The Application and Research of Intelligent Mobile Terminal in Mixed Listening and Speaking Teaching of College English

Bo Jiang(✉)

Navigation Technology Department, Tianjin Maritime College, Tianjin 300000, China
LX190625@126.com

Abstract. With the rapid development of mobile technology, the coverage of Wlan, 3G and 4G networks is expanding day by day, and intelligent mobile terminal assisted English teaching and learning has become a hot research field. This study explores the application of intelligent mobile terminals in mixed listening and speaking teaching of college English from three aspects. The first aspect analyzes the application of mobile terminals in the collection of listening and speaking teaching resources. The second aspect analyzes the application of mobile terminals in the recommendation of listening and speaking teaching resources. The third aspect analyzes the application of intelligent mobile terminals in listening and speaking teaching scoring: intelligent mobile terminals extract the relevant features of students' input voice, and use SVR to give students' listening and speaking practice scores, which are presented on the mobile terminal learning page. The results show that the average absolute error is less than 1, indicating that the application of intelligent mobile terminals in the recommendation of college English mixed listening and speaking teaching resources is better. The correlation degree is more than 0.5, which indicates that the accuracy of the evaluation results is high, and the resource recommendation time is always below 80ms, proves the application effect of intelligent mobile terminals in college English mixed listening and speaking teaching.

Keywords: Intelligent Mobile Terminal · College English · Mixed Listening and Speaking Teaching · Application Analysis

1 Introduction

In the "College English Curriculum Requirements" revised by the Ministry of Education in July 2007, the teaching goal of college English is set to "cultivate students' comprehensive English application ability, especially listening and speaking ability". At the same time, it also points out that "colleges and universities should be supported by modern information technology, especially network technology, so that English teaching and learning can be developed towards personalized and self-help learning without being limited by time and place to a certain extent." College English teaching should pay more

L. Yun et al. (Eds.): ADHIP 2023, LNICST 548, pp. 277–294, 2024.
https://doi.org/10.1007/978-3-031-50546-1_19

attention to developing students' listening and speaking skills, At the same time, the ability and habit of autonomous learning in listening and speaking should be cultivated. At present, college English listening and speaking teaching in domestic colleges and universities generally faces the following problems: large class teaching (40–50 class), part of the content in listening and speaking textbooks is disconnected from social development, multimedia language classrooms and other teaching facilities are insufficient, and hardware and software equipment are aging and not updated in time. Influenced by these factors, there are few opportunities for students to really exercise their listening and speaking ability in the listening and speaking class and they can't get feedback from teachers in time, which leads to the decline of students' learning enthusiasm, let alone the cultivation of their habit and ability of autonomous learning. The emergence and popularization of mobile technology has provided favorable conditions for English listening and speaking teaching: mobile devices such as smart phones can make students no longer limited to a fixed learning time, space and mode, and can choose appropriate ways to learn languages at any time and place.

Under the above background, this research will integrate the application of mobile technology after class with the teaching of English listening and speaking in class, and try to build a mixed teaching model of college English listening and speaking courses based on mobile technology, supplemented by a teaching record model that combines the whole process assessment and electronic teaching files, in order to further cultivate students' autonomous learning ability and improve their listening and speaking ability.

2 The Application of Mobile Terminal in the Collection of Listening and Speaking Teaching Resources

For a long time, the classroom teaching of college English for non English majors in China has adopted the lecturing intensive reading mode, ignoring the cultivation of college students' English listening and speaking ability. College students often fail to achieve good results when they really need to communicate in English in their future work and life. Today, with the rapid development of information technology, the deep integration of information technology and English courses has become the core of the current college English teaching reform. How to build an English listening and speaking teaching model that conforms to the characteristics of the subject and the learning rules, and how to cultivate students' English communication ability, has become a major issue facing the current.

The rapid development of mobile terminal technology provides a new opportunity for the reform of college English listening and speaking teaching mode. Mobile terminal devices include smart phones, laptops, tablets, on-board smart terminals, wearable devices and other specific forms. In terms of technology and function realization, mobile terminals have multimedia functions such as audio and video, and intelligent tools supporting data transmission and processing capabilities. It can access the Internet to browse and download information, as well as submit data and interact with roles. At the same time, the mobile terminal is a good helper for learning. It can be equipped with a visual operating system, and can install customized learning software and intelligent companion for various applications. Mobile terminal technology supports learners

to use mobile devices for anytime and anywhere learning. Mobile technology assisted language learning has incomparable advantages in expanding learning time and space, enriching learning interaction, improving learning efficiency, etc. Blending Learning has been studied at home and abroad for a long time. Margret believes that blended learning is the combination or mixing of network technology based schools to achieve a certain teaching goal. It is the combination of multiple teaching methods and teaching technologies to achieve the best teaching results together. It is the combination of teaching technology and specific teaching classes. This paper studies how to effectively use intelligent mobile terminal applications to carry out English listening and speaking teaching. The research is divided into three parts, namely, the application of mobile terminals in the collection of listening and speaking teaching resources, the application of mobile terminals in the recommendation of listening and speaking teaching resources, and the application of intelligent mobile terminals in listening and speaking scoring.

With the rapid development of information technology, computer network has been used more and more in various teaching processes. Among them, the construction of online teaching resources has attracted more and more attention. For example, the gradual development of teaching resource database, teaching website and online course construction has become one of the core contents of education informatization. Therefore, a large number of relevant college English mixed listening and speaking teaching resources can be obtained on the network by using the general search engine of mobile terminals (see Fig. 1).First, we should determine the target teaching content, and then use the general search engine of intelligent mobile terminal to obtain the listening and speaking teaching resources that have a certain degree of relevance to the target teaching content on the network; On this basis, the teaching resources obtained on the network are optimized and sorted, high-quality teaching resources with high relevance and quality to the target teaching content are selected, and a listening and speaking teaching resource database is constructed to facilitate users to query and access at any time[1].

The core of resource database design is to solve the problem of database classification management. According to the classification method of teaching resources of the Ministry of Education, listening and speaking teaching resources are divided into five types according to the types of documents. It covers text, multimedia and file resources. The specific classification is shown in Fig. 2.

Due to the variety of resources, the commonness of resources can be extracted from many resources, and the data structure of resources can be analyzed for resource storage. Therefore, according to the above five types of resources, we abstract two types of data for management, one is text information, the other is file information, and store them in the cloud storage space of intelligent mobile terminals. The connotation of cloud storage is storage virtualization and storage automation. Through cluster application, grid technology or distributed file system and other functions, a large number of different types of storage devices in the network are gathered together to work together through application software to jointly provide external data storage and business access functions. In other words, cloud storage is no longer storage but a service. Its core is to combine application software and storage devices, and realize the transformation of storage devices to storage services through application software. In general, cloud storage is a new concept extended and developed from the concept of cloud computing. Cloud

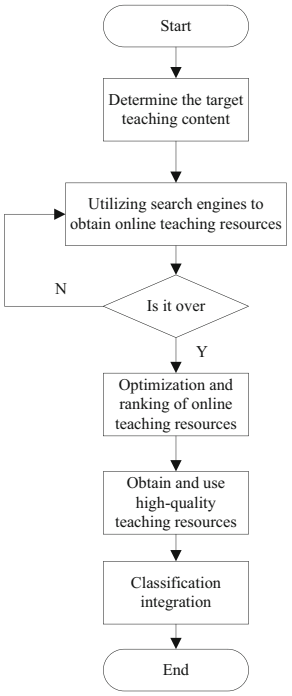

Fig. 1. Collection process of listening and speaking teaching resources based on intelligent mobile terminal general search engine

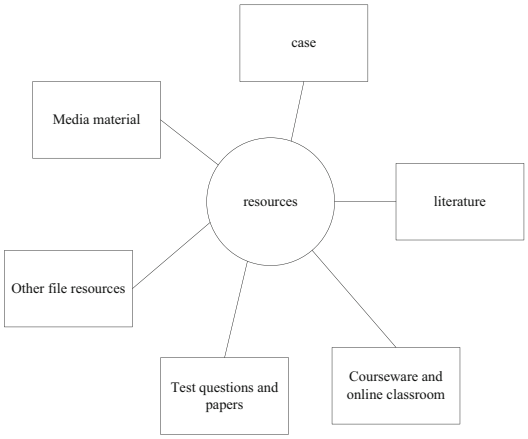

Fig. 2. Classification of College English Mixed Listening and Speaking Teaching Resources

storage is to store resources in a secure large-scale storage server through the network. Wherever you go and use any intelligent mobile terminal, you can access the listening, speaking and teaching resources[2]as long as you can connect to the storage server. The data structure of college English hybrid listening and speaking resources stored in the cloud of intelligent mobile terminal is shown in Fig. 3 below.

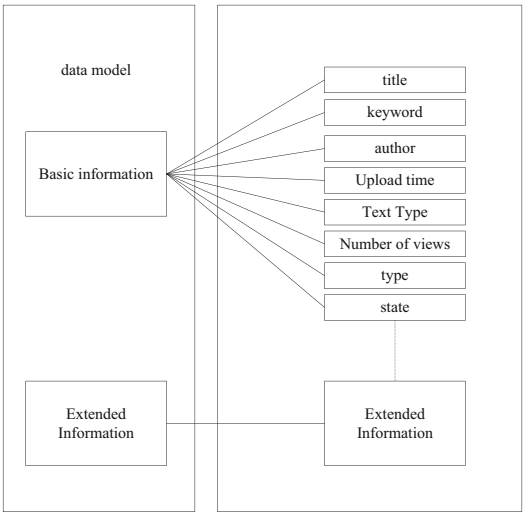

Fig. 3. Listening and speaking resource storage structure based on intelligent mobile terminal

The node consists of two parts. The first part is the basic information part. Each resource has its own basic information to facilitate the administrator's query and statistics. The basic information includes title, keyword, author, upload time, text type, number of views, type, and status. The second part is the extension information. The secondary extension information is divided into a variety of defined types, including text and file class structures. At the same time, the user-defined node type is reserved to prepare for future resource expansion. For text resources, information is directly stored in the defined data structure as an entity, sealed in data objects, and for file type resources, we use the form of storing file data objects, including the original name, current name, size, and physical address of the file.

3 Application of Mobile Terminal in Recommendation of Listening and Speaking Teaching Resources

Personalized recommendation is an important branch of the application of intelligent mobile terminals in college English mixed listening and speaking teaching. It analyzes the user's behavior characteristics, the context of mobile devices, social network relationships and other information through the recommendation program in mobile terminals, predicts applications that users may be interested in, helps users filter resources

in advance, and tries to find applications that match user preferences [3]. The recommendation program in mobile terminal is developed based on the collaborative filtering algorithm. Collaborative filtering algorithm is a commonly used recommendation algorithm. It first calculates the similarity between users, and then recommends items to target users according to the preferences of similar users, so as to complete personalized recommendation, which has certain practicability and effect. Collaborative filtering recommendation algorithms are divided into BaseItemCF (project-based system filtering algorithm) and BaseUserCF (user based collaborative filtering algorithm). These two algorithms calculate user similarity and item similarity, respectively. BaseItemCF is to recommend similar items of interest to users BaseUserCF is to recommend similar items of interest to users. In the collaborative filtering technology, the following assumption is true: if user A and user B have similar interests, then the items that are of interest to Party A may also be of interest to Party B.

The collaborative filtering algorithm is to apply this idea in life to the recommendation system [4]. The application process of mobile terminal recommendation program in listening and speaking teaching resource recommendation is as follows:

(1) Student interest expression

The common interests of students are the basis for recommendation. Therefore, in collaborative filtering algorithms, data processing is mainly based on students' scores of teaching resources rather than content.

Assume that the number of students is m, the number of teaching resources is n, Scoring recommendation matrix A is a $m \times n$ the matrix of i line No j elements of columns a_{ij} indicates a student i for teaching resources j the scoring result of. The design of student teaching resource scoring matrix is shown in Table 1.

Table 1. Scoring matrix of student teaching resources

Students/Resources	1	2	…	n
1	a11	a12	…	a1n
2	a21	a22	…	a2n
…	…	…	…	…
m	am1	am2	…	amn

In the student teaching resource scoring matrix as shown in Formula (1), a_{ij} it is a score of 1–5, representing students i for teaching resources j the scoring result of. a_{ij} the higher the value of, the higher the student's rating of the teaching resource.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \tag{1}$$

(2) Nearest neighbor set selection

The generation of "nearest neighbor set" is to calculate the similarity between students' scores of teaching resources, and establish the nearest neighbor set of students' interests, which is the core of implementing collaborative filtering based recommendation algorithm.The Person formula is mainly used to calculate the relevant similarity of two students i and k similarity between $S_{(i,k)}$ the calculation formula is shown in Formula (2) [5].

$$S_{(i,k)} = \frac{\sum_{i \in C_{(i,k)}} \left(a_{ij} - \overline{a}_i\right)\left(a_{kj} - \overline{a}_k\right)}{\sqrt{\sum_{i \in C_{(i,k)}} \left(a_{ij} - \overline{a}_i\right)^2} \sqrt{\sum_{k \in C_{(i,k)}} \left(a_{kj} - \overline{a}_k\right)^2}} \tag{2}$$

where, $C_{(i,k)}$ on behalf of students i and k common scoring item set of; a_{ij}、a_{kj} on behalf of students i and k listening and speaking teaching resources j scores; $\overline{a}_i$、$\overline{a}_k$ on behalf of students i and k the average score of.

(3) Generate recommendation results

According to the set of students' "favorite neighbors" realized in the previous step, recommend teaching resources to students from the following two aspects:

① Predict students' preference for all teaching resources;

② According to the students' preference for teaching resources, the teaching resources are sorted and N teaching resources that students most like are recommended, namely Top-N recommendation set.

Hypothetical students i the collection of assessed listening and speaking teaching resources is B_i for any teaching resources b does not belong to B_i the formula of the predictive value of the preference degree of is shown in Formula (3).

$$D = \frac{\vec{i} \sum_{k=1}^{K} \sin k + \sum_{k=1}^{K} \sin k(\text{cour}k - \vec{u})}{\sum_{k=1}^{K} \sin k} \tag{3}$$

Among them, $\vec{i}$ indicates a student i the average score of the assessed set of listening and speaking teaching resources, k for students i students of the nearest neighbor set, $\sin k$ indicates a student k and i similarity, $\vec{u}$ for students i for listening and speaking teaching resources b the average recommended score of.

Based on the preference prediction formula as shown in Formula (3), grade all teaching resources' preference, rank the scoring results, and select the highest recommendation set for recommendation[6].

4 Application of Intelligent Mobile Terminal in Listening and Speaking Teaching Scoring

The task of this chapter is to effectively apply the recommended high-quality teaching resources to the process of English listening and speaking teaching. The purpose of English listening and speaking teaching is to train students' listening and speaking ability. Listening and speaking teaching scoring is a technology that students listen to the specified text, then follow or translate into the target language simultaneously, and finally the intelligent mobile terminal feeds back scores according to students' practice results. Its goal is to endow the intelligent mobile terminal with the ability to act as a virtual teacher, conduct a fair, objective and efficient evaluation of students' listening and speaking practice, and alleviate the serious shortage of professional listening and speaking teachers. In learning, it can help students better understand the pronunciation level, improve the efficiency of listening and speaking learning and promote self-study; In examinations, it can assist or replace manual marking of listening and speaking tests, greatly improving the efficiency and quality of marking. The core of the application of intelligent mobile terminal in listening and speaking teaching scoring is to use the relevant algorithms of the voice automatic recognition module in the terminal. The specific application process is as follows:

4.1 Resource Retrieval Module

The intelligent mobile terminal has the functions of course classification, course resource navigation and course index, which provides students with recommended courses of different levels and popularity. Students can learn specific courses according to their own application level and learning needs to improve their sense of interaction with the system; You can also enter the corresponding text through the search box to conduct fuzzy query, presenting the most appropriate and personalized course search results. See Chapter 2 for the specific process.

4.2 Speech Processing

The generation and perception of speech depend on human voice system and auditory system. The speaker first generates voice information in his mind, and then packages the information in the form of rhythm, loudness, pitch cycle rise and fall, that is, language coding operation. After coding, the speaker sends out sound through the cooperation of the vocal organs, and then transmits the voice signal to the listener's ear through the sound wave as the medium. Its auditory system transmits the processed signal to the brain center and converts it into language coding, thus generating semantic information [7]. Speech processing mainly includes two aspects: one is processing speech signals, such as preprocessing speech signals to eliminate most of the useless information; On the other hand, the speech signal is analyzed and the feature parameters are extracted for subsequent learning. It mainly includes the following three aspects:

(1) Preemphasis

Since the energy loss caused by lip radiation is concentrated in the high frequency part, it is necessary to emphasize the high frequency part of speech to make the spectrum smooth. The "pre emphasis technology" is to pass the sampled signal through a FIR high pass filter, and its transfer function is as follows:

$$f(d) = 1 - \beta d^{-1} \tag{4}$$

Among them, β is the pre weighting coefficient, usually $0.9 < \beta < 1.0$. if t the input signal at the moment is $s(t)$, the output signal after pre emphasis $y(t)$ for:

$$y(t) = s(t) - \beta s(t-1) \tag{5}$$

(2) Framing windowing

Speech signal is a typical time-varying signal. It is difficult to study a long segment of speech signal, but in reality, when people speak, the movement of the mouth and throat is a continuous action, and the speed is not fast. According to this characteristic, a long speech signal is usually divided into several short segments using differential thinking for research, and these short segments are called "analysis frames" [8].Although there can be no overlap between frames when framing, this may cause the calculated pitch to jump. In order to prevent jumping and make the voice signal after framing more stable, it is necessary to overlap a part between the two analysis frames. This part is called frame shift. The frame shift should not be too long, generally less than 1/2 of the frame length. The current frame length is 256 and the frame shift is 80.Some window functions are needed for framing. The commonly used window functions are rectangular window and Hamming window. The voice signal can be segmented by moving the window for weighting. The window function selected in this section is Hamming window. The following is the introduction of rectangular window and Hamming window.

Window function of rectangular window $p(l)$ for:

$$p(l) = \begin{cases} 1, & 0 \le l \le L-1 \\ 0, & otherwise \end{cases} \tag{6}$$

where, L indicates the window length; l represents a voice signal frame.

The window function of Hamming window is:

$$p(l) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi l}{L-1}, & 0 \le l \le L-1 \\ 0, & otherwise \end{cases} \tag{7}$$

(3) Endpoint detection

Generally, in order to ensure the integrity of voice information, an intelligent mobile terminal will leave a blank voice segment when recording voice signals. Therefore, the process of endpoint detection is to detect the information of voice segments and eliminate noise segments, so as to determine the starting and ending points of effective voice information, so as to improve the accuracy of subsequent operations [9]. In this

section, the double threshold comparison method is used for endpoint detection. Two performance indicators are needed: short-time energy and short-time zero crossing rate. They are introduced below.

In reality, people's vocal organs are inertial, so the state of voice signals will not change abruptly, and the energy contained in voice signals is different. For example, the energy contained in voiceless and voiced sounds is obviously different, so short-term energy can be used to express the personality characteristics of voice signals. The extraction formula of short-term energy is as follows:

$$E(l) = \sum_{t=1}^{\infty} \left[s(t)p(l-t) \right]^2 \tag{8}$$

Among them, $E(l)$ it represents the No l frame voice signal $s(t)$ short-term energy.

In the time domain diagram of the voice signal, if the voice signal is continuous, when the waveform crosses the time axis, it indicates that zero crossing has occurred; If it is a discrete voice signal, it is necessary to find adjacent sampling points. If one voice signal is positive and the other is negative, it is also considered that zero crossing has occurred. Short time zero crossing rate is defined as:

$$\begin{aligned} e(l) &= \sum_{-\infty}^{\infty} |\mathrm{sgn}s(t) - \mathrm{sgn}s(t-1)|p(l-t) \\ &= |\mathrm{sgn}s(t) - \mathrm{sgn}s(l-1)|p(l) \end{aligned} \tag{9}$$

Among them, sgn[] is a pseudo symbolic function whose expression is:

$$\mathrm{sgn}[s(l)] = \begin{cases} 1, s(l) \geq 0 \\ -1, otherwise \end{cases} \tag{10}$$

Among them, $p(l)$ is a window function, l is the window length. $p(l)$ the expression is:

$$p(l) = \begin{cases} \dfrac{1}{2L}, 0 \leq l \leq L-1 \\ 0, otherwise \end{cases} \tag{11}$$

The double threshold method first sets two thresholds for short-term energy and zero crossing rate, so that endpoint detection can be divided into four stages:

(1) Mute stage: if one of the energy and zero crossing rate is below the low threshold, it is the mute stage;
(2) Transition section: if both parameters exceed the low threshold but none of them enter the high threshold, the transition section will be entered;
(3) Voice segment: if either of the two parameters exceeds the high threshold, the voice segment will be entered;
(4) End segment: if both the energy and zero crossing rate parameters are reduced below the low threshold, and the minimum time threshold is greater than the total time length, it will be marked as a noise signal, and then continue scanning. If the minimum time threshold is less than the total time length, it will be marked as the end point.

4.3 Voice Scoring

When students complete the English listening or voice following test, the intelligent mobile terminal will immediately extract the relevant features of the students' input voice and the corresponding features of the standard voice, and give the score of the students' input voice according to certain scoring rules, which will be displayed on the learning page. This module allows students to understand their own pronunciation level, and constantly improve their English listening and speaking level according to this scoring feedback.

(1) Intelligent mobile terminal extracts voice features

Feature extraction is a very important step before machine learning, which determines the credibility and accuracy of the scoring model. As long as the feature selection is accurate enough, even if the SVR model is not optimal, a scoring result with small error can still be obtained. SVR is a Nonlinear regression algorithm, which can handle the feature extraction task of nonlinear relationship. The model established by SVR has good anti noise performance, can accurately extract features from noisy data, and improves the accuracy of feature extraction. In the scoring model based on SVR studied in this paper, two types of features are mainly extracted: voice features and text features. Speech features are extracted directly from speech signals, and text features are extracted from the output of speech recognition engine of intelligent mobile terminal. Before using relevant technologies to extract features, feature screening[10]is often required. This article refers to the following guidelines when screening features:

① The importance of each feature can be measured by calculating its Pearson correlation coefficient with the manual score. Generally, features with a correlation coefficient lower than 0.2 should not be selected;

② The intelligent scoring system should describe the examinee's spoken language from multiple dimensions. Features with high similarity should not be included at the same time, so we calculate the correlation between the selected features. For each pair of features with a correlation coefficient greater than 0.9, delete one of them;

③ Since this paper is aimed at the mixed listening and speaking teaching score of college English, there is usually no fixed reference answer to this kind of question, so when making feature selection, this paper mainly selects universal features rather than features strongly related to the answers of the reference text, in addition to semantic similarity features. Table 2 shows a brief description of the features finally selected for use in this article.

In this paper, four phonetic features are extracted to evaluate the pronunciation quality, fluency and content richness of candidates' spoken English. Speech speed is mainly used to describe oral fluency, which can be calculated by the following formula:

$$V = \frac{N}{T - T'} \tag{12}$$

where, V represents speaking speed; N it represents the total number of words in students' spoken language, T it indicates the total duration of oral recording, T' indicates the mute duration in the recording. In addition to the speed characteristics, the number of silences in the recording can also reflect the oral fluency of the tester to some extent.

Table 2. Speech Features

Feature category	Feature Name
Phonetic features	articulationRate
	numSilence
	posteriorScore
	speakingRatio
Text-based features	eassyLength
	uniqueWords
	parseTreeDepth
	semanticSimilarity
	goodGrammerRatio

In terms of pronunciation quality evaluation [11], the posterior probability feature of pronunciation is used by many oral scoring systems. This paper also uses this feature to describe the accuracy of examinees' pronunciation. In addition, the time ratio of extracting pronunciation can also reflect the richness of oral content to a certain extent.

$$H = \frac{T - T'}{T} \tag{13}$$

where, H stands for pronunciation time ratio.

In the traditional oral evaluation for reading aloud questions, the standard oral sequence corresponding to the reference text is usually used as the label to force alignment the test speech, and then the average posterior probability of each phoneme is calculated through the classic GOP (Goodness of Pronunciation) algorithm. However, there is no reference text in the open oral scoring, so it is necessary to combine the speech recognition engine and an acoustic model trained with standard English pronunciation to calculate the average posterior probability as the pronunciation quality feature.

At the text level, grammar is the most basic criterion to distinguish the language proficiency of examinees. We use part of speech tags to determine whether there are grammatical problems in the use of words in spoken content. The famous English original novels generally do not have grammatical errors. This paper uses the method in the open source composition scoring system (EASE) Enhanced AI Scanning Engine) for reference to extract ternary tags and quaternary tag combinations from Sherlock Holmes' novel collections after tagging sentence, word and part of speech tags, and store the extracted results locally as a tag combination query database. In feature extraction, we will extract three tag combinations and four tag combinations of each sentence from speech recognition text. For each combination, we query in the tag combination library. If we cannot find it, it is considered that there is a syntax error. We use the following formula to calculate the text syntax accuracy μ (ς is the total number of ternary label combinations and quaternary label combinations contained in the text, ϖ indicates the total number of correct label combinations).

$$\mu = \frac{\varsigma}{\varpi} \tag{14}$$

SVR cannot directly recognize text data and audio data, so it cannot directly use speech recognition text as the input data of scoring model. We need to convert the above data into the numerical tensor form that SVR can handle. Text vectorization refers to the process of transforming text into numerical tensor. This paper will use word embedding technology to vectorize speech recognition text. This paper uses the pre trained word embedding model G1oVe to transform speech recognition text into vector representation. The whole text vectorization process: first, through data cleaning of speech recognition text, onomatopoeia and repeated words in the text due to recognition errors are removed. Then the cleaned text is segmented. Then, the embedded text file of G1oVe words is parsed. The file is a. txt file. Each line word string in the file and its corresponding vector representation. Finally, we build a word embedding matrix that can be loaded into SVR.

(2) Intelligent mobile terminal evaluates listening and speaking quality

Building an appropriate and efficient voice acoustic model is the last stage of the speech recognition system, which has a significant impact on the performance of the system. An ideal speech recognition network should have strong generalization ability and learning ability of sample features. It can learn a large number of training samples, so as to mine the corresponding relationship between various speech feature parameters and speech semantic information, and achieve accurate classification of test samples.

Support Vector Regression (SVR) algorithm is a machine learning algorithm based on structural risk minimization criteria. It makes full use of the advantages of machine learning, and can learn complex data patterns with only limited training samples, thus mapping feature scores to target scores. Therefore, this paper uses SVR algorithm as regression model to achieve effective fusion of multidimensional evaluation features. It is introduced below. Given training data set $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, where, $x_i \in X^n$ indicates that from the i extracted from segment reading voice n dimensional eigenvector, $y_i \in Y$ yes x_i corresponding manual scoring, m is the total number of samples in the training data set. Our goal is to train all sample pairs in the dataset (x_i, y_i), find a regression function that is as flat as possible $y = F(x)$ to approach the relationship between them and minimize the prediction error.

For non-linear data x because it is difficult to be linearly separable in the original space, the SVR algorithm uses a nonlinear function to solve this problem $R(x)$, will x map to high-dimensional feature space for processing. The regression function is defined as:

$$F(x) = \langle v, R(x) \rangle + r \tag{15}$$

Among them, v is the weight vector, r is offset $\langle , \rangle$ it is an inner product operation.

The most widely used SVR algorithm is ε Type of insensitive loss function, defined as:

$$q\big[F(x)-y\big]=\begin{cases}0, |F(x)-y|<\varepsilon\\ |F(x)-y|-\varepsilon, otherwiswe\end{cases} \tag{16}$$

where, positive number ε is the preset error, when the regression function $F(x)$ for the actual target value: ε within the range, i.e. ε in the insensitive zone, the loss is recorded as 0.To measure ε the deviation on both sides of the insensitive zone defines two relaxation variables, namely g_i, h_i, the objective function is:

$$\min \frac{\|v\|^2}{2}+J\sum_{i=1}^{m}(g_i+h_i) \tag{17}$$

The following constraints are met:

$$\begin{cases}y_i-F(x_i)\le\varepsilon+g_i\\ F(x_i)-y_i\le\varepsilon+h_i\\ g_i,h_i\ge 0\end{cases} \tag{18}$$

In the formula, constant J is the penalty coefficient for prediction errors. It can be seen from Formula (16) and Formula (17) that this is an optimization problem, so Lagrange function is constructed. Main variables of Lagrange function v, r、g_i, h_i calculate the partial derivative in turn, and make its value 0.The result of partial derivation is substituted into Lagrange function and transformed into dual optimization problem. Finally, the SVR regression function obtained by solving is:

$$F(x)=\sum_{i=1}^{m}\big(\phi_i+\phi_i'\big)K(x_i,x)+r \tag{19}$$

Among them, $K(x_i,x)$ is a kernel function; ϕ_i、ϕ_i' are two Lagrange multipliers.

Finally, the pronunciation quality evaluation process based on support vector regression algorithm is as follows:

1) Based on students' spoken English pronunciation data, two types of features are extracted respectively, and feature scores are calculated;
2) The cubic polynomial function is used to normalize each feature score calculated to make it consistent with the range of manual scoring;
3) Construct SVR training sample set with multi-dimensional evaluation feature score as input and manual score as output;
4) Training parameters in SVR scoring model;
5) Use the same method to extract the evaluation features of each dimension of the pronunciation to be tested, and then use the trained SVR scoring model to fuse the features, so as to achieve an effective evaluation of the overall pronunciation quality of students.

5 Application Test

5.1 Experimental Configuration

(1) Front end pretreatment configuration

The pre emphasis coefficient used in pre emphasis is 0.97. The Hamming window with a window length of 25 ms is used to smooth the voice frame signal. The duration of each voice signal frame is 25 ms, and the overlap between adjacent voice frames is 15 ms.

(2) Knowledge base configuration

The acoustic model uses a context independent monophone model, and the HMM model corresponding to each phoneme consists of three emission states from left to right. The probability distribution on the 1V} 'CC acoustic eigenvector associated with the HMM state is simulated using GMM containing eight Gaussian components. The language model uses the ternary language model, and the pronunciation dictionary uses the CMU pronunciation dictionary of Carnegie Mellon University (CMU).

(3) Speech recognition engine

This paper is based on the open source intelligent mobile terminal of CML University to develop an assessment model of English pronunciation quality suitable for Chinese students.

5.2 Recommended Test of Listening and Speaking Teaching Resources

Two data sets are used to verify the effectiveness of smart mobile terminals in recommending college English mixed listening and speaking teaching resources. The quality of recommendation is measured by its prediction results, mainly by measuring the accuracy between the system's recommendation results and users' real scores. There are many existing evaluation strategies, among which the average absolute error MAE is easy to understand and easy to calculate, which is the most widely used measurement standard. Therefore, the average absolute error is recommended as the application effect. The results are shown in Table 3 below,

Table 3. Average Absolute Error

Data set	1	2
mean absolute error	0.54	0.27

It can be seen from Table 3 that the average absolute error is less than 1, which indicates that the application effect of intelligent mobile terminals in the recommendation of college English mixed listening and speaking teaching resources is good.

5.3 Listening and Speaking Quality Evaluation Test

The sentences with balanced phoneme coverage in the CMU ARCTIC corpus are selected as the reading text corpus. Every five sentences form a reading passage with a length of about 50 words. Ten students from Guilin University of Electronic Science and Technology are invited to read these passages at a normal speed, and the pronunciation is as clear as possible. Finally, 10 pieces of reading speech data are recorded, Save as 16 kHz sampling rate and 16bit mono WAV format. Invite English teachers from Foreign Languages Institute to give a full score of 100 points to the overall pronunciation quality of these voice data in terms of two types of characteristics. When evaluating pronunciation quality, the manual scoring is usually taken as the reference standard, and the system performance is evaluated by measuring the correlation between machine scoring and manual scoring. The results are shown in Fig. 4 below.

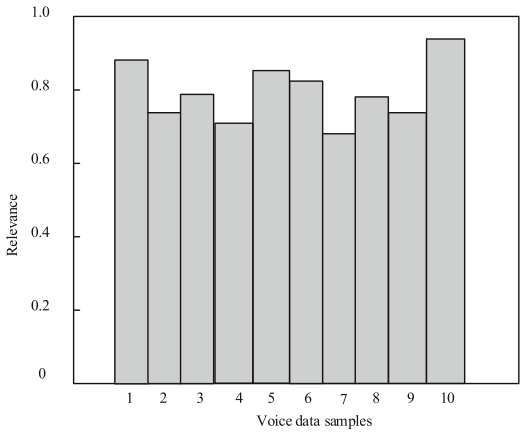

Fig. 4. Correlation

As can be seen from Fig. 4, the correlation degree exceeds 0.5, indicating that the evaluation results are highly accurate, which proves the application effect of intelligent mobile terminals in college English mixed listening and speaking teaching.

5.4 Recommended Time Test for Listening and Speaking Teaching Resources

Based on the above experimental configuration, the recommended time for listening and speaking teaching resources in this method was determined in five groups of experiments, and the specific results are shown in Fig. 5.

From Fig. 5, it can be seen that the resource recommendation time of the method in this article is always below 80ms, and the curve variation is relatively stable, proving that the recommendation efficiency of intelligent mobile terminals in mixed listening and speaking teaching of college English is relatively high.

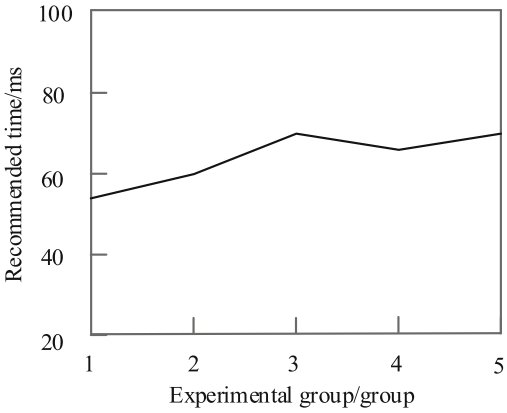

Fig. 5. Recommended time test

6 Conclusion

With the improvement of China's international status and the increasing frequency of international exchanges, the development of the country and society has put forward higher requirements for college students' English ability. The College English Curriculum Requirements issued by the Ministry of Education in 2007 clearly states that the teaching goal of college English is to cultivate students' comprehensive English application ability, especially listening and speaking ability. At present, college English courses in China are in the process of actively exploring the transformation from the traditional "teacher centered" teaching model to the "student centered" teaching model. As the main course to cultivate students' listening and speaking ability, college English listening and speaking courses are no exception. But generally speaking, listening and speaking courses are subject to large class teaching, and the teaching model is single The influence of various factors, such as insufficient teaching facilities and technical means, has not achieved much in improving students' listening and speaking ability, and even led to students' learning enthusiasm for listening and speaking courses getting worse. At the same time, compared with reading and writing courses, many teachers are not so calm when facing listening and speaking courses. However, with the rapid development of information technology and mobile technology, the coverage of Wlan, 3G and 4G networks is expanding, and mobile electronic terminal devices such as smart phones, PDAs and tablets are also becoming more and more popular. Mobile technology assisted English teaching and learning has become a hot topic in English education. Therefore, How to use mobile technology to build a college English listening and speaking teaching model that conforms to the characteristics of English subjects and language learning laws has also become a major research topic. The main contents of this study are summarized as follows:

(1) In this paper, the recommendation algorithm in intelligent mobile terminal is applied to practice to achieve personalized recommendation of listening and speaking teaching resources.

(2) SVR algorithm is introduced to fuse the evaluation features of different dimensions, which significantly improves the overall performance of pronunciation quality evaluation.

There are still some areas that can be improved and expanded in this study. Here are the following as the development directions of future research:

(1) In this paper, the method of feature comparison is used in intonation assessment, that is, each paragraph of students' reading voice needs to have a corresponding reference standard voice. In the future, we can explore the intonation assessment method using statistical modeling.
(2) The input parameters of the scoring model are improved.For the scoring model that needs feature engineering, the dimension of features can be appropriately increased, while removing some features that are less relevant to manual scoring.For the "end-to-end" scoring model, voice data and text data can be converted into better vector representations.

References

1. Gerd, K.: Writing virtual reality teaching resources. Phys. Teacher **61**(2), 107–109 (2023)
2. New Resources in TRAILS: The teaching resources and innovations library for sociology. Teach. Sociol. **51**(1), 108–111 (2023)
3. Yuan, X.: A balanced allocation method of English MOOC teaching resources based on QoS constraints. Inter. J. Continuing Eng. Educ. Life-Long Learn. **33**(1), 84–98 (2023)
4. Lin, Y.: A neural network-based approach to personalized recommendation of digital resources. Comput. Inform. Mech. Syst. **5**(4), 97–101 (2022)
5. Zou, F., Chen, D., Xu, Q., et al.: A two-stage personalized recommendation based on multi-objective teaching–learning-based optimization with decomposition. Neurocomputing **452**(6), 716–727 (2021)
6. Jiang, S., Ding, J., Zhang, L.: A Personalized recommendation algorithm based on weighted information entropy and particle swarm optimization. Mob. Inf. Syst. **2021**(4), 1–9 (2021)
7. Kang, Z., Sadeghi, M., Horaud, R., et al.: Expression-preserving face frontalization improves visually assisted speech processing. Int. J. Comput. Vis. **131**(5), 1122–1140 (2023)
8. Yoo, H., Seo, S., Im, S.W., Yong, G.G.: The Performance evaluation of continuous speech recognition based on Korean phonological rules of cloud-based speech recognition open API Inter. J. Netw. Distrib. Comput. **9**(1), 10 (2021)
9. HyeongJu, N., JeongSik, P.: Accented speech recognition based on end-to-end domain adversarial training of neural networks. Appl. Sci. **11**(18), 8412 (2021)
10. Ahmed, A., et al.: Connecting Arabs: B ridging the gap in D ialectal speech recognition. Commun. ACM **64**(4), 124–129 (2021)
11. Lee, D., Kim, D., Yun, S., Kim, S.: Phonetic variation modeling and a language model adaptation for Korean English code-switching speech recognition. Appl. Sci. **11**(6), 2866 (2021)
12. Liu, S., He, T., Li, J., et al.: An effective learning evaluation method based on text data with real-time attribution - a case study for mathematical class with students of junior middle school in China. ACM Trans. Asian Low-Resour. Lang. Inform. Process. **22**(3), 1–22 (2023)
13. Tan, S., Sun, L., Song, Y.: Prescribed performance control of Euler-Lagrange systems tracking targets with unknown trajectory. Neurocomputing **480**, 212–219 (2022)