A Spoken English Teaching System Based on Speech Recognition and Machine Learning

https://doi.org/10.3991/ijet.v16i14.24049

Fengming Jiao (✉), Jiao Song, Xin Zhao,
Ping Zhao, Ru wang
Jitang College of North China University of Science and Technology,
Tangshan, China
tstsdah@163.com

Abstract—The learning model and environment are two major constraints on spoken English learning by Chinese learners. The maturity of computer-aided language learning brings a new opportunity to spoken English learners. Based on speech recognition and machine learning, this paper designs a spoken English teaching system, and determines the overall architecture and functional modules of the system according to the system's functional demand. Specifically, MATLAB was adopted to realize speech recognition, and generate a speech recognition module. Combined with machine learning algorithm, a deep belief network (DBN)-support vector machine (SVM) model was proposed to classify and detect the errors in pronunciation; the module also scores the quality and corrects the errors in pronunciation. This model was extended to a speech evaluation module was created. Next, several experiments were carried out to test multiple attributes of the system, including the accuracy of pronunciation classification and error detection, recognition rates of different environments and vocabularies, and the real-timeliness of recognition. The results show that our system achieved good performance, realized the preset design goals, and satisfied the user demand. This research provides an important theoretical and practical reference to transforming English teaching method, and improving the spoken English of learners.

Keywords—Speech recognition, machine learning, spoken English, teaching system

1 Introduction

With the gradual acceleration of the economic globalization, English, as an international language, has become an important tool in people's communication. However, constrained by the learning model and the environment, most English learners in China find it very difficult to learn spoken English. When speaking English, students may not put their tongues in the right position due to the influence of their mother tongue, leading to inaccurate pronunciation [1]. What is more, teachers often focus on teaching grammar and vocabulary in English classes and tend to overlook students' pronunciation. Even though greater importance has been attached to spoken English teaching, due to the limited class time, teachers find it difficult to give one-to-one tutoring to students, and learners do not know what pronunciation problems they have and how to correct them. Therefore, it is quite challenging for students to improve their spoken English.

Speech recognition is a technology that processes and converts human's spoken language into text or commands that can be recognized by machines [2]. China began the research on speech recognition as early as the 1950s. In the 1980s, the Hidden Markov Model began to be widely used in the field of speech recognition. The emergence of artificial neural networks also brought new hope to machine learning. With the unremitting efforts of Chinese scientists, speech recognition has gradually matured and basically reached the international level [3]. At present, many spoken English learning systems and Apps have been developed based on speech recognition at home and abroad, but they can only be used to practice spoken English and cannot detect or correct pronunciation errors [4]. Automatic pronunciation error detection originated from the computer-assisted language teaching system, which provides the feature to automatically detect, feedback and correct pronunciation errors and greatly improves the efficiency of learners in learning spoken English [5]. The statistical speech recognition method, and the linguistic knowledge and distinguishing feature method are the two main methods currently used for pronunciation error detection at home and abroad, but the latter is not suitable for all pronunciation error detection cases, so the statistical speech recognition method is more widely applied [6]. Most learners like to be directly guided on where they are wrong and how to correct the problems when learning English.

To improve the existing problems of the spoken English systems, this paper designs a spoken English teaching system based on speech recognition and the machine learning algorithm, which uses speech recognition to help students practice spoken English, and also applies the machine algorithm to classify and detect the 6 common types of errors in English, and provides corrections and feedbacks to improve the efficiency of learners in learning spoken English.

2 Requirement Analysis and Overall Design of the Spoken English Teaching System

2.1 Analysis of the functional requirements for the system

In traditional English classroom teaching, the teacher is the main role in class. Due to many reasons such as the inaccurate pronunciation of the teacher and the limited time in the class, students have few opportunities to speak English in class, and many are even reluctant to speak because they are not confident about their pronunciation or afraid of making mistakes. Over time, they have become antagonistic against or afraid of learning spoken English [7]. Therefore, the spoken English teaching system needs to be able to teach students pronunciation, recognize students' spoken English, evaluate their pronunciation, and at the same time, correct their pronunciation and give feedbacks, and provide man-machine interaction, so that students will have the opportunity to practice spoken English. In addition, when a student inputs a spoken Chinese sentence, the system can also convert Chinese into English and give voice broadcast. Figure 1 shows the data flow diagram of the spoken English learning system designed according to English learning needs.

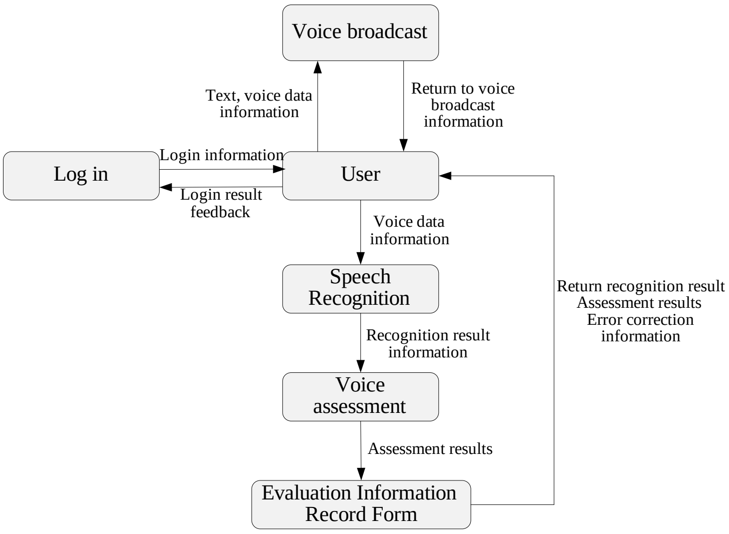

Voice broadcast
Text, voice data information
Return to voice broadcast information
Log in
Login information
User
Login result feedback
Voice data information
Speech Recognition
Return recognition result
Assessment results
Error correction information
Recognition result information
Voice assessment
Assessment results
Evaluation Information Record Form

Fig. 1. Data flow diagram of the spoken English learning system

2.2 Overall design of the system

Overall architecture of the system: According to the analysis of the functional requirements and the data flow diagram, the overall architecture of the spoken English teaching system is established, as shown in Figure 2. At the same time, in order to allow students to practice spoken English anytime and anywhere, the system should support computer and mobile login.

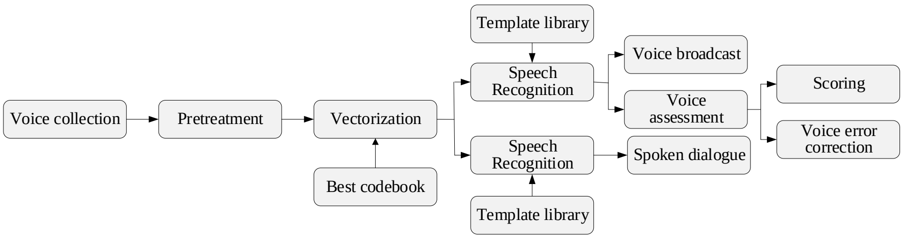


Fig. 2. Overall structure of the spoken English teaching system

Functional modules of the system: Figure 3 shows the functional modules of the spoken English teaching system.

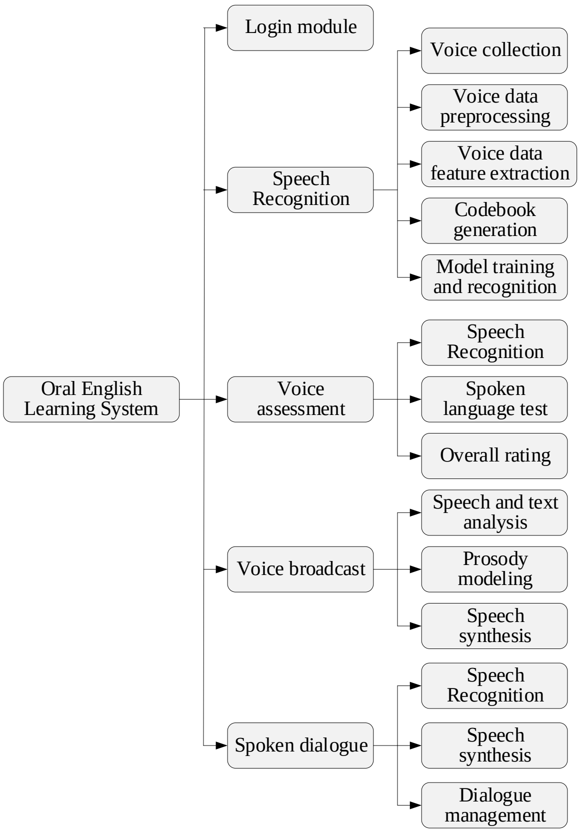


Fig. 3. Functional modules of the spoken English teaching system

Login module: The login module is used to provide student users with access to the system. New users need to register before they can log in. Existing users need to enter the correct user's name and password to enter the system.

Speech recognition module: The speech recognition module is one of the core modules of the system. Its main task is to recognize the speech data and information input by the user into the system, and convert it into information that can be recognized by the system. It also uses the interface to complete the interaction with the user. Therefore, the speech recognition module specifically consists of the following sub-modules:

Speech acquisition: It uses the microphone to collect the user's speech information, and mark each frame of the data collected to facilitate the subsequent speech feature extraction

Speech data pre-processing: It mainly performs operations such as noise reduction, and speech signal emphasis, framing and windowing on the speech information acquired by the speech acquisition module to ensure the accuracy and recognizability of the speech data [8].

Speech data feature extraction: It is used to remove redundant information that is irrelevant to speech processing, and extract key feature parameters that only describe semantic information.

Model training: It is used to train and modify the speech models in the corpus so that they can be used in application scenarios such as speech error detection, voice broadcast, and spoken dialogue provided by the system.

Model recognition: After a series of processing, the speech data are converted to the results that can be recognized by the system and allow the system to detect errors and evaluate the judgment rules, thereby achieving the speech recognition function.

Speech evaluation module: The speech evaluation module is also one of the core modules of the system. It is used to detect and score the speech recordings or spoken dialogues of the users through the spoken English error detection function based on machine learning, and at the same time feedback the evaluation results and the error information to the users to correct their English pronunciation [9].

Voice broadcast module: This module is used to recognize and convert the text or speech input by the users into the system and then broadcast it in English a natural, accurate and fluent way.

Spoken dialogue module: The function of this module is to create an English dialogue scene for students. Students can have English dialogues with the system through human-machine interaction. The system will also evaluate and correct students' pronunciation. This module can provide the language environment that students do not have in real life.

3 Design and Implementation of the Spoken English Teaching System Based on Speech Recognition and Machine Learning

With the requirement analysis and the overall design, a spoken English teaching system based on speech recognition and machine learning was established. It can be used on both computer and mobile phone so that students can learn and practice spoken English anytime and anywhere. Due to limited space, this paper only elaborates on the detailed design and implementation of two core modules of the system, i.e. speech recognition and speech evaluation. The former mainly uses MATLAB to implement speech recognition [10], and the latter mainly adopts the pronunciation classification and error detection model based on DBN-SVM to perform error detection, scoring and correction of the speech data.

3.1 Design and implementation of speech recognition

Design and implementation of the speech acquisition module: The speech acquisition module allows a user to input speech information into the system through a microphone. When the user finishes recording, the system will save the speech information input by the user as a speech file. Figure 4 shows the flow chart of speech acquisition [11].

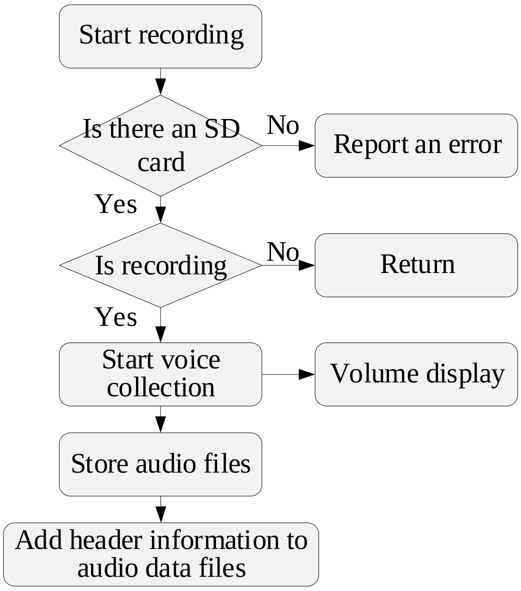


Fig. 4. Flow chart of speech acquisition

Design and implementation of the pre-processing module: This module mainly converts the acquired information into digital signals that can be recognized by the system, and then pre-processes the speech information by framing, windowing, and endpoint detection. The specific process is shown in Figure 5 [12].

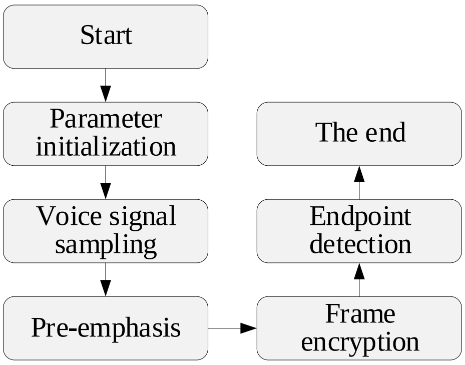


Fig. 5. Pre-processing flow chart

Design and implementation of the feature extraction module: After the speech data are pre-processed, MFCC (Mel Frequency Cepstral Coefficient) is extracted as the feature parameter of the speech. Figure 6 shows the specific flow chart of MFCC feature extraction [13].

Speech acquisition and recognition interface: Figure 7 shows the speech acquisition and recognition interface of the spoken English teaching system (for mobile phone) designed in this paper. After the learner logs in the system, he can choose the sentence that he would like to read. By clicking "Start recording", he will input a speech piece, which is the sentence he reads after the system. When the recording is over, he may click the "End recording" button, and the system will automatically process the input speech data. If the user clicks the "Play" button, the system will play the speech that has just been input, and at the same time display the input volume. If the user is satisfied with the input speech data, he may click the "Score" button, and the system will score the speech. The user can also check the "Pronunciation suggestions". The system will give the user the correct pronunciation so that the user will know what is wrong with his own pronunciation and how to improve it.

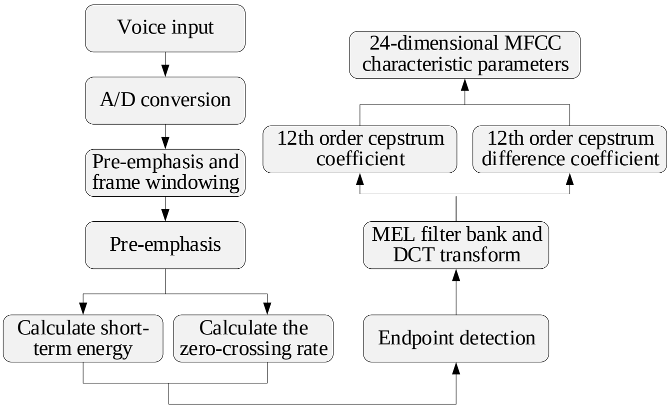
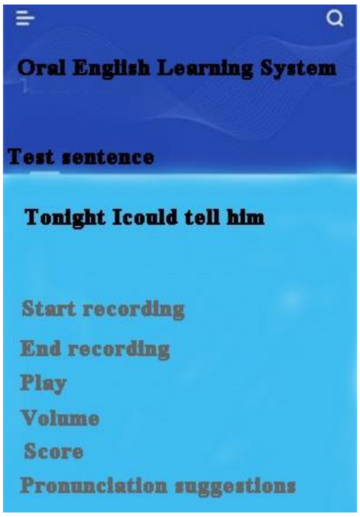

Voice input
A/D conversion
Pre-emphasis and frame windowing
Pre-emphasis
Calculate short-term energy
Calculate the zero-crossing rate
Endpoint detection
MEL filter bank and DCT transform
12th order cepstrum coefficient
12th order cepstrum difference coefficient
24-dimensional MFCC characteristic parameters

Fig. 6. Specific flow chart of MFCC feature extraction


Oral English Learning System

Test sentence

Tonight Icould tell him

Fig. 7. Speech acquisition and recognition interface

3.2 Design and implementation of spoken English error detection based on machine learning

Construction of the pronunciation classification and error detection model based on DBN-SVM: The accuracy of pronunciation is one of the main indicators of whether a person's spoken English is authentic or not. As is known, pronunciation is the result of the interaction of multiple organs, and the oral cavity plays the main role in adjusting the pronunciation. Research results indicate that the tongue movement has a great impact on the quality of pronunciation [14]. During the learning of spoken English, students may not put their tongues in the right position due to the influence of their mother tongue, leading to inaccurate pronunciation. In classroom teaching, although teachers will give instructions on the articulation and the shape of the mouth, students can only imitate the pronunciation according to their own experience, without knowing whether they have made any mistakes or not, as there is no one-to-one tutoring. This is not an effective way to improve students' spoken English [15]. Based on this, a pronunciation classification and error detection model was constructed based on the Deep Belief Network and Support Vector Machine (DBN-SVM) [16] to identify learners' pronunciation problems and propose corrective suggestions for them. Figure 8 shows a flowchart of how the DBN-SVM-based pronunciation classification and error detection model is constructed.

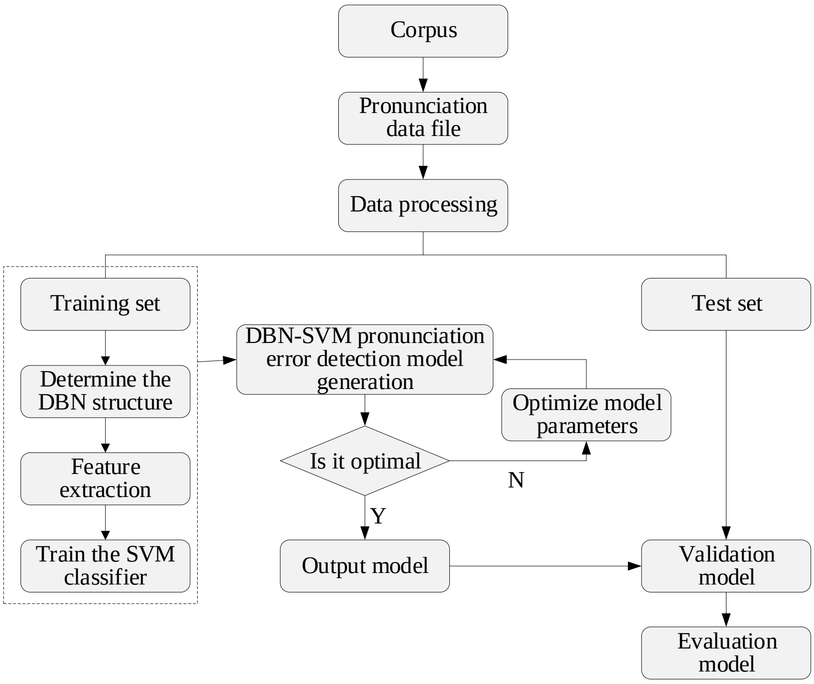


Fig. 8. Flowchart of how the DBN-SVM-based pronunciation classification and error detection model is constructed

The DBN-SVM-based pronunciation classification and error detection model first uses the Hidden Markov Model Toolbox to force the alignment of the pronunciation data files collected from the corpus with the reference text to obtain the alignment time information at the sound velocity level, and uses it as a dataset for the pronunciation classification and error detection model [17]. The second step is to pre-process the data through min-max scaling to limit the fluctuations of the data within a certain range, thereby reducing the differences among the data, and at the same time, divide the data into two parts - the training set and the test set. In order to obtain the deeply hidden features of the pronunciation data, the third step is to first determine an optimal DBN structure with the training samples, and perform deep learning of the pronunciation data. The fourth step is to use the contrastive divergence algorithm to perform parameter training on the RBMs, the basic components of the DBN. The fifth step is to output the features extracted from the deep network, and repeat the fourth step until all RBMs are trained. The sixth step is to establish a classification support machine vector model for 6 types of errors based on the output error features as the basis. The seventh step is to input the test set data into the established DBN-SVM-based pronunciation classification and error detection model, and calculate the classification and detection accuracy and error for each error type. The eighth step is to evaluate the error detection ability and performance of the constructed model.

Data acquisition and parameter setting: In this paper, CSTR VCTK Corpus was selected as the corpus for the system and the speech data manually tagged by speech experts as the data source. According to the speech experts, the speech data can be divided into 7 categories, one of which is the correct pronunciation, and the other six are the wrong ones [18]. Figure 9 shows the distribution of the selected 6 types of pronunciation error samples in the training set and test set, with 1290 samples in the training set and 258 in the test set (a total of 1290 pronunciation samples).

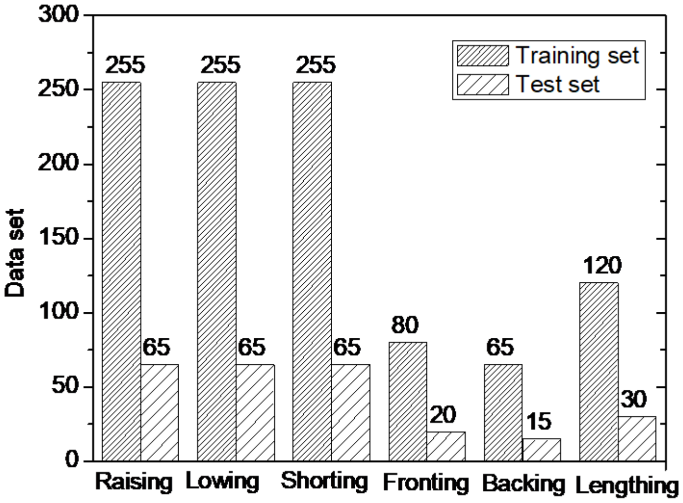


Fig. 9. Sample distribution

The parameter settings of DBN and SVM have significant impacts on the accuracy and precision of the DBN-SVM-based pronunciation classification and error detection model. Therefore, this paper determined the values of the main parameters of DBN and SVM by reference to relevant literatures at home and abroad in conjunction with the experimental method [19], which are shown in Table 1.

Table 1. Main parameter settings of DBN and SVM

Parameter	**DBN**					**SVM**		
	Hidden layer mature	***Number of nodes in each layer***	***Learning rate***	***Epochs***	***Activation function***	***Kernel function***	***Penalty coefficient***	***Gamma***
Value	2	100-130	10^{-5}	250	Sigmoid	rbf	0.5	0.01

In order to verify the effectiveness of the DBN-SVM-based pronunciation classification and error detection model in the identification of pronunciation errors, the linear discriminant analysis (LDA) and the principal component analysis (PCA) combined with the support vector machine classifier [20] were used for comparison in terms of error detection accuracy and training time with respect to three error types - raising, lowing, and shorting. Figure 10 shows the comparison results of the three algorithms in terms of training time, and Figure 11 shows the comparison results in terms of error detection accuracy. It can be seen that the DBN-SVM algorithm has the highest accuracy – above 80%, better than those of the other two algorithms, but the training time is also the longest. This is mainly because the deep neural network mechanism used in the algorithm has more iterations in the computation, which take a long time. Considering accuracy, the DBN-SVM-based pronunciation classification and error detection model was applied in the speech evaluation module of the spoken English teaching system.

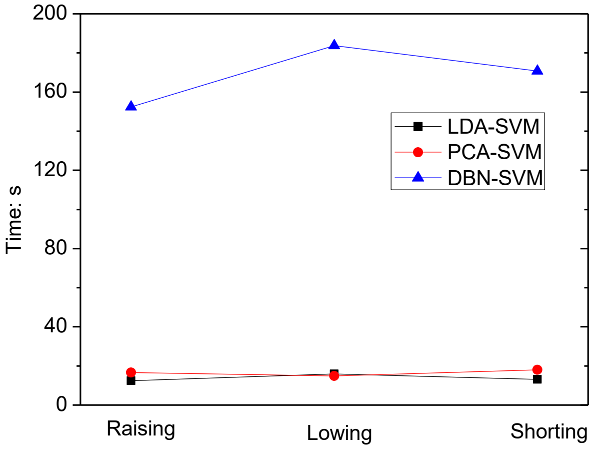


Fig. 10. Comparison results of training time of three algorithms

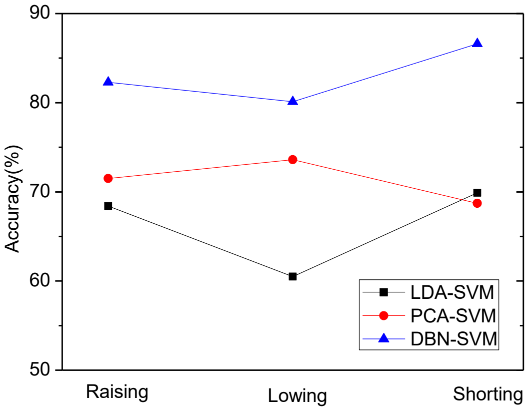
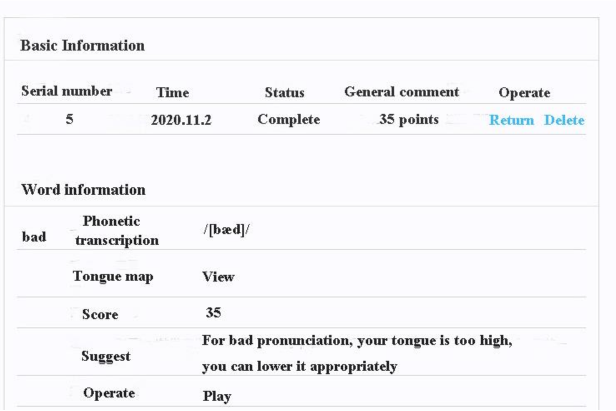


Fig. 11. Comparison results of error detection accuracy of three algorithms

3.3 Testing and analysis of the system

Functional test of the system: The functional modules of the system were all tested to ensure they can operate properly. The test results show that each module can operate in a stable manner. Figure 12 shows the detailed diagram of the pronunciation error correction interface (computer version).



Basic Information

Serial number	Time	Status	General comment	Operate
5	2020.11.2	Complete	35 points	Return Delete

Word information

bad	Phonetic transcription	/[bæd]/
	Tongue map	View
	Score	35
	Suggest	For bad pronunciation, your tongue is too high, you can lower it appropriately
	Operate	Play

Fig. 12. Detailed diagram of the pronunciation error correction interface

Analysis of the effect of vocabulary size on the recognition rate of the system: Analysis was also performed on the effect of vocabulary size on the recognition rate of the spoken English teaching system in different environments. Three environments with distinctive features - a quiet laboratory, an outdoor area with relatively few people, and a school area or a noisy restaurant with a large number of people – were selected as the experimental sites. 100 random tests were performed on 50, 100, and 200 isolated words, respectively, with the recognition rate test results shown in Figure 13.

It can be seen that the recognition rate of the system decreased as the environment became noisier and the number of words increased, but the recognition rate of the system stayed at more than 90% all the way.

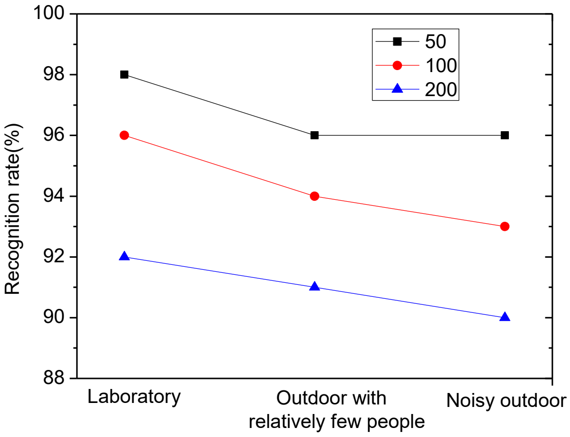


Fig. 13. Test results of the recognition rate

Real-time test on the recognition performance of the system: The real-time recognition of the system means that the system needs to give the recognition result as soon as a user inputs the speech information. This is an important experience for the users and an important indicator for evaluating the performance of the system. This paper took the speech data of 5 isolated words as the research objects, and recorded the time from the endpoint detection to the point when the speech recognition result was returned. Figure 14 shows the real-time test results. It can be seen that the average response time for the 5 words was 311 milliseconds, which can meet the needs of the system users.

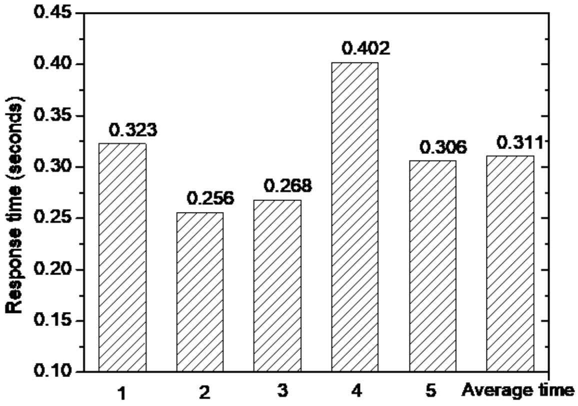


Fig. 14. Real-time test results

4 Conclusion

In order to meet the needs of spoken English learners and give them more effective suggestions on pronunciation to improve their learning quality and efficiency, this paper designed and implemented a spoken English teaching system based on speech recognition and machine learning. The specific conclusions are as follows:

1. The paper analyzed the functional requirements for the spoken English teaching system from the perspective of English learners, and based on this, built a data flow chart of the spoken English learning system, and also determined its overall architecture and functional modules.
2. It elaborated on the detailed design and implementation of the two core modules of the system - speech recognition and speech evaluation. Specifically, MATLAB was adopted to achieve speech recognition and built the speech recognition module, and with the aid of the machine learning algorithm, a DBN-SVM-based pronunciation classification and error detection model was established to perform the error detection, scoring and correction of the speech, and form the speech evaluation module.
3. Several tests were carried out on the attributes of the system, including the pronunciation classification and error detection accuracy, the recognition rate under different environments and with different vocabulary sizes, and the real-timeliness of recognition. The test results show that the system achieved good performance, realized the preset design goals and satisfied the users' requirements.

5 Acknowledgement

Building the "Golden Course" of College English under the U Campus Platform (2019GJJG605).

6 References

[1] Hai, Y.F. (2020). Computer-aided teaching mode of spoken English intelligent learning based on speech recognition and network assistance. Journal of Intelligent and Fuzzy Systems, 39(4): 5749-5760. https://doi.org/10.3233/jifs-189052

[2] Elouali, A., Elberrichi, Z., Elouali, N. (2020). Hate speech detection on multilingual twitter using convolutional neural networks. Revue d'Intelligence Artificielle, 34(1), 81-88. https://doi.org/10.18280/ria.340111

[3] Huang, W. (2021). Simulation of English teaching quality evaluation model based on gaussian process machine learning. Journal of Intelligent and Fuzzy Systems, 40(2): 2373-2383. https://doi.org/10.3233/jifs-189233

[4] Zhang, F. (2020). Innovation of English teaching model based on machine learning neural network and image super resolution. Journal of intelligent and Fuzzy Systems, (2): 1-12. https://doi.org/10.3233/jifs-179953

[5] Wang, S., Mu, M. (2021). Exploring online intelligent teaching method with machine learning and SVM algorithm. Neural Computing and Applications, (6): 1-14. https://doi.org/10.1007/s00521-021-05846-6

[6] Lin, Q., Zhu, Y., Zhang, S., Shi, P., Guo, Q., Niu, Z. (2019). Lexical based automated teaching evaluation via students' short reviews. Computer Applications in Engineering Education, 27(1): 194-205. https://doi.org/10.1002/cae.22068

[7] Liu, H., Chen, R., Cao, S., Lv, H. (2020). Evaluation Of College English Teaching Quality Based On Grey Clustering Analysis, International Journal of Emerging Technologies in Learning, 16(2): 173-187.

[8] Skowronski, M.D., Harris, J.G. (2006). Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. Journal of the Acoustical Society of America, 119(3): 1817. https://doi.org/10.1121/1.2166948

[9] Batle, J., Ciftja, O., Naseri, M., Ghoranneviss, M., Farouk, A., & Elhoseny, M. (2017). Equilibrium and uniform charge distribution of a classical two-dimensional system of point charges with hard-wall confinement. Physica Scripta, 92(5): 055801.

[10] Zhou, N. R., Liang, X. R., Zhou, Z. H., & Farouk, A. (2016). Relay selection scheme for amplify ‐ and ‐ forward cooperative communication system with artificial noise. Security and Communication Networks, 9(11): 1398-1404.

[11] Sechidis, K., Fusaroli, R., Orozco-Arroyave, J.R., Wolf, D., Zhang, Y.P. (2021). A machine learning perspective on the emotional content of parkinsonian speech. Artificial Intelligence in Medicine, 115: 102061. https://doi.org/10.1016/j.artmed.2021.102061

[12] Mathew, J. L., John, T.J., Parakh, A. (2015). Intermittent short course rifapentine-isoniazid combination for preventing tuberculosis in children. Indian Pediatrics, 52(5): 424-5. https://doi.org/10.1007/s13312-015-0648-4

[13] Singh, M.K., Nandan, D., Kumar, S. (2019). Statistical analysis of lower and raised pitch voice signal and its efficiency calculation. Traitement du Signal, 36(5): 455-461. https://doi.org/10.18280/ts.360511

[14] Li, H. (2020). Improved fuzzy-assisted hierarchical neural network system for design of computer-aided English teaching system. Computational Intelligence. https://doi.org/10.1111/coin.12362

[15] Duan, R., Wang, Y., Qin, H. (2020). Artificial intelligence speech recognition model for correcting spoken English teaching. Journal of Intelligent and Fuzzy Systems, 40(1): 1-12. https://doi.org/10.3233/jifs-189388

[16] Wang, Y., Gales, M.J.F., Knill, K.M., Kyriakopoulos, K., Malinin, A., Dalen, R.C.V., Rashid, M. (2018). Towards automatic assessment of spontaneous spoken English. Speech Communication, 104: 47-56. https://doi.org/10.1016/j.specom.2018.09.002

[17] Roe, D.B., Sproat, R.W., Pereira, F.C.N., Riley, M.D., Macarron, A. (1992). A spoken language translator for restricted-domain context-free languages. Speech Communication, 11(2-3): 311-319. https://doi.org/10.1016/0167-6393(92)90025-3

[18] Qi, Y., Dong, B., Ge, F., Yan, Y. (2012). Text-independent pronunciation quality automatic assessment system for English retelling test. The Journal of the Acoustical Society of America, 131(4): 3234. https://doi.org/10.1121/1.4708063

[19] Sakti, S., Markov, K., Nakamura, S. (2007). Incorporating knowledge sources into a statistical acoustic model for spoken language communication systems. IEEE Transactions on Computers, 56(9): 1199-1211. https://doi.org/10.1109/tc.2007.1069

[20] Lin, L., Liu, J., Zhang, X., Liang, X. (2021). Automatic translation of spoken English based on improved machine learning algorithm. Journal of Intelligent & Fuzzy Systems, (Preprint), 40(2): 1-11. https://doi.org/10.3233/jifs-189234

7 Authors

Fengming Jiao, female, was born on 10th, October, 1991 in Tangshan city, Hebei Province. She received her M.A. from Yunnan Minzu University and started her teaching career since 2016 in North China University of Science and Technology. She has published 10 papers on English teaching, attended 4 projects of provincial level, and is in charge of 1 project of provincial level now. tstsdah@163.com .

Jiao Song, female, was born on 17th, Oct.1987 in Tangshan City, Hebei Province. She received B.A. degree from Hebei Normal University in 2010 and M.A. degree of Foreign Language and Applied Linguistics from Hebei United University and started her teaching career since 2013. She has published 10 papers on English teaching and translation strategy. Up to now, she led 1 project of provincial level and 2 projects of municipal level. dianasong123@126.com

Xin Zhao, received her master's degree in foreign language and applied linguistics in 2011 from Yanshan University. Now she is a lecturer in North China University of Science and Technology. Her current research interests include intercultural communication and college English teaching. dongdonglove526@163.com

Ping Zhao, female, was born on 28th, March,1985 in Hengshui city, Hebei Province. She received her B.A. from Jitang College of North China Coal Medical University in 2009 and M.M. from Hebei United University in 2014. She started her teaching career in 2013 in Jitang College of North China University of Science and Technology. She has published 4 papers on English Teaching, led 1 project of provincial level. tstsdah@163.com

Ru Wang, female, was born on 19th August, 1980 in Inner Mongolia, China. She received her M.A. of Foreign Linguistics and Applied Linguistics in Hebei United University and started her teaching career since 2013. She has published over 10 papers on English teaching and led 1 project of provincial level. wangrubobo@163.com

Article submitted 2021-04-19. Resubmitted 2021-05-28. Final acceptance 2021-06-02. Final version published as submitted by the authors.