International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 3593-3603

https://doi.org/10.70517/ijhsa463303

A Study on the Application of Audio Classification and Speech Matching Techniques Based on Cluster Analysis in the Teaching of Spoken English

Fangfang Yu[1,*], Leilei Chen[1] and Jiqin Wu[1]
[1] School of International Trade, Jiangxi Tourism and Commerce Vocational College, Nanchang, Jiangxi, 330100, China
Corresponding authors: (e-mail: Themil2007@163.com).

Abstract In this paper, a scientific and effective scoring method plays an important role in the improvement of students' English speaking level in the teaching of spoken English. In this paper, we design an English speaking scoring method that integrates spectral clustering algorithm and speech data. The spectral clustering algorithm effectively integrates the feature information in the speech data by constructing a similarity matrix, and divides the students' spoken English samples into different categories. The categorized data are inputted into the scoring algorithm for speech matching, and the reference spoken English speech data are used as the standard, and the difference between the two is calculated by the dynamic time regularization algorithm, reflecting the difference between the students' spoken English speech and the reference speech, and scoring the students' spoken English performance. The spectral clustering algorithm in this paper is able to achieve a higher degree of accuracy and reduction in the classification of students' spoken English samples compared with comparative algorithms such as K-mean clustering. And based on this paper's automatic English speaking scoring algorithm, the mean difference between algorithmic scoring and manual scoring is only 0.2698 points, and there is no significant difference in the scoring level between the two. The application in English teaching can reduce teachers' workload while improving students' English speaking learning effect, which provides a more intelligent method for English speaking teaching.

Index Terms Spectral clustering algorithm, Speech data features, Similarity matrix, Dynamic time regularization, Spoken English scoring

I. Introduction

As China becomes more and more connected to the world, especially with the Beijing Olympics and the upcoming Shanghai World Expo, many people have been pushed to the top of the English learning trend [1]. Many learners do not have the time and opportunity to receive full-time systematic English training, and computer-assisted language learning (CALL), where learning time and location are not restricted, provides them with a new alternative [2]-[4]. Initially, assisted learning was mainly applied to the training of textual ability and comprehension. With the development of speech technology, more and more CALL studies and applications began to focus on English spoken pronunciation learning [5].

Most of the current computer-assisted language learning systems focus on the learning of words and grammar [6]. Only some spoken language learning software has a single function, which can only give learners an overall rating of their pronunciation. However, because of the limitations of self-learners, it is difficult for them to find errors and correct incorrect pronunciation by themselves [7], [8]. The function of software pronunciation error correction can help learners correct their pronunciation errors in time and avoid the errors from becoming habitual through many repetitions [9]. In view of the current problems of oral English teaching, oral English teaching in the network environment based on speech matching technology is to take the Internet as the platform and environmental basis of oral English teaching, use speech synthesis and speech matching technology, combine with online rich speaking resources for learning, and realize human-computer dialogue oral practice, online communication, homework correction and oral testing [10]-[14].

Aiming at the shortcomings of traditional spoken English scoring methods such as objectivity and low efficiency, this paper designs an automatic scoring algorithm for spoken English pronunciation based on the spectral clustering algorithm and the linguistic feature matching scoring method. The spectral clustering algorithm deals with complex spoken English sample data by constructing a similarity matrix, which divides the linguistic features in the data into multiple levels. Then the short-time average amplitude curve and fundamental frequency trajectory feature parameters of the data are extracted, and the feature difference degree between the audio data and the standard

data is calculated by the dynamic regularization algorithm to circumvent the non-essential differences caused by the randomness of the spoken English data. A variety of comparison algorithms are used to compare with the spectral clustering algorithm of this paper, and then based on the actual scoring effect of the algorithm, the design of the algorithm of this paper is tested for the possibility of its application in actual spoken English teaching.

II. English Spoken Pronunciation Scoring Method Based on Speech Matching Technology

Spectral clustering algorithm plays a key role in automatic scoring of spoken English pronunciation, which provides an objective basis for the scoring process by dividing the spoken English pronunciation samples into different categories through a data-driven approach.

II. A. Application of Automatic Pronunciation Scoring Algorithms in English Language Teaching

The automatic pronunciation scoring algorithm in this paper has a great application in secondary school English speaking teaching, teachers can record the students' pronunciation in class, and teachers can give assessment and feedback to the students' recordings in class, so as to find out the students' deficiencies in spoken English, and then conduct targeted speaking teaching in the next class. Students can also listen to the standardized speech several times in class, and then get the scores in real time through the comparison algorithm, which can increase the students' interest in learning English and improve the efficiency of students' oral learning. Students' independent language monitoring learning and getting real-time pronunciation feedback can effectively identify deficiencies in pronunciation, and this approach can improve learners' motivation.

Recording students' speech output in the classroom can provide a strong basis for analyzing students' output. Under the guidance of the teacher, using students' oral recordings to compare with standardized voice recordings for oral learning also reflects the teaching concept of combining classroom teaching and independent learning. Practice has shown that teaching using automatic pronunciation scoring algorithms is much more effective than traditional teaching methods in terms of listening comprehension and speaking fluency. In this way, we can change the way of learning spoken English in secondary schools by means of information technology, and the automatic pronunciation scoring algorithm based on cluster analysis not only facilitates teachers to carry out their teaching tasks, but also facilitates students to communicate with teachers on a one-to-one basis, and makes the learning efficiency more efficient.

II. B. Classification of Spoken English Audio Samples Based on Spectral Cluster Analysis

SC algorithm [15] is one of the most widely used clustering methods. Compared with the traditional K-means algorithm, SC algorithm has a better scope of application and can cluster data of arbitrary shape. It has the advantages of good clustering effect, stronger adaptability, and simple and fast realization. The basic idea is to treat all data as points in space, which are connected to each other by edges. Edges connected by more distant points have lower weights, while edges connected by closer points have higher weights. Clustering is achieved by cutting the graphs that make up the data points, with the expectation that the smaller the weight of the edges between subgraphs and the larger the weight of the edges within subgraphs after cutting.

II. B. 1) Knowledge related to spectral clustering

(1) Undirected weighted graph G

Given the similarity $s_{ij} \geq 0$ between the sample points $x_1, \cdots, x_n$ and all the data points x_i and x_j, the intuitive goal of clustering is to classify the data points into groups, so that the points in the same group are similar and the points in different groups are different. The graph $G = (V, E)$. Each vertex v_i in the graph represents a data point x_i. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is positive or greater than a certain threshold and the edges are weighted by s_{ij}. The clustering problem can now be reformulated in terms of similarity graphs: to find a division of the graph, different groups have low weights between them and edges within groups have high weights. Taking the following undirected weighted graph as an example, define $V = \{A, B, C, D, E\}$ and $E = \{e_1, e_2, e_3, e_4, e_5\}$, which corresponds to edges $e1$, $e2$, etc. in the graph, in that order.

(2) Degree matrix D

The degree matrix D is defined as a diagonal matrix with degrees $d_1, \cdots, d_n$, and D is represented as shown in equation (1). d_i denotes the sum of the weights of all vertices connected to vertex v_i. $|A|$ denotes the number

of vertices contained in A, and $vol(A)$ denotes the sum of the degrees of all vertices contained in A, as shown in the following equation:

$$D=\begin{bmatrix} d_1 & 0 & 0 & \cdots \\ 0 & d_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} \tag{1}$$

$$vol(A)=\sum_{i\in A} d_i \tag{2}$$

(3) Neighborhood matrix W

The adjacency matrix W can be obtained from the distance between vertices, W is represented as shown in equation (3).

$$W=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{3}$$

The three commonly used ways of constructing the adjacency matrix W are as follows:

$\varepsilon-$ Neighborhood graph: connects all points with distances less than ε, otherwise disconnects. Since all connected points are roughly the same distance from each other (at most ε), weighted edges only indicate connectivity. Therefore, the $\varepsilon-$ neighborhood graph is usually considered an unweighted graph, as shown in equation (4).

$$w_{ij}=\begin{cases} 0 & s_{ij}\le\varepsilon \\ \varepsilon & s_{ij}>\varepsilon \end{cases} \tag{4}$$

where ε is the threshold parameter, s_{ij} is the similarity between graph vertices, and ε-neighborhood graph only reacts to whether it is connected or not, without other information, in real life, it is rarely used.

k-nearest-neighbor graph: if v_i is in the k-nearest-neighbor of v_j or v_j is in the k-nearest-neighbor of v_i, it connects the two vertices v_i and v_j as shown in Eq. (5), which is called the k-nearest-neighbor method. If v_i is in the k-nearest-neighbor of v_j and v_j is in the k-nearest-neighbor of v_i, the two vertices v_i and v_j are connected as shown in equation (6), which is known as the mutual k-nearest-neighbor method.

$$w_{ij}=w_{ji}=\begin{cases} 0 & v_i\notin KNN(x_j) \text{ and } v_j\notin KNN(x_i) \\ \exp\left(-\dfrac{\|v_i-v_j\|^2}{2\sigma^2}\right) & v_i\in KNN(x_j) \text{ or } v_j\in KNN(x_i) \end{cases} \tag{5}$$

$$w_{ij}=w_{ji}=\begin{cases} 0 & v_i\notin KNN(x_j) \text{ and } v_j\notin KNN(x_i) \\ \exp\left(-\dfrac{\|v_i-v_j\|^2}{2\sigma^2}\right) & v_i\in KNN(x_j) \text{ and } v_j\in KNN(x_i) \end{cases} \tag{6}$$

where $KNN(x_i)$ denotes the k nearest neighbors of the data point x_i.

Fully connected graph: here simply connect all the points that have positive similarity to each other and then weight all the edges as shown in equation (7).

$$w_{ij}=w_{ji}=\exp\left(-\frac{\|v_i-v_j\|^2}{2\sigma^2}\right) \tag{7}$$

where the Gaussian similarity function $s(x_i, x_j) = \exp\left(-\left\|x_i - x_j\right\|^2 / (2\sigma^2)\right)$, with the parameter σ controlling the neighborhood's width .

(4) Laplace Matrix L

The Laplace matrix is an important tool for spectral clustering, defining the non-normalized graph Laplace matrix as:

$$L = D - W \tag{8}$$

There are two matrices known as normalized Laplace matrices. These two matrices are closely related to each other and are defined as:

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{9}$$

$$L_{rw} = D^{-1} L = I - D^{-1} W \tag{10}$$

(5) The cut-value function Cut

Given a similar graph with adjacency matrix W , the simplest and most direct way to divide the graph is to solve the minCut problem. The two most common cut functions are Ratio Cut and Ncut, where the size of a subset A of the graph is measured by the number of vertices $|A|$, and in Ncut, the size of a subset A of the graph is measured by the weights of its edges $vol(A)$. The definitions are as follows:

$$RatioCut(A_1, \cdots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \overline{A}_i)}{|A_i|} = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{|A_i|} \tag{11}$$

$$Ncut(A_1, \cdots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \overline{A}_i)}{vol(A_i)} = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{vol(A_i)} \tag{12}$$

where $W(A_i, \overline{A}_i) = \sum_{i \in A_i} \sum_{j \in \overline{A}_i} w_{ij}$, the cost function $cut(A_i, \overline{A}_i) = \frac{1}{2}\sum_{i=1}^{k} W(A_i, \overline{A}_i)$, and $\overline{A}_i$ denotes the set of nodes that do not belong to the i th class.

II. B. 2) English Spoken Audio Clustering Process

SC algorithm is an is an algorithm developed from graph theory. In general, building the similarity matrix is the most crucial basis for spectral clustering algorithms to influence the final clustering performance. The commonly used method to construct the neighbor matrix W is the fully connected approach based on Gaussian kernel distance. And to the last commonly used clustering method is K-means [16]. This paper is based on NJW algorithm to improve, the following NJW algorithm to summarize the spectral clustering algorithm process.

(1) According to equation (7), find the similarity matrix S .

(2) Find the degree matrix D . Sum the elements of each column of matrix S to get the total number, and place them on the diagonal to form a diagonal matrix, and the rest of the positions are zero to get the degree matrix.

(3) Find the Laplace matrix L . $\left(L = D^{-1/2} S D^{-1/2}\right)$.

(4) Find the K largest eigenvalues of L and the corresponding eigenvectors V .

(5) Perform K-means clustering or any other algorithm on the eigenvectors V to cluster them into K clusters.

II. C.Automatic Scoring Algorithm for English Spoken Pronunciation

The research on speech recognition based algorithms for learning spoken English is extensive. Speech recognition is the key to carry out pronunciation learning, but it can directly play a role in improving suitable English pronunciation.

II. C. 1) Scoring method based on reference speech data

Scoring based on Spoken English Reference Phonology data utilizes Spoken English Reference Phonology data as a reference standard. This means that there exists a standardized voice that serves as the best answer. The similarity between the learning voice and this voice will be used as the basis for scoring. This approach utilizes the idea of “template matching” in speech recognition. The more similar the learning voice is to the chosen reference

voice, the higher the score. After dynamic temporal regularization, time-varying speech features can be compared to reflect feature gaps. Despite the unfairness of this criterion, the method serves to motivate students to imitate the reference pronunciation.

Specifically in terms of operationalization, the main method used is feature comparison. Various features of the spoken English learning voice are compared with the corresponding features of the reference voice, the degree of difference between the learning voice and the reference voice is calculated, and the learning voice is scored accordingly.

The scoring process using the reference speech data is shown in Figure 1, which is divided into three main parts: the first part is the extraction of feature parameters, the second part is feature comparison, and the third part is the scoring mechanism.

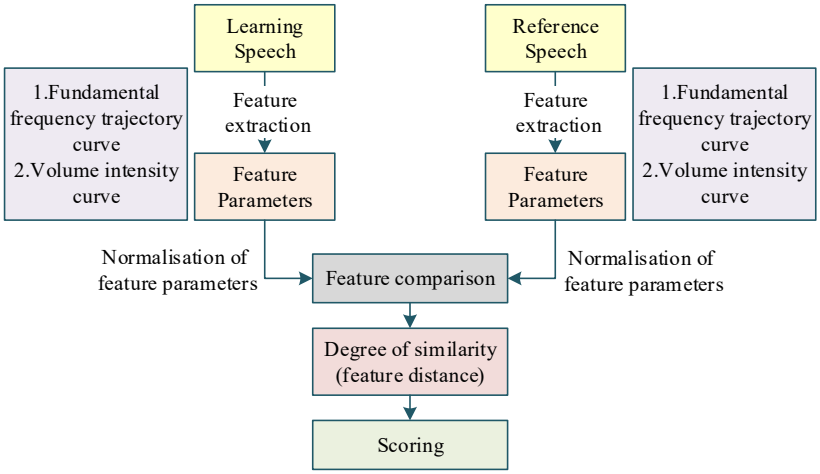


Figure 1: Feature Comparision Flow

II. C. 2) Selection of characteristic parameters

Differences in speech taking fundamental and amplitude, respectively, are used as the basis for comparison. In the part of scoring using standard speech data this paper uses the following two characteristic parameters, which are the short-time average amplitude curve and the fundamental frequency trajectory.

The fundamental period is the interval of periodic motion caused by the vibration of the vocal folds when pronouncing a turbulent sound. The fundamental period is the reciprocal of the vocal fold vibration frequency F0, which is not only an important parameter for speech signal analysis, but also an important parameter of the excitation source in the digital model of speech production, carrying very important discriminative information.

There are various methods for the detection of the fundamental tone period. Time-domain methods include autocorrelation and average amplitude difference methods [17]. Transform domain methods such as inverse spectral analysis. The latter is the most effective, but is more complex to compute.

The short-time autocorrelation function and the short-time average amplitude difference function both reflect the periodicity of the signal, but the former has a large amount of arithmetic, and the accumulation of the product has a higher requirement for the storage of the self word length. The short-time amplitude difference function has similar efficacy, so it has gained wide application in speech signal processing.

When the window function uses a rectangular window, the short-time average amplitude difference function is defined as follows:

$$F_n(k)=\frac{1}{R}\sum_{n=0}^{N-1}\left|x(n)-x(n+k)\right|\ k=0,1,\cdots N-1 \tag{13}$$

Obviously, the average magnitude difference function requires only addition, subtraction and absolute value operations and is more efficient than the autocorrelation function.

The short-time average energy is too sensitive to high levels and is prone to word length overflow. Therefore, the short-time average amplitude is used to characterize the energy of the speech signal in most cases. The short-time average amplitude is defined as follows:

$$aveMag(n)=\frac{1}{M}\sum_{m=0}^{M-1}\left|x_n(m)\right|,\ n=0,1,\cdots,N-1 \tag{14}$$

where $m=0,1,\cdots,M-1$, $n=0,1,\cdots,N-1$, N is the total number of frames, i.e. the length of the volume intensity curve, and M is the frame size.

II. C. 3) Similarity Comparison Method DTW

In speech processing, it is not possible to simply make direct comparisons between input features and templates, because speech signals have considerable randomness. Even if the same person reads the same sentence aloud, it is unlikely to have exactly the same duration, for example, as the speed of vocalization becomes faster, the duration of the stable part of the vowel will be shorter, while the duration of the consonant or transitional part of the vowel will remain essentially the same. Therefore time regularization is essential. Dynamic time regularization is a nonlinear regularization technique that combines time regularization and distance measure computation. Suppose: the sequence of reference template feature vectors is $a_1,a_2,a_3,\cdots,a_m,\cdots,a_M$, and the sequence of feature vectors of the input speech is $b_1,b_2,b_3,\cdots,b_n,\cdots,b_N$, and $M\neq N$. Then the purpose of dynamic regularization is to find a temporal regularization function $m=w(n)$ that nonlinearly maps the time axis n to the reference template time axis m such that:

$$D(n,w)=\min_{w(n)}\sum_{n=1}^{N}d\left[n,w(n)\right] \tag{15}$$

where $d[n,w(n)]$ denotes the distance between the n th and input feature vectors and the $w(n)$ th reference template vector, and it is clear that $w(n)$ should be a non-decreasing function.

Dynamic time regularization is an optimization problem, and dynamic programming techniques are commonly used to solve this problem, exploiting the concept that a local optimum can lead to an overall optimum, and the goal of the solution is to seek the optimal time regularization function $w(n)$ and the corresponding $D(n,w)$. The DTW function [18] satisfies certain constraints in the specific problem.

Boundary conditions:

$$w(1)=1,w(N)=M \tag{16}$$

Continuity conditions:

$$w(n+1)-w(n)=\begin{cases}0,1,2 & w(n)\neq w(n-1)\\ 1,2 & w(n)=w(n-1)\end{cases} \tag{17}$$

Recursive formulas can be introduced:

$$\begin{aligned}D(n+1,m)&=d[n+1,m]\\&\quad+\min[D(n,m)g(n,m),D(n,m-1),D(n,m-2)]\end{aligned} \tag{18}$$

In the formula:

$$g(n,m)=\begin{cases}1 & w(n)\neq w(n-1)\\ \infty & w(n)=w(n-1)\end{cases} \tag{19}$$

Since computing each point $D(n+1,m)$ requires computing the value of D for all three points on the n column, it is very time-consuming to compute the temporal regularization using dynamic programming techniques.

In pattern recognition, it is often necessary to calculate the distance between features. In speech recognition, the degree of similarity between the reference pattern and the input pattern is determined based on the distortion measure that constitutes the distance between the two frames. It is a measure that reflects the difference between signal features and is denoted by $D(x,y)$.

The distance measure chosen here should satisfy the following mathematical properties:

(1) Positivity: $D(x,y)\geq 0$; when $x=y$ there is $D(x,y)=0$.

(2) Symmetry: $D(x,y)=D(y,x)$.

(3) Triangular inequality: $D(x,y)\leq D(x,z)+D(z,y)$.

In the calculation of DTW distances, the absolute value average distance is used:

$$D(x,y)=\frac{\sum_{i=1}^{N}|x_i-y_i|}{N} \tag{20}$$

II. C. 4) Feature comparison scoring and scoring parameter tuning

DTW distance is not directly used as a score for spoken English pronunciation; a reasonable mapping that will go from distance to score must be found. It is assumed that the relationship between distance and score satisfies the following equation:

$$score=\frac{100}{1+a(dist)^b} \tag{21}$$

Clearly this formula maps distances to a range of fractions from 0 to 100. Solving for the unknown parameters a and b in the formula requires knowing some score and distance pairs. It is possible to solve for the above parameters from the score values and DTW distances of some experts in the experiment. Using the formulas of this thesis even if the distances are larger or smaller than what appeared in the experiment it is reasonable to convert the scores into the interval from 100 to 0.

The actual score estimation formula is slightly more complex due to the fact that 2 feature parameters are used, and the final score is a weighted sum of the two.

$$score=w_1\cdot\frac{100}{1+a_1(dist_1)^{b_1}}+w_2\cdot\frac{100}{1+a_2(dist_2)^{b_2}} \tag{22}$$

The unknown parameters in Eq. satisfy certain constraints: $a_1,a_2,b_1,b_2>0$ and $w_1+w_2=1$. The a_1,a_2,b_1,b_2 are the parameters of the distances converted to scores, and w_1,w_2 are the weights of the three features.

III. Results and Analysis of Oral Pronunciation Scoring in English Language Teaching

III. A. Subjects

The test subjects for this experiment were selected from a randomly selected class of students in the first grade of a school, the total number of students in the class was 50, of which the ratio of male to female was similar, from which 15 boys and 15 girls were randomly selected as test subjects. In order to test the accuracy of the algorithm's classification, 500 segments of English listening audio from the English listening test of the grade were selected to be used as a comparison of the algorithm's classification results. In order to test the accuracy of the algorithm's scoring, eight masters of English majors of the school were invited to manually score the test students' speaking in the testing process. The test words were words commonly used in the first grade, which were divided into 3 categories, monosyllabic, disyllabic and polysyllabic words, and the number of words was 50, and the test words for each student were randomly selected during the test. The accuracy of the algorithmic scoring was determined by the difference in similarity between the manual scoring and the algorithmic scoring.

III. B. Experimental steps

This paper is through the standard voice acquisition, play the standard voice, student recordings and other experimental steps process to achieve voice comparison, experiments with the standard voice of the word for the KDDI, the synthesized voice relative to the standard voice file, the test object is 30 students, first of all, collect 500 English listening audio used as a classification of English teaching spoken audio, 500 audio is divided into three categories. Then, students imitated the standard speech, and eight master's degree students of English majored in the university scored the students' imitated speech, and finally calculated the average score of the eight artificial scores, and the similarity scoring mechanism was 100 points. Secondly, the algorithm comparison software is used for automatic scoring, and finally the gap between the algorithm's scoring and the manual scoring score is obtained, so that a better comparison can be made as to whether the scoring mechanism of the pronunciation auto-scoring algorithm based on cluster analysis is in line with the subjective feeling of human beings. This voice comparison is carried out on android platform.

III. C. Comparison of Audio Classification Results of Spoken English Teaching

In this paper, we compare the results obtained by K-mean clustering, clustering algorithm for fast finding density peaks, and spectral clustering algorithm in this paper.The results obtained by the three clustering methods are compared with the type distribution map of the original English listening audio data, where the original audio type is

the audio type of English exam listening in a certain school. For example, if a certain English listening audio is named as Daily Conversation category, all English listening of that audio is set as Daily Conversation English exam listening. Finally, each point is colored and drawn on a two-dimensional plane, which is the type distribution map. Figure 2 shows the type distribution map of 500 English listening audios. The hollow stars are the daily conversation English listening audio. The solid stars are English listening audio for entertainment and life. The squares indicate the more serious and regular academic English listening audio. Because the characteristics of daily conversation and entertainment audio are more consistent and the language is more emotional, they are grouped into one category, so that the original audio data can be viewed as containing 2 clusters (k=2). The results obtained by the three clustering algorithms are shown in Fig. 3 to Fig. 5, Fig. 4 (a), (b) for the clustering algorithm for fast finding density peaks for the decision diagram and 2d non-classical multimodal degree scaling diagram, respectively.

As can be seen from the figure, compared with other clustering methods, the clustering results obtained by the spectral clustering algorithm are closer to the actual distribution of English listening audio data. Specifically, the comparison of the clustering results of the three clustering algorithms with the real results can be seen that the category differentiation of the Fast Find Peak Density algorithm is more accurate, but the exclusion of noise is too much, and the K-mean clustering does not eliminate noise, but the clustering effect is less satisfactory. The clustering effect of spectral clustering is the most ideal among the three, clustering the 500 English listening audio segments into audio containing 2 clusters in a more restored way.

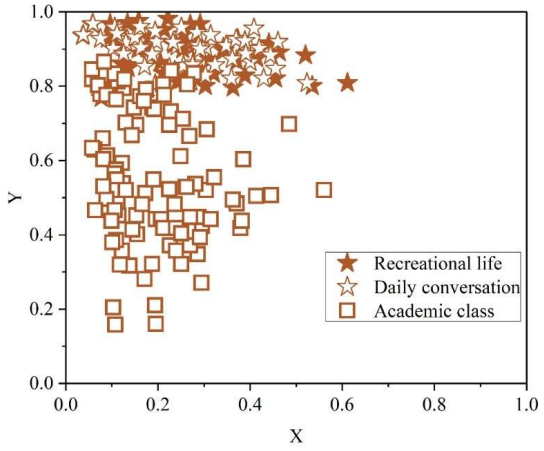


Figure 2: Original sample distribution

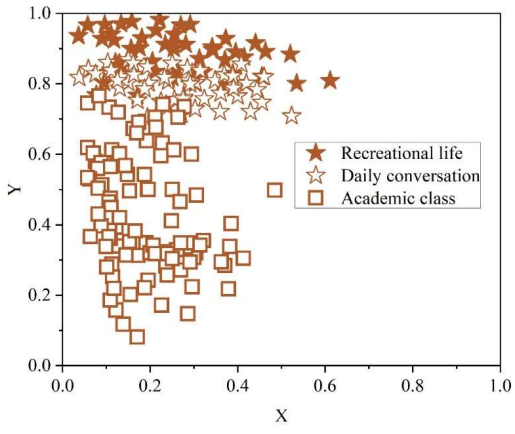
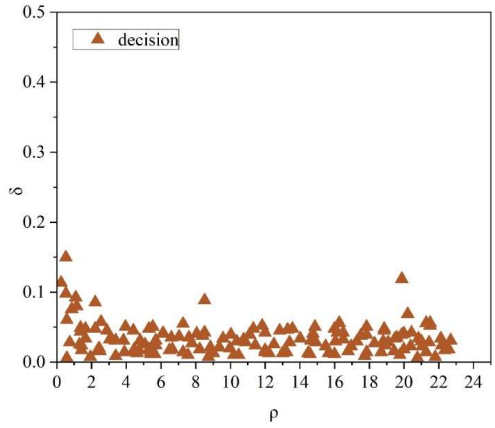
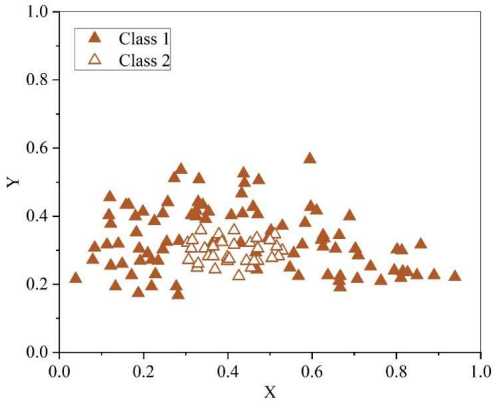


Figure 3: K-mean clustering results



(a)Decision graph



(b)2D nonclassical multidmensional scaling

Figure 4: Fast finding clustering algorithm of density peak

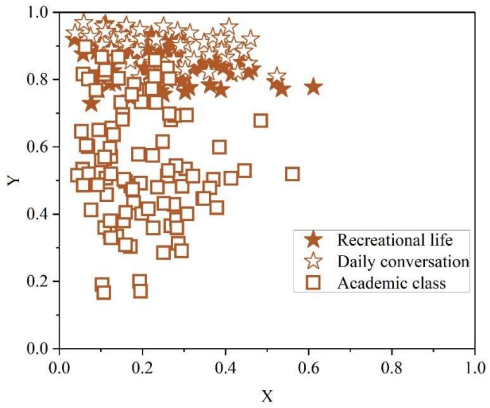


Figure 5: The spectral clustering algorithm of this paper is obtained

III. D. Results and Analysis of Speaking Scores

The specific experimental data are shown in Table 1. Table 1 shows some of the data of the tests conducted on 30 students (male and female students) on monosyllabic, disyllabic, and polysyllabic words of first-year English words respectively. Stu1-Stu3 refers to the data of 3 randomly selected students tested on monosyllabic words, with words randomly selected from a list of 30 monosyllabic words.Stu4-Stu6 refers to the data of 3 randomly selected students tested on disyllabic words, with imitation words for each student also randomly selected from a list of 30 disyllabic words.Stu7, Stu8 refers to 2 randomly selected students tested on multisyllabic words, and each student's imitation words were also randomly selected from a list of 30 multisyllabic words. The manual scoring represents the average score obtained by the 8 English majors who scored the students' imitation speech, and the algorithmic scoring represents the algorithmically derived similarity scores after the students performed the imitations.

In order to analyze whether there is a difference between algorithmic scoring and manual scoring, independent samples t-test was conducted with scoring forms (algorithmic and manual) as factors. Firstly, the test hypothesis was designed and the test level was clarified. h1: $\mu 1 = \mu 2$, i.e., there is no difference in the level of students' scores by different scoring forms. h2: $\mu 1$ and $\mu 2$ are unequal, i.e., there is a difference in the level of students' scores by different scoring forms, with $\alpha=0.05$. The group statistics and the results of the independent samples test are shown in Tables 2 and 3, respectively. t denotes the data of manual scoring, and G denotes the data of algorithmic scoring.

The mean value of sample T is 83.5612, which is basically equal to the mean value of sample G 83.2914. for both, F corresponds to Sig equal to 0.258 greater than 0.05, so the variance is chi-square, corresponding to the two-sided test P = 0.785 is greater than 0.05, which indicates that the scores of samples T and sample G do not have a significant difference at the level of 0.05, so the test hypothesis H2 is rejected, i.e., there is no difference at the level of the scores of samples T and Sample G scores are not different.

After analyzing the data of manual scoring and algorithmic scoring above, it can be seen that no matter whether the test is conducted on monosyllabic, bisyllabic as well as polysyllabic words, the similarity results of algorithmic scoring and the scoring of the school's master's degree in English (manually) do not differ much, which is basically in line with the human's subjective feeling. Overall, after the corresponding test, the algorithm is basically qualified to meet the requirements of secondary school English word voice comparison, which is conducive to students' better learning of English word pronunciation in the classroom, and better than the traditional just listen to the English voice without giving real-time feedback on the students' oral learning effect. There is an error between the scoring of the speech evaluation algorithm and the manual scoring, but the error is not big, on the one hand, there is a subjective error in the manual scoring itself, on the other hand, the algorithm needs to be further optimized and improved. In short, this algorithm is basically completed in the expectation.

Table 1: Experimental results

Student	Word number	A	B	C	D	E	F	G	H
Stu1	Algorithm rating	81.2	73.4	78.8	78.9	79.8	81.3	83.3	86.2
	Manual grading	80.1	72.3	79.2	78.3	80.2	82.7	83.6	85.9
Stu2	Algorithm rating	84.7	82.1	86.4	91.9	90.5	84.6	91.4	82.8
	Manual grading	86.1	82.5	85.9	91.2	90.7	84	93.1	82.9
Stu3	Algorithm rating	94.2	93.6	85.3	87.3	86.3	86.5	84.8	86.8
	Manual grading	95.3	94.1	85.6	86.7	87.1	86.3	85.2	87.2
Student	Word number	a	b	c	d	e	f	g	h
Stu4	Algorithm rating	80.9	72.8	79.2	78.6	79.4	82.2	83.2	85.5
	Manual grading	79.9	76.7	75.4	80.8	82.7	80.1	81.1	82.1
Stu5	Algorithm rating	85.8	82.2	86.3	91.2	90.5	84	92.1	83.2
	Manual grading	84.2	83.9	86.3	88.2	92.4	85	90.5	82.8
Stu6	Algorithm rating	95	93.6	85.1	87.5	87	86.9	85	86.9
	Manual grading	95.1	92.8	84.3	88	90.5	91	83.4	85.5
Student	Word number	x1	x2	x3	x4	x5	x6	x7	x8
Stu7	Algorithm rating	81.5	73.4	79.6	78.8	79.2	82.3	83.1	85
	Manual grading	79.7	75.9	74.9	80.4	82.4	80.8	80.1	82.1
Stu8	Algorithm rating	84.4	82.2	86.2	92.3	90.3	84.5	91.6	82.8
	Manual grading	86.5	82.4	86.6	90.8	90.9	84.8	93.7	82.8

Table 2: Group statistic

Sample		N	Mean	Standard deviation	Standard error of mean
Score	T	30	83.5612	8.4215	0.9546
	G	30	83.2914	8.0564	0.8954

Table 3: Independent sample inspection

		Levin variance equivalence test		T test						
		F	Sig.	t	df	Sig.2	MD	SE	95%CI Upper limit	95%CI Lower limit
Score	Equal.Var.	1.235	0.258	0.265	59	0.785	0.33541	1.28546	-2.45121	2.84512
	Unequal.Var.			0.265	58.145	0.785	0.33541	1.28546	-2.45121	2.84512

IV. Conclusion

This paper organically combines the spectral clustering algorithm and the feature comparison scoring method based on reference speech, clusters pronunciation samples of different levels into different categories, and accurately scores the samples of different feature categories.

The article selected K-mean clustering, fast finding density peak clustering algorithm as the comparison algorithm, based on the school's first grade 500 English listening audio for comparison experiments, to verify the effectiveness of this paper based on spectral clustering algorithm of English listening samples clustering. The algorithm in this

paper asks the basis of excessive denoising, 500 segments of English listening audio is accurately clustered into audio containing 2 clusters, the application of the spectral clustering algorithm to English language teaching, a large number of samples used for scoring of unequal levels of accurate classification, can largely improve the scoring accuracy of the subsequent automatic scoring algorithms for English oral pronunciation, and make a good foundation for the enhancement of the level of students' English performance.

In the oral scoring experiment, this paper's automatic pronunciation scoring algorithm based on cluster analysis, algorithmic scoring and manual scoring mean difference of only 0.2698 points, independent samples test results show that the algorithmic scoring and manual scoring there is no significant difference, which shows that this paper's algorithm basically realizes the secondary school monosyllabic, disyllabic and multi-syllabic words in a variety of different classes of English words oral-speech comparison. The algorithm in this paper basically realizes the comparison of spoken English words of different classes in secondary schools. Secondly, in terms of application, the automatic scoring algorithm of pronunciation based on cluster analysis in this paper can better realize the promotion of English teaching and learning, which can help teachers and students to a certain extent in their teaching and learning tasks, and can promote students' learning of spoken English.

Funding

The Research on the Reform of Online Practical Teaching for Business English Major in Higher Vocational Colleges under the Guidance of "Reality and Virtuality Complementary, Creator Education" - Taking "Business English" Course as an Example.

Application of "EPMC" Model in the Cultivation of Business English Talents in the Context of Cross - Border E-commerce.

References

[1] Cui, T., & Yang, Y. (2022). Social relationships and grit in English as a foreign language learning among high school students: A three-wave longitudinal study. Frontiers in Psychology, 13, 1038878.

[2] Mutlu, A., & Eroz-Tuga, B. (2013). The role of computer-assisted language learning (CALL) in promoting learner autonomy. Eurasian Journal of Educational Research, 51, 107-122.

[3] Chapelle, C. A. (2010). The spread of computer-assisted language learning. Language teaching, 43(1), 66-74.

[4] Enayati, F., & Gilakjani, A. P. (2020). The Impact of Computer Assisted Language Learning (CALL) on Improving Intermediate EFL Learners' Vocabulary Learning. International Journal of Language Education, 4(1), 96-112.

[5] Rasekh Eslami, Z., & Zohoor, S. (2023). Second language (L2) pragmatics and computer assisted language learning (CALL). Technology Assisted Language Education, 1(3), 1-17.

[6] Tafazoli, D., María, E. G., & Abril, C. A. H. (2019). Intelligent language tutoring system: Integrating intelligent computer-assisted language learning into language education. International Journal of Information and Communication Technology Education (IJICTE), 15(3), 60-74.

[7] Shak, P., Lee, C. S., & Stephen, J. (2016). Pronunciation problems: A case study on English pronunciation errors of low proficient students. International Journal of Language Education and Applied Linguistics.

[8] Caisaguano Tigasi, G. A. (2024). English pronunciation errors in english major students (Doctoral dissertation, Ecuador: Pujilí: Universidad Técnica de Cotopaxi (UTC)).

[9] Ramasari, M. (2017). Students pronounciation error made in speaking for general communication. Linguistic, English Education and Art Journal, 1(1), 37-48.

[10] Jiao, F., Song, J., Zhao, X., Zhao, P., & Wang, R. (2021). A spoken English teaching system based on speech recognition and machine learning. International Journal of Emerging Technologies in Learning (iJET), 16(14), 68-82.

[11] Wang, J. (2020). Speech recognition of oral English teaching based on deep belief network. International Journal of Emerging Technologies in Learning (Online), 15(10), 100.

[12] Zhang, Y., & Liu, L. (2018). Using computer speech recognition technology to evaluate spoken English. Educational Sciences: Theory & Practice, 18(5).

[13] Li, M. (2021, September). The application of computer speech recognition technology in oral English teaching. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 1271-1274).

[14] Liu, L., & Zhou, R. (2020). Optimization of oral English teaching system based on computer-aided technology. Computer-Aided Design and Applications, 18, 147-157.

[15] Xin Chen,Naiwei Kuai,Wenwei Fu,Zhiqiang Zhang,Tong Guo,Tao Liu & Cong Liu. (2025). Automated physics parameter identification of tuned vibration absorber in offshore wind turbines based on unsupervised spectral clustering and SSI. Ocean Engineering,328,121052-121052.

[16] Duygu Selin Turan & Burak Ordin. (2025). The incremental SMOTE: A new approach based on the incremental k-means algorithm for solving imbalanced data set problem. Information Sciences,711,122103-122103.

[17] So Shinae,Lee Kang Hee,You Kwang Bock,Lim Ha Young & Park Jisu. (2017). A Study of the Pitch Estimation Algorithms of Speech Signal by Using Average Magnitude Difference Function (AMDF). Asia-pacific Journal of Multimedia services convergent with Art, Humanities, and Sociology,7(4),235-242.

[18] Zhonghai Chen & Tengyu Zhang. (2025). Evaluation of basic sports actions for students based on DTW posture matching algorithm. Systems and Soft Computing,7,200196-200196.