ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

English listening and speaking test platform for Chinese high school students

Zirui Liu

Southwest University, Zhenyuan, China

Abstract. In the era of rapid technological development, the use of artificial intelligence technology to help education can not only quickly improve the efficiency of work, but also help students improve their grades and academic performance. This study proposes to build an English listening and speaking test platform that provides automatic scoring of high school English speaking exams for Chinese high school students and gives scores and final scores for each scoring index, thus reducing the workload of examiners and allowing students to practice for their questions to improve their performance.

Keywords: Automatic scoring system； Automatic speech recognition； Decision tree ； PCA； GM(1,1).

1. Introduction

The Chinese college entrance exam in English consists of four parts: listening, reading comprehension, language knowledge application, and writing. In countries where English is not the official language, there is a great demand for speaking practice, and in the context of quality education, students must be able to conduct daily conversations in addition to written English. To help Chinese high school students improve their English-speaking performance, provide more opportunities for students to practice, and improve the efficiency of examiners' marking, this study designs an automatic English-speaking scoring system to help complete the teaching of English-speaking.

This study focuses on the human-computer mode of the English-speaking test. In the "human-computer" mode, the computer plays the role of the examiner instead of the English teacher, and all questions and instructions are given by the computer. The computer automatically records and saves all the voice messages of the candidates when they answer the questions in the test. No human factors are involved in the entire process of the candidate's examination. At the end of the test, all test takers' test recordings will be centralized and the markers will conduct uniform centralized marking according to the online marking requirements of the Ministry of Education[1]. This study proposes an English listening and speaking test platform to learn the marking requirements of the marking teachers to achieve automatic scoring of the speaking test, the test platform interface is shown in Figure 1, such as travel type questions, candidates click on the play button, the black part for the computer to automatically play the audio when to the blue space part need to be clicked by the candidate to record the button to start to answer.

This study is divided into 5 main parts. The first part is to use automatic speech recognition technology to recognize and convert the recorded information of the test takers' responses, use a third-party software development kit to convert the speech information into text, and finally save the text in the form of text and train the machine to perform feature extraction to score the test takers. In the second part, a principal component analysis and a grey prediction model were developed to rank the importance of several indicators in the scoring criteria and to predict the students' performance after using the platform based on the existing test scores to evaluate the

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

effectiveness of the platform in helping students to improve their performance. The third section uses decision trees to analyze students' scores on each of the scoring criteria and to guide students' preparation for the test to improve their performance on the speaking test. The fourth part is an evaluation model of the listening and speaking test platform using a questionnaire to evaluate the platform designed for this study. The fifth part is an experiment with controlled variables to investigate the effect of this platform on students' performance by using this test platform for one group of subjects to study and another group of subjects to study independently while ensuring that other conditions are the same.

Figure 1 English listening and speaking test platform interface

2. Build an automatic scoring system

2.1 Automatic Speech Recognition (ASR)

Speech recognition research began in the 1950s, to the 1960s to form the initial ideas then to the 1980s HMM models and artificial neural networks. 90s speech recognition technology began to move from the laboratory to practice, after the introduction of the Hidden Markov Model to speech recognition technology and began to gradually appear in Microsoft, Google, and other language recognition systems[2]. So far, the development of speech recognition technology has matured, and now more people use algorithms and models such as HMM, deep learning, and neural networks to implement speech recognition[16].

However, in practice, technicians use different ways to implement speech recognition, and the operation of these methods often involves a lot of subjective judgments as well as the need to consume a lot of time and effort when extracting features[14].

SDK (Software Development Kit) is a software toolkit integrated for creating application software, generally containing proprietary software packages, frameworks, operating systems, and hardware platforms[15]. The KODA SDK is a toolkit from KODA based on a collection of multiple functions. In this study, we have access to the KDDI SDK and can easily convert large segments of speech into text by simply programming the speech set file, converting each test taker's answer into text, and saving it as a text set with high accuracy and fast conversion of speech information into text information. For example, when a test taker answers "Yes, I like it." then yes, i like it." is output to the text which is saved in lowercase to reduce errors in the language processing later.

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

2.2 Automatic scoring system

Regardless of the format, the scoring rubric is an operational criterion that reflects the test designer's understanding of the language performance of test takers at different levels of proficiency[17,18]. There are various ways to develop scoring criteria, such as the quantitative approach. The CEFR was developed in the unique socio-economic and educational culture of Europe and has been adapted to the socio-economic and educational culture of the European Union[19], but it was developed through a multifaceted Rasch analysis of a large pool of ready-made, intuitively based descriptors[20]. The empirically derived, binary-choice, boundary definition scales (EBB) was chosen for this study[21,22]. The difference is that it is not developed by meticulously analyzing the candidate corpus, but by drawing on Thurstone's Method of Paired Comparisons and Kelly's Repertory Grid Technique. Grid Technique)[23], in which experts judge the level of candidates' real oral corpus and formulate key features that can classify the sample into specific levels.

In this study, students were scored for each speaking test according to the 3 dimensions of coherence, accuracy, and task fulfillment according to the scoring scale developed by EBB.

For the scoring of coherence, the computer automatically detects the time of no language between every two words of the student's answer, and if there is a gap of more than 5 seconds, it means that the student has a language incoherence, and the score will be deducted according to the number of times of incoherence.

For the scoring of accuracy, two parts were scored: grammar and whether the use of words was on-topic. For the grammar section, the study used Grammarly, a third-party grammar-checking software, to check the grammar of the text set, which Grammarly claims to be able to not only correct errors but also improve the style of the text. After the Grammarly check, the number of errors in terms of engagement, correctness, and clarity was obtained, and if the total number of errors was too high, it indicated that the student had a poor grasp of grammar and scored low in the grammar section. For the deduction section, the weighting was calculated for words or phrases in the text using Term Frequency-Inverse Document Frequency (TF-IDF)[4]. Some words may occur frequently in the text, such as the word "it", but they also occur frequently in other texts, so their TF-IDF values will be small. If a word appears many times in the text and rarely in other texts, such as the word "biology", it will be given a high TF-IDF value and will be selected as a keyword to characterize the text. The five keywords with high TF-IDF values were selected to calculate the mean cosine distance between them and the keywords of the topic given by the computer, and the smaller the mean value, the higher the relevance of the vocabulary used by the students to the requirements of the topic.

For task fulfillment scoring, the scoring is judged from the number of words answered by the test taker in the allotted time and the vocabulary diversity. Mikolov et al. proposed word2vec based on NNLM to convert the words in the text into word vectors. [5]A word vector means digitizing the characters and representing each word as an n-dimensional vector, with different words having a value of 1 in one dimension and 0 in the other dimensions. e.g. "love" is represented as [0,0,0,1,0,0,0,...,0], and "hate" is represented as [0,0,0,0,0,...,0]. "hate" is represented as [0,0,0,0,0,1,0,0,0,...,0][4]. word2vec gives a word vector to each word after training the corpus, and if the vector distance of two words is shorter, the more similar the meaning of the two words is. For example, the vector distance between "potato" and "sweet potato" is shorter than the vector distance between "potato" and "watermelon". The vector

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

distance between "potato" and "sweet potato" is shorter than that between "potato" and "watermelon. The K-means algorithm[6] is used for word clustering. 5 keywords selected by TF-IDF are used as cluster centers, and each cluster center represents a category. The vector distance between the word vector of each word and the word vector of the cluster center is calculated, and each word is assigned to the category to which it belongs according to the distance, and the set of 5 clusters is output. Two features can be extracted from each category after word clustering, which is the number of words under the category and the distribution of words. The distribution of words refers to the diversity of words used. If multiple words in the text express the same category of words, such as "very" and "extremely", it means that the candidate has a good vocabulary. Students' task fulfillment scored according to the number of features extracted, and the higher the number of features, the higher the score for the section.

2.3 Scoring Criteria

An assessment scale is a two-dimensional form (e.g. Table 1) that is a set of criteria developed by a teacher or assessment professional to evaluate student performance in terms of both assessment indicators and grade levels[5].

Table 1 Evaluation scales of English speaking tests in Chinese high schools

Evaluation Indicators	A (exemplary)	B (developing)	C (poor)
coherence	The number of inconsistencies is less than 3, the number of Grammarly errors is less than 3, the mean value of the IF-1DF part is less than 1, the number of words is greater than 50, and the vocabulary diversity feature value is more than 5.	The number of inconsistencies is 3-5 times, the number of Grammarly errors is 3-5, the average value of the IF-1DF part is 1-2, the number of words is about 50, and the vocabulary diversity feature value is above 3.	The number of inconsistencies is more than 5, the number of Grammarly errors is more than 5, the mean value of the IF-1DF part is greater than 2, the number of words is less than 40, and the characteristic value of vocabulary diversity is less than 3.
accuracy			
task fulfillment			

3. Principal component analysis model and gray prediction model

3.1 Principal component analysis model

Principal component analysis (PCA) is a method to reduce multi-index and multi-variable to a few main indexes.[8] Its basic idea is to keep the main information of the original data, at the same time, transform the effective feature components with less dimension to represent the multi-dimensional data set, to achieve the purpose of optimal variance.

The design of the sample matrix (assuming that there are n evaluation objects and m evaluation indexes), then a single sample is:

$$X_j = (X_{i1}, X_{i2}, X_{i3}, \cdots, X_{im})$$

The whole sample space is constructed as follows：

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

$$X = \begin{bmatrix} X_{i1} & \cdots & X_{im} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix}$$

• By standardizing the sample matrix, the influence of each index is eliminated：

$$X = \begin{bmatrix} a_{i1} & \cdots & a_{im} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}$$

Among： $a_{ij} = \frac{(x_{ij} - \bar{x_j})}{\sqrt{\frac{(x_{ij} - \bar{x_j})^2}{n}}}$ $\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ (i=1,2, ⋯,n; j=1,2, ⋯,m)

• Analyze the correlation of the index system. The statistical method to study the closeness of each evaluation index is called correlation analysis. After feature extraction of evaluation indexes, the information of evaluation objects is often reused, which is determined by the correlation between various quantitative indexes, but this situation will reduce the scientificity and rationality of evaluation. To avoid this situation, it is necessary to carry out correlation analysis on the evaluation indexes, that is, through the analysis of the correlation coefficient of each index, calculate their correlation coefficient, and subtract the evaluation indexes with larger correlation coefficient according to the correlation principle, to eliminate the influence of information repetition reflected by them on the results and ensure the rationality and effectiveness of the evaluation index system. The correlation coefficient matrix of the sample matrix is obtained as follows:

Among： $r_{ij}$is the original variablex_i and x_j, the correlation coefficient between them. The matrix model is as follows:

$$r_{ij} = \frac{\sum_{k=1}^{n} (a_{ki} - \bar{a_i})(a_{k_i} - \bar{a_j})}{\sqrt{\sum_{k=1}^{n} (a_{ki} - \bar{a_i})^2 \sum_{k=1}^{n} (a_{ki} - \bar{a_j})^2}}$$

Among:$\overline{a_1} = {(\sum_{k=1}^{n} a_{ki})}/{n}$ (i=1,2, ⋯,n; j=1,2, ⋯,m)

• Eigenvalues and eigenvectors of correlation coefficient matrix. We can find the eigenvalue λ by solving the characteristic equation$|\lambda k - R|$. They are arranged in the order of size, i.e.$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda m \geq 0$; and then calculate the corresponding eigenvaluesλ_i and eigenvectoru_i， M new index variables are composed of eigenvectors:

$$y_m = u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + \cdots + u_{mm}\tilde{x}_m$$

• Calculate the contribution rate and cumulative contribution rate of the key evaluation indicators of higher education quality (that is, the contribution degree of the key evaluation indicators to the original evaluation indicators).

$b_j = \frac{\lambda_i}{\sum_{k=1}^{m} \lambda_k}$ (j=1,2,3, ⋯,m)

Main component y_i information contribution rate :

$$\alpha_P = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

The cumulative contribution rate of the main component is y_p. When α_P is close to 1 (generally α_P = 0.85, 0.90, 0.95), the first P index variables y are selected as the y_p principal

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

components instead of the original m index variables, so that the P principal components can be compositely analyzed.

Calculate the composite score:

$$Z = \sum_{j=1}^{P} b_j y_j$$

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

The scores corresponding to the 3 scoring indicators of each student after taking multiple tests on the platform were analyzed by PCA, and the combined scores of the 3 indicators were calculated and ranked. The ranking of the combined scores represents the ranking of the importance of each indicator in influencing the final score of the candidate, and the indicator with the highest score is the indicator that has the greatest influence on the candidate's performance. If a student scores PCA on the 3 indicators after 5 exams and gets the scores on the 3 indicators if the student has the highest score on the number of words and the lowest score on semantics, it means that the number of words that the student said in each exam has the greatest impact on the student's current performance and is the main reason for the student's current performance if the student wants to further improve the exam performance, he/she should strengthen the semantic content If the student wants to further improve his or her test score, he or she should strengthen the practice of semantic content, such as improving the connection between the answer and the question, and making the answer more relevant to the topic of the question.

3.2 Grey prediction model

The gray prediction method first generates numbers, and the commonly used generation equation in gray theory is AGO(Accumulated Generating Operation).

An AGO accumulation is performed on the original data to provide intermediate information for modeling and weaken the randomness of the original time series. [9]

Suppose the time series $X^{(0)}$ has n observations:

$$X^{(0)} = \{X^{(0)}(1),\ X^{(0)}(2),\ \cdots,\ X^{(0)}(n)\}$$

Generate a new sequence by AGO accumulation:

$$X^{(1)} = \{X^{(1)}(1),\ X^{(1)}(2),\ \cdots,\ X^{(1)}(n)\}$$

among them, $X^{(1)} = \sum_{t=1}^{i} X^{(0)}(t), (i = 1,2,\cdots,\ n)$

Then the corresponding differential equation of the GM (1,1) model is:

$$\frac{\mathrm{d}X^{(1)}}{\mathrm{d}t} + \alpha X^{(1)} = \mu$$

Construct cumulative data matrix B and constant vector Y_n:

$$B = \begin{bmatrix} -\frac{1}{2}[X^{(1)}(1) + X^{(1)}(2)] & 1 \\ -\frac{1}{2}[X^{(1)}(2) + X^{(1)}(3)] & 1 \\ \vdots & \\ -\frac{1}{2}[X^{(1)}(2) + X^{(1)}(n)] & 1 \end{bmatrix}, \quad Y_n = \begin{bmatrix} X^{(0)}(1) \\ X^{(0)}(2) \\ \vdots \\ X^{(0)}(n) \end{bmatrix}$$

Use the least squares method to obtain the development gray number α and the endogenous control gray numberμ

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

$$\widehat{\alpha} = \begin{bmatrix} \alpha \\ \mu \end{bmatrix} = (B^T B)B^{-1} Y_n$$

Substituting gray number equations into time differential equations $\frac{\mathrm{d}X^{(1)}}{\mathrm{d}t} + \alpha X^{(1)} = \mu$, Solve the differential equation to find the time function

$$\widehat{X}^{(0)}(t+1) = \widehat{X}^{(1)}(t+1) - X^{(1)}(t)$$

Accuracy test, using the residual test to find residual percentage.

According to the forecast, the model to calculate $\widehat{X}^{(1)}(t)$, and b-b generated $\widehat{X}^{(0)}(t)$, and then calculate the original sequence $X^{(0)}(t)$ and $\widehat{X}^{(0)}(t)$sequence of absolute error and relative error sequence:

$$\Delta^{(0)}(i) = \left|X^{(0)}(t) - \widehat{X}^{(0)}(t)\right|, (i = 1,2,\cdots,n)$$

$$\Phi(\mathrm{i}) = \frac{\Delta^{(0)}(\mathrm{i})}{X^{(0)}(t)} \times 100\%, (i = 1,2,\cdots,n)$$

The standard deviation of the original sequence: $s_1 = \sqrt{\frac{\sum [X^{(0)}(t) - \overline{X}^{(0)}]^2}{n-1}}$

Absolute error standard deviation: $s_2 = \sqrt{\frac{\sum [\Delta^{(0)}(t) - \overline{\Delta}^{(0)}]^2}{n-1}}$

The variance is the ratio of the absolute error standard deviation to the original sequence standard deviation: $C = \frac{s_1}{s_2}$

Small error probability: $P = p\{|\Delta^{(0)}(\mathrm{i}) - \overline{\Delta}^{(0)}| < 0.6475 s_1\}$

The evaluation criteria are shown in Table 5-2:

Table 2 Inspection and evaluation criteria table

Precise level	Excellent	Good	Qualified	Unqualified
P	>0.95	>0.80	>0.70	≤0.70
C	<0.35	<0.60	<0.65	≥0.65

If the accuracy meets the requirements, the model can be used for prediction. If the accuracy cannot meet the requirements, a residual correction model is still needed to improve the accuracy, and then the modified model is used for prediction.

Establish a residual model, using the residuals in the original series:

$$g^{(0)}(t) = X^{(1)}(t) - \widehat{X}^{(1)}(t)$$

The residual model is combined with $\widehat{X}^{(0)}(t+1)$, which is the revised model:

$$\widehat{X}^{(1)}(t+1) = \left[X^{(0)}(1) - \frac{\mu}{\alpha}\right] \mathrm{e}^{-\alpha t} + \frac{\mu}{\alpha} + \delta(t-i)(-\alpha)\left[g^{(0)}(1) - \frac{\mu}{\alpha}\right] \mathrm{e}^{-\alpha t}$$

If a student's future score is predicted to be high and stable, it means that the student has a good grasp of the test content and only needs to continue to maintain it. If a student's score is predicted to decrease, the student needs to practice more, and the specific training can be based on the weakness indicators derived from the principal component analysis model.

4. Decision Tree-based Education Data Mining

A decision tree is a method in machine learning with a tree-like structure that approximates a discrete-valued structure, where each internal node represents a judgment on an attribute, each branch represents the output of each judgment result, and finally, each leaf node represents the result of classification.[10]

Firstly, sample data were collected and all students using this English listening and speaking test platform were selected as the data sample target, and the data contained the

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

number of users, students' age, gender, number of times they took the test, each index score, and final score.

The second step is to perform sample data aggregation.

The age, gender, number of tests taken, individual scores, and final scores of each indicator for each student were summarized in an excel sheet, and the summary data were preprocessed. Since decision trees are more accurate for numerical data analysis, all data were Americanized, such as 1 for male students and 0 for female students in the gender item.

The third step is to define the decision tree model training objectives and representative features. The final performance level is defined as the decision tree training objective, and the age, gender, number of tests taken, and each index score are used as representative features for decision tree training.

The fourth step is to train the decision tree model, calling the Decision Tree Classifier in the python algorithm package for decision tree model construction, to prevent the decision tree over-fitting, the node size of the number of leaves is given in advance.

The fifth step is to extract the decision tree classification rules, and the completed decision tree can be visualized to derive the classification rules, such as the final score of B for those who have taken the test more than 5 times and have a semantic score level of B, etc. According to these rules, students' behavior in using the testing platform can be analyzed to further derive what factors affect students' performance.

5. Embedded experiments

Twenty students from the same natural class in the same grade were randomly selected from a high school in a region of China, and these 20 students were divided into 2 groups of 10 students each. For a period of 3 months, one group of students was trained twice a day using the English listening and speaking test platform designed for this study, while the other group studied independently for the same amount of time without using any other application platform and received the same content tutoring from the teacher in the classroom for the rest of the time.

The scores of 4 scoring indexes and final scores of 20 students before and after the experiment were collected, and after 3 months of experimentation, the students' scores were visualized and analyzed by plotting the score dynamics of each score tracing points of each student, and the student’s performance was assessed by the slope size and positive or negative according to the calculated image slope, and if the slope was positive and the number was larger, it indicated that the student's performance had significantly improved before and after the experiment.

The results showed that the slope of the straight line was positive after fitting the dynamic graph of 20 students' performance, indicating that each student's performance improved when using the automatic scoring system, proving that the testing system is useful and reliable in improving students' oral performance.

6. Evaluation Model

The purpose of this study is to realize the automatic scoring of English-speaking tests in Chinese high schools and help high school students to improve their speaking performance and improve the efficiency of examiners in marking papers[24]. The questionnaire of this automatic

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023

scoring system was created based on the questionnaire survey of the MU platform and was distributed to 20 students who participated in the test in three dimensions: scientific, feasible, and user-friendly. ", "average", "unsatisfactory", and "very unsatisfactory", with the values of 5, 4, 3, 2, and 1 respectively. Each questionnaire is worth 40 points.

After the questionnaires were returned and the questionnaire scores were calculated, the average score exceeded 37, indicating that the test platform was rated relatively high overall. The average score of the user-friendly dimension is more than 4, which means the platform is easy to operate and easy to use. The average score on the scientific dimension is just over 13.5, which can be more to include various provinces and years of questions for the students to practice. For feasibility, the average score exceeds 18, which means that the platform's scoring standard is reliable and the error between the automatic scoring and the actual exam is within the acceptable range of students.

Table 3 Questionnaires on the platform for speaking tests

Dimension	Question
Humanized	Are you satisfied with the ease of operation of the test platform?
Scientific	Are you satisfied with the abundance of test platform question bank resources?
	Are you satisfied with the difficulty setting of the test platform questions?
	Are you satisfied with the number of test platform questions?
viability	Are you satisfied with the testbed scoring rules?
	Are you satisfied with the test platform results?
	Are the weakness dimensions proposed by the test platform accurate?
	Are you satisfied with the difficulty of the test platform and the actual exam?

7. Conclusion

In this study, by building an English listening and speaking test platform, the use of speech automatic recognition technology to convert students' recorded information into words, the training machine scores the text information. To better identify students' deficiencies in the exam, the principal component analysis and gray prediction model are used to analyze the scores and final scores of students in various scoring indicators to promote the development of student performance. To better derive the rules that affect students' grades, decision trees are used to mine data in an educational environment, and information such as students' age, gender, number of tests taken, and grades are mined and analyzed. To enable the public to better evaluate the reliability of the platform, the platform is evaluated using questionnaire surveys, and the platform is further debugged according to the feedback of examiners, classroom teachers, and students. This platform can use the use and not applicable students of the grade control and the use and application of the same score but a group of analysis according to the platform recommendations, a group of self-study to carry out control experiments, according to the results of these two experiments can get the test platform can improve student performance and whether it can allow accurate students to correct deficiencies.

References

[1] Ma Jiyang,Ma Linlin. Exploration of computer-aided spoken English examinations in college entrance examinations--a case study in Ningxia Hui Autonomous Region[J]. Journal of Ningxia Normal College,2019,40(05):78-81. (Chinese)

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023
[2] Li Xiao ridge,Wang Mengjie. The construction and research of teaching mode of simultaneous interpretation ability cultivation based on speech recognition APP--Take KODA Xunfei Language Note APP as an example[J]. Foreign language e-learning,2018(01):12-18.(Chinese)

[3] Luan Di. Design and implementation of an automatic scoring system for English essays in Chinese examinations [D]. Huazhong University of Science and Technology,2021.DOI:10.27157/d.cnki.ghzku.2021.005358. (Chinese)

[4] Liu Lu,Peng Tao,Zuo Wanli,Dai Yaokang. A clustering-based approach to PU active text classification[J]. Journal of Software,2013,24(11):2571-2583. (Chinese)

[5] Liu, Haokun. Research and design of automatic scoring algorithm for English composition[D]. University of Science and Technology of China,2018.(Chinese)

[6] Tadayoshi Fushiki. Estimation of prediction error by using K-fold cross-validation[J]. Statistics and Computing, 2011, 21(2) : 137-146.

[7] Zhang ZY, Yang TZ, Xu AR. Research on building a mathematical electronic technology course assessment system for engineering certification based on Rubric[J]. University Education,2020(09):21-24. (Chinese)

[8] Zhang Qijia. Short-term load forecasting of power system based on PCA neural network [D]. Lanzhou Jiaotong University,2017. (Chinese)

[9] Li Yanru,Li Meng,Liu Shuang. Prediction of birth population development trend and influencing factors in Shanghai - based on GM(1,1) gray prediction model[J]. Journal of Economic Research,2022(24):72-74. (Chinese)

[10] Yang Xiaojuan. Research on the application of decision tree algorithm in the analysis of students' course grades[D]. Yunnan Normal University, 2021. DOI:10.27459/d.cnki.gynfc.2021.001105. (Chinese)

[11] Huang Yan. Research on the application of data mining in the analysis of predicted performance of university examinations [D]. Anhui University,2014. (Chinese)

[12] Ryan Green,Fu Ying. Artificial intelligence encounters grammar problems[J]. English World,2022,41(01):50-53. (Chinese)

[13] Duong, Minh et al. "Two methods for assessing oral reading prosody." ACM Trans. Speech Lang. Process. 7 (2011): 14:1-14: 22.

[14] Gu, Y. W.. Research on speech recognition methods in the context of deep learning for artificial intelligence [J]. Software,2022,43(05):122-124. (in Chinese)

[15] Cheng H-B, Ouyang H-K. Implementation of speech recognition technology based on Linux platform[J]. Internet of things technology,2022,12(10):89-91. doi:10.16667/j.issn.2095-1302.2022.10.026. (Chinese)

[16] Wang Q. Research on Chinese speech recognition system based on deep learning [D]. Shenyang University of Technology, 2022. doi:10.27322/d.cnk

[17] i.gsgyu.2022.000669. (Chinese)

[18] DAVIES A,BROWN A,ELDER C,et al. Dictionary of Language Testing [Z].Cambridge:Cambridge University Press,1999.

[19] MISLEVY R,STEINBERG L,ALMOND R.On the structure of educational assessments[J].Measurement:Interdisciplinary Research and Perspectives.2003,1(1):3-62.

[20] FULCHER G. Testing Second Language Speaking [M].London:Longman,2003.

[21] NORTH B. The Development of a Common Framework Scale of Language Proficiency [M].Peterlang,2000.

[22] TURNER C E. Listening to the voices of rating scale developers:identifying salient features for second language performance assessment[J]. Canadian Modern Language Review,2000,56(4):555-584.

[23] TURNER C E,UPSHUR J A.Rating scales derived from student samples:effects of the scale maker and the student sample on scale contentand student scores[J].TESOL Quarterly,2002,36(1):49-70.

[24] POLLITT A,MURRAY N.What raters really pay attention to[C]//MILANOVIC M,SAVILLE N.Studies in Language Testing 3:Performance testing, cognition and assessment.Cambridge:University of Cambridge Local Examinations Syndicate and Cambridge University Press,1996:74-91.

ISSN:2790-167X DOI: 10.56028/aehssr.4.1.337.2023
[25] Lin ZR, Song JIAN. Research on the satisfaction and influencing factors of teaching reform of college Civics under the background of MU class--an empirical analysis based on 1249 questionnaires from Wuhan University[J]. Ideological and Political Science Course Research,2020(02):112-120.(Chinese)