pronunciation correction is shown in Table 1. The studies listed in Table 1 have all focused-on methods for recognizing English speech. Some emphasized improving speech quality to enhance recognition accuracy, while others have already applied speech recognition algorithms to English pronunciation practice and verified their auxiliary role. This article also approaches the topic from the perspective of speech recognition and applies algorithms to English pronunciation correction. The principle behind this correction is to utilize speech recognition algorithms to convert speech into text and then compare it with the actual text of the recognized speech in order to identify errors. In terms of speech recognition, this article mainly employs the long short-term memory (LSTM) algorithm and introduces more intuitive lip feature points for improved accuracy.

Main authors	Research content	Research results
Gang [4]	Based on artificial emotion recognition and high-speed hybrid models, they analyze and filter various types of noise that affect speech quality in order to enhance students' English speech recognition abilities.	The results demonstrated that the model they constructed performed well.
Sidgi et al. [5]	They conducted a study on the effectiveness of ASR eyeespeak software in improving pronunciation for Iraqi English learners.	The research results showed that the software could significantly improve students' English pronunciation.
Dai [6]	They designed an intelligent system based on speech recognition technology to correct students' English pronunciation errors.	Comparative experiments verified the practical application value of the system.

Table 1: A summary of related works

2 Introduction

Pronunciation errors have always been one of the challenges that students face because they not only affect communication effectiveness [1] but also can lead to a

decrease in learners' interest and confidence in learning English [2]. Traditional methods for correcting spoken English mainly rely on teacher guidance and repetitive practice; however, this training method is limited by time and location, making it unable to meet the personalized needs of students for autonomous learning. Additionally, correcting oral mistakes is also a time-consuming and tedious task for teachers. The development of deep learning technology has brought new solutions to the field of speech processing [3]. Through extensive data training, deep learning technology enables voice recognition and analysis, facilitating oral correction. The article provides a brief introduction to the process of recognizing pronunciation errors in spoken English based on deep learning. In this recognition process, audio features of pronunciation were combined with video features of lip movements during pronunciation to improve error identification performance. Subsequently, simulation experiments were conducted on the error detection algorithm, and a case study was performed on 100 freshmen from Huihua College at Hebei Normal University to verify the effectiveness of the algorithm in correcting pronunciation.

3 Deep learning-based English pronunciation error recognition

During English oral training, students typically mimic the pronunciation of standard texts. However, they frequently face challenges when trying to accurately pronounce certain sounds during practice sessions. It is difficult for them to adjust themselves and enhance their pronunciation on their own due to limited guidance from teachers caused by time and location restrictions, leading to low effectiveness. The advent of deep learning technology offers a novel approach for correcting pronunciation [7]. In the field of audio, the smallest unit is a phoneme, which is represented by 'phonetic symbols' in English pronunciation. Therefore, when using deep learning techniques to correct pronunciation, it involves recognizing students' imitated pronunciation's phoneme sequence and comparing it with the standard text's phoneme sequence to achieve identification and correction of pronunciation errors.

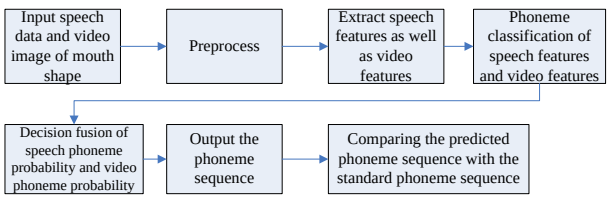

Figure 1: The recognition flow of deep learning-based English pronunciation errors.

Figure 1 illustrates the process of English pronunciation error recognition based on deep learning. In this recognition process, not only audio features are utilized but also mouth shape video features are incorporated [8] to enhance the accuracy of phoneme sequence recognition. The specific steps are as follows.

① The English pronunciation audio data of students and the corresponding synchronized video data of pronunciation are inputted.

② The audio and video data are preprocessed [9].

③ Features are extracted from the audio and video data. Mel-frequency cepstral coefficients (MFCC) are employed to extract features from the audio data. Firstly, fast Fourier transform (FFT) transformation is performed on the audio signal [10], then MFCC features are extracted from it. The corresponding formula is:

$$\begin{cases} Y(k)=\sum_{n=0}^{N-1} y(n)\cdot e^{\frac{-2j\pi kn}{N}} \\ P(\omega)=|Y(k)|^2 \\ S(m)=\ln\left(\sum_{k=0}^{N-1} P(\omega)\bullet H_m(k)\right) \\ \sum_{m=0}^{M-1} H_m(k)=1 \\ c(l)=\sum_{m=1}^{M-1} S(m)\cos\left(\frac{\pi l(2m+1)}{2M}\right) \quad l=1,2,3,\cdots,L \end{cases} , (1)$$

where $Y(k)$ stands for the frequency domain signal after FFT [11], $y(n)$ stands for the original time-domain signal, k stands for the serial number of the sampled point, n represents the time sampling point of the time-domain signal, $P(\omega)$ is the instantaneous energy of $Y(k)$, $H_m(k)$ is the frequency response of a triangular filter, m is the serial number of a group of triangular filters, totally M, $c(l)$ is the L-order MFCC feature parameter, and $S(m)$ is the energy spectral function of frequency domain signal after filter processing. The Dlib algorithm is used for video data feature extraction, which utilizes gradient-boosted regression trees to extract and recognize 68 feature points in face images. The corresponding formula is:

$$\begin{cases} S_t=(x_1^t,x_2^t,\cdots,x_{68}^t) \\ S_{t+1}=S_t+r_t(I,S_t) \end{cases} , (2)$$

where S_t represents the current set of facial feature points, x_i^t is the i-th feature point in the current facial image, $r_t()$ is the t-grade cascade regressor that is used for calculating the residual error between the current facial key points and the real face and updating the current facial key points according to the residual error, and S_{t+1} represents the set of facial feature points after $r_t()$ updating. In the recognition process, pronunciation is identified through lip movements. Therefore, only 20 feature points in the lip area are needed to avoid

interference from other feature points. Additionally, these 20 feature points are also normalized [12].

④ Both audio and video features are used for phoneme classification, and LSTM is employed to recognize these two types of features. Compared to ordinary neural network structures, LSTM not only utilizes the current input data but also takes advantage of previous state data, making it more suitable for handling sequence problems. Pronunciation phoneme recognition is also a kind of sequence problem. The calculation formula within the hidden layer of LSTM is:

$$\begin{cases} f_t = \sigma(b_f + u_f x_t + \omega_f h_{t-1}) \\ s_t = f_t s_{t-1} + g_t \sigma(b + u x_t + \omega h_{t-1}) \\ g_t = \sigma(b_g + u_g x_t + \omega_g h_{t-1}) \\ h_t = \tanh(s_t) q_t \\ q_t = \sigma(b_q + u_q x_t + \omega_q h_{t-1}) \end{cases}, (3)$$

where f_t, s_t, g_t, q_t are the output results of the forget, circulating, input, and output gates [13], h_t is the hidden state in the calculation process, $\omega, \omega_f, \omega_g, \omega_q$ are the weights of the gated recurrent, forget gate, input gate, and output gate units for hidden state h_{t-1} at the last moment, u, u_f, u_g, u_q represent the weight of the gated recurrent, forget gate, input gate, and output gate units for current input data x_t, and b, b_f, b_g, b_q represent the bias terms of the gated recurrent, forget gate, input gate, and output gate units.

⑤ After performing forward computation separately on the audio features and video features using LSTM, probability distribution sequences of phonemes are obtained for each. Then, the phoneme probability distribution sequences from both audio and video are weighted and summed together. The weight allocation between them is usually fixed based on empirical knowledge, but this paper adopts a gating mechanism to adaptively adjust the weights. Finally, the highest probability phoneme sequence is obtained from the combined phoneme probability distribution sequence [14].

⑥ The calculated predicted phoneme sequence is compared with the standard phoneme sequence corresponding to the input audio or video in order to detect any inconsistencies and provide suggestions.

4 Case study

4.1 Experimental environment

The article initially examined the algorithm designed to identify English pronunciation errors and subsequently evaluated its efficacy in correction training. The recognition algorithm was tested on a laboratory server.

4.2 Algorithm test setup

The audio and video dataset used for simulation experiments was a self-built dataset. The data were collected from 100 sophomore students who were randomly selected from Huihua College of Hebei Normal University, including 52 male students and 48 female students. The pronunciation of twenty sentences by these participants was recorded at a sampling rate of 16 kHz. During the process of collecting pronunciations, the facial changes of the participants were also simultaneously recorded at a frame rate of 60 fps. Consent has been obtained from the subjects for the collection of audio data and facial video data, and the purpose of the data has been explained to them, with an assurance that it will not be used for any other purposes.

The relevant parameter settings for the identification algorithm used in this article are shown in Table 2. The number of nodes in the input layer of the LSTM algorithm depended on the dimensionality of the input data, while the number and activation function type of hidden layer nodes were obtained through orthogonal experiments. The number of output layer nodes depended on the phoneme labels as well as the quantity of blank symbols and termination symbols in the speech dataset. In addition, to further validate the recognition algorithm proposed, comparative experiments were conducted with two other algorithms. One algorithm only used lip video for recognition while the other only used audio. The two algorithms only differed in the recognition features used, with both algorithms utilizing LSTM as the main component. The relevant parameters of the two algorithms were solely dependent on the input feature dimensions in terms of the number of nodes in the input layer, while all other parameters remained consistent with Table 2.

Table 2 Settings of relevant parameters of the proposed recognition algorithm

Name of parameter	Value	Name of parameter	Value
Number of MFCC feature dimension	39	Number of lip feature dimension	46
Number of nodes in the input layer of LSTM	85	Number of hidden layers	3
Number of nodes in the hidden layer	200	Activation function of the hidden layer	Sigmoid
Number of nodes in the output layer	50	Maximum training number	300

In addition, to test the robustness of the algorithm in this article against noise, white noise was added to the audio file, and then speech recognition was performed on the noisy audio.

4.3 Test on the correction effectiveness of the algorithms

A total of 100 freshmen from Huihua College of Hebei Normal University were randomly selected and divided into two groups: the control group and the experimental group. Both groups underwent a pronunciation test before receiving correction training, with a maximum score of 10 points. Afterwards, both the control group and the experimental group received two weeks of correction training, followed by another pronunciation test [15].

Traditional teaching methods included: ① conventional classroom teaching, where students follow the teacher's reading; ② students formed groups and engaged in English communication on specific topics.

The improved teaching method includes assigning homework for students to practice oral skills using the algorithm proposed in this article, in addition to the above two activities. The algorithm helped identify pronunciation errors and allowed students to adjust their pronunciation based on standard pronunciation.

Statistical analysis was conducted on the test scores of two groups of students before and after correction training using SPSS software, followed by an independent t-test. A P-value less than 0.05 indicated significant differences.

4.4 Test results

The convergence curves of the three algorithms for pronunciation error recognition are shown in Figure 2. From Figure 2, it can be observed that all three algorithms converged as the number of iterations increased. The video-based LSTM algorithm achieved stability after approximately 160 iterations, while the audio-based LSTM algorithm converged after about 120 iterations. The audio-video based LSTM algorithm reached stability after approximately 80 iterations. The video-based LSTM algorithm had the highest value in terms of the loss function, followed by the audio-based LSTM algorithm, and finally, the audio-video based LSTM algorithm had the lowest value when convergence was reached.

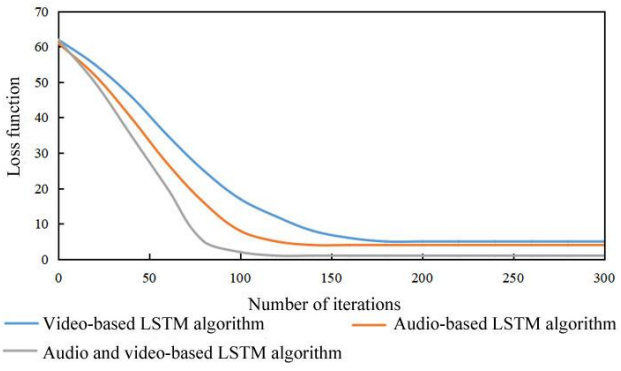

Figure 2: Convergence curves of three pronunciation error recognition algorithms.

The accuracy of three algorithms for phoneme recognition and pronunciation error detection is shown in Figure 3. It can be observed that the LSTM algorithm based on audio-video data had the highest accuracy, followed by the LSTM algorithm based solely on audio, and the lowest accuracy was seen in the LSTM algorithm relying solely on video. Furthermore, it was noted that the phoneme recognition accuracy of each algorithm surpassed that of pronunciation error detection. This is because when detecting pronunciation errors, the phoneme sequence of the pronunciation was recognized firstly, and then the recognized sequence was compared to the standard sequence, which reduced the accuracy.

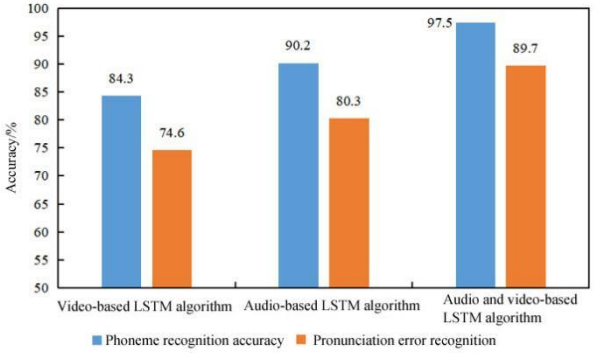

Figure 3: Phoneme recognition accuracy and pronunciation error detection accuracy of three pronunciation error recognition algorithms.

In order to test the noise resistance of the recognition algorithm, white noise was added to the audio files for recognition, and the results are shown in Figure 4. Overall, even with the addition of white noise in the audio files, the LSTM algorithm based on audio-video achieved the highest accuracy in phoneme recognition, followed by the audio-based LSTM algorithm and the video-based LSTM algorithm with lower accuracy. Compared with the same recognition algorithm before and after adding white noise, it can also be observed that the LSTM algorithm based on audio and video had a slight decrease in phoneme recognition accuracy when facing white noise interference, but the decrease was not significant. On the other hand, the other two algorithms had a larger decrease, especially the LSTM algorithm based on video.

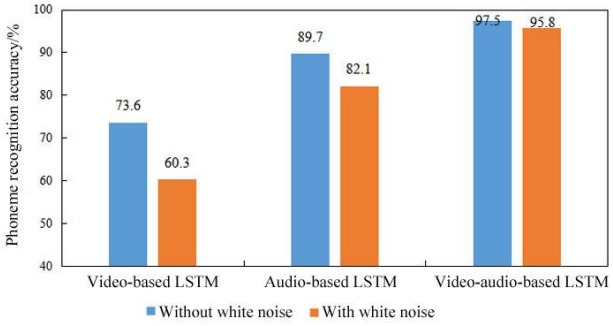

Figure 4: The phoneme recognition accuracy of three recognition algorithms under noise interference.

After English oral training, the distribution of test scores for the control group and experimental group is

shown in Figure 5. From Figure 5, it can be observed that prior to English oral training, there was minimal disparity in the distribution of test scores between the control group and experimental group, with a majority of scores concentrated within the range of five to six points. However, following oral English training, a noticeable distinction emerged in the distribution of test scores between the two group. The control group exhibited little change compared to their pre-training performance, whereas the range of score that most subjects achieved increased to eight points. The descriptive statistics of scores for the control group and experimental group before and after oral training are presented in Table 3. It can be observed that prior to the training, the P value for the average score ± standard deviation of the two groups was 0.784, i.e., the difference was not remarkable. After the oral training, the average score ± standard deviation of the two groups was significantly different, and the average score of the experimental group was significantly higher, showing a P value of 0.011. In addition, comparing the performance before and after the training within the same group, it was found that the P value of the control group was 0.698, i.e., the difference was not significant; the P value of the experimental group was 0.014, i.e., the difference was significant. Therefore, it was concluded that the use of the LSTM algorithm based on audio-video data effectively assisted students in correcting pronunciation errors during oral practice.

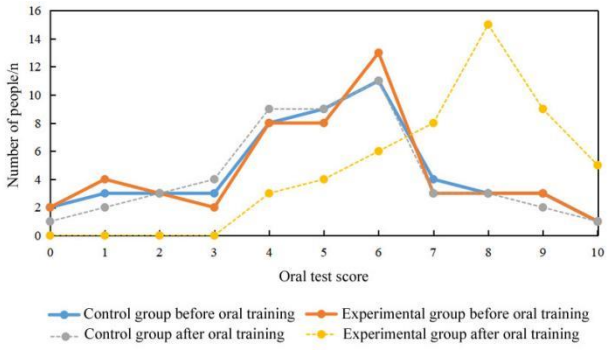

Figure 5: The distribution of oral test scores of the control and experimental groups before and after oral training.

Table 3 Descriptive statistics of the control and experimental groups before and after oral training

Group	Before oral training	After oral training	P value
Control group	5.00 ± 2.33	4.80 ± 2.09	0.698
Experimental group	4.96 ± 2.37	7.50 ± 1.63	0.014
P value	0.784	0.011	

5 Discussion

In the process of language learning, pronunciation is a crucial aspect. For many non-native English speakers, English pronunciation can be quite challenging. Deep learning technology offers new possibilities for addressing this issue. By utilizing deep learning, it becomes possible to model and analyze large-scale speech data, enabling more accurate and rapid speech recognition while also correcting any errors present. This article primarily employs LSTM for speech recognition and introduces lip movement features during pronunciation in order to enhance the algorithm's accuracy. Afterwards, the performance of the algorithm was tested and applied to oral training to examine its auxiliary role in oral training, as shown in the previous section. Among the single video-based algorithm, the single audio-based algorithm, and the audio-video-based algorithm, the audio-video-based algorithm converged fastest during training and had the smallest error when stable; similarly, this algorithm also demonstrated the best recognition performance compared to other test results. When applying the algorithm proposed to oral training, the experimental group that utilized this algorithm demonstrated significant improvement in their oral scores after training, whereas the control group that employed the traditional oral training method did not exhibit significant improvement.

Analyzing the reasons behind the above results, it can be observed that the algorithms based on single video and single audio rely solely on lip shape features and MFCC features respectively. Different pronunciations exhibit distinct lip shape characteristics; however, in practical applications, slight deviations in lip shape variations during continuous speech may occur, which consequently reduce the accuracy of these features. MFFC features are characteristics of audio that can more directly reflect the properties of the sound compared to lip shape features. Therefore, in both training and practical testing, MFFC features outperformed single video-based speech recognition algorithms. The video-audio-based algorithm combined lip shape features with MFFC features, resulting in superior performance during training and practical testing compared to the other two algorithms. When applying this algorithm to spoken language training, it could more accurately identify errors in user pronunciation and provide targeted corrections. Moreover, with the use of this speech recognition algorithm, users can practice anytime and anywhere on their own, making it more convenient compared to traditional methods of oral practice.

6 Conclusion

The article provides a brief introduction to the process of recognizing English pronunciation errors based on deep learning. In this recognition process, audio features of pronunciation were combined with video features of lip movements during pronunciation to improve error identification performance. Subsequently, simulation experiments were conducted on the error detection algorithm, and a case study was performed on 100

freshmen from Huihua College at Hebei Normal University to verify the effectiveness of the algorithm in correcting oral English. (1) Compared to the LSTM algorithm based solely on video or audio, the LSTM algorithm based on audio and video converged faster and had the smallest loss function when convergence was stable. (2) Whether it was the accuracy of phoneme recognition or pronunciation error detection, the LSTM algorithm based on audio-video achieved the highest accuracy, followed by the algorithm based solely on audio, while the algorithm based solely on video had the lowest accuracy. (3) Although the accuracy of phoneme recognition using the audio-video-based LSTM algorithm was slightly reduced when facing white noise interference, the reduction was not significant; however, the other two algorithms showed a great decrease in accuracy, especially the video-based LSTM algorithm. (4) Before the oral training, there was no significant difference in the distribution of scores between the control group and experimental group, as well as in their average score, highest/lowest score, and standard deviation. However, following the oral training, the control group did not show any noticeable change, while the experimental group exhibited a shift towards higher score ranges in terms of distribution. Additionally, there were improvements in their average and lowest scores in the experimental group. Additionally, a decrease in standard deviation was noted.

The contribution of this article lies in the introduction of lip shape features on top of using MFCC characteristics for speech recognition, enhancing the accuracy of the algorithm and providing an effective auxiliary tool for English pronunciation correction.

References

[1] Igarashi K, Wilson I (2020). Improving Japanese English pronunciation with speech recognition and feed-back system. *SHS Web of Conferences*, 77, pp. 1-5. https://doi.org/10.1051/shsconf/20207702003

[2] Li D S (2020). English Speech Recognition and Multidimensional Pronunciation Evaluation. *Education Research Frontier*, 010, pp. 184-188.

[3] Kleynhans N, Hartman W, van Niekerk D, van Heerden CJ, Schwartz R, Tsakalidis S, Davel M (2016). Code-switched English pronunciation modeling for Swahili spoken term detection. *Procedia Computer Science*, 81, pp. 128-135. https://doi.org/10.1016/j.procs.2016.04.040

[4] Gang Z (2021). Quality evaluation of English pronunciation based on artificial emotion recognition and gaussian mixture model. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 40, pp. 7085-7095. https://doi.org/10.3233/JIFS-189538

[5] Sidgi L F S, Shaari A J (2017). The Effect of Automatic Speech Recognition EyeSpeak Software on Iraqi Students' English Pronunciation: A Pilot Study. *Advances in Language & Literary Studies*, 8, pp. 48-54. https://doi.org/info:doi/10.7575/aiac.alls.v.8n.2p.48

[6] Dai M (2021). Intelligent Correction System of Students' English Pronunciation Errors Based on Speech Recognition Technology. *WSEAS Transactions on Advances in Engineering Education*, pp. 192-198. https://doi.org/10.1142/S0219649222400135

[7] Lim D Y, Kim S G, Chong K T (2018). Development of a Real-time Lip Recognition for Improving English Pronunciation using Deep Learning. *Journal of Institute of Control Robotics and Systems*, 24, pp. 327-333. https://doi.org/10.5302/J.ICROS.2018.18.8003

[8] Liu X, Xu M, Li M, Han M, Chen Z, Mo Y, Chen X, Liu M (2019). Improving English pronunciation via automatic speech recognition technology. *International Journal of Innovation and Learning*, 25, pp. 126-140. https://doi.org/10.1504/IJIL.2019.097674

[9] Giantari K, Sabarudin S, Zahrida Z (2020). Pronunciation Recognition of –ed Ending Words by the Students of English Education Study Program of the University of Bengkulu. *Journal of English Education and Teaching*, 4, pp. 278-293. https://doi.org/10.33369/jeet.4.2.278-293

[10] Zhao L, Liu Y, Chen L, Zhang J, Jonathan KG (2019). English oral evaluation algorithm based on fuzzy measure and speech recognition. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 37, pp. 241-248. https://doi.org/10.3233/jifs-179081

[11] Wu H, Sangaiah A K (2021). Oral English Speech Recognition Based on Enhanced Temporal Convolutional Network. *Intelligent Automation and Soft Computing*, 28, pp. 121-132. https://doi.org/10.32604/iasc.2021.016457

[12] Evers K, Chen S (2021). Effects of Automatic Speech Recognition Software on Pronunciation for Adults With Different Learning Styles. *Journal of Educational Computing Research*, 59, pp. 669-685. https://doi.org/10.1177/0735633120972011

[13] Cao Q, Hao H (2021). Optimization of Intelligent English Pronunciation Training System Based on Android Platform. *Complexity*, 2021, pp. 1-11. https://doi.org/10.1155/2021/5537101

[14] Zhan W, Chen Y (2020). Application of machine learning and image target recognition in English learning task. *Journal of Intelligent and Fuzzy Systems*, 39, pp. 5499-5510. https://doi.org/10.3233/JIFS-189032

[15] Peng S (2018). Research on Interactive English Speech Recognition Algorithm in Multimedia Cooperative Teaching. *International English Education Research*, pp. 79-82. https://doi.org/10.1109/ICITBS.2018.00095