Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Design and implementation of a personalized vocal music teaching system assisted by artificial intelligence algorithms

Jialu Qin[1,] ✉

[1] *College of Education, Hubei Business College, Wuhan, Hubei, 430079, China*

ABSTRACT

The rapid development of artificial intelligence technology has made its application in the field of education increasingly widespread. The purpose of this paper is to design and implement a personalized vocal music teaching system based on artificial intelligence algorithms to solve the problems of single teaching method and lack of personalized guidance that exist in traditional vocal music teaching. The overall architecture of the system is constructed by analyzing the demand for vocal music teaching and combining deep learning and other artificial intelligence technologies. The key algorithms involved in the system are elaborated in detail, including the personalized recommendation algorithm of the learning path fused with the long and short-term memory network (LSTM) and the attention mechanism, and the intelligent evaluation algorithm that includes the evaluation of pitch, rhythm and timbre. Through practical application cases, it is verified that the system in this paper can effectively improve the teaching effect of vocal music and students' vocal music professionalism, providing an important auxiliary role and key ideas for the innovative development of vocal music teaching.

Keywords: artificial intelligence algorithm, deep learning, LSTM, attention mechanism, vocal music teaching

1. Introduction

The rapid development of information technology has brought profound changes to contemporary society, and vocal music teaching is no exception. With the support of information technology, the traditional mode of vocal music teaching is being subverted and reconstructed. The development of technology has had a great impact on the transformation of the traditional concept of music education. Nowadays, computer technology has a close connection with music education, and the application of artificial intelligence in vocal music teaching is a very distinctive example. Whether in terms of

✉ Corresponding author.
E-mail address: qinjialu2025@163.com (J. Qin).
Received 10 January 2024; Revised 25 May 2024; Accepted 20 October 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-182

teaching ideas, teaching methods, or in terms of teaching design, the personalized teaching system assisted by artificial intelligence algorithms greatly improves the quality of vocal teaching [1-2].

Education is not only to impart knowledge to students, but also to cultivate students to form good habits of thinking and learning habits, it can be said that the scientific form of artificial intelligence has a high degree of relevance to education in nature [3]. The purpose of artificial intelligence applied to vocal music teaching is not only to give full play to the role of intelligent tools, but also to provide teachers with some new ideas to solve teaching problems [4-6]. Under the general framework of music education, intelligent tools can collect sound information appearing in the teaching process and analyze it, and then combine with neural network algorithms to generate results that meet teaching needs, which can provide a new path for students to feel and study music [7-10]. From a practical point of view, artificial intelligence is very suitable for vocal music teaching, which is mainly reflected in two aspects. That is, the artificial intelligence-based feedback system can effectively improve students' practice efficiency, in addition to the digital processing of sound information is also convenient for teachers to carry out vocal research, so as to develop a more scientific and rational and personalized teaching program [11-13].

In this paper, the main framework of the vocal music teaching aid system is constructed based on the artificial intelligence technology and the demand analysis of vocal music teaching. The long and short-term memory network is used to capture the time-dependence of vocal learners' behavioral sequences, and the attention mechanism is integrated to dynamically adjust the algorithm's attention to different time steps, so as to construct a personalized learning path recommendation model, which is used for personalized recommendation of vocal learning tasks to students. Intelligent evaluation of students' singing practice is realized by calculating the physical modeling of students' singing pitch, extracting rhythmic features, and classifying and identifying timbre, and comparing them with standard pitch, rhythm and timbre. At the same time, the evaluation results of pitch, rhythm and timbre are integrated, and the weighted sum is used to get the comprehensive score of the student's singing. The system in this paper is applied to actual vocal music teaching, and through the analysis of personalized learning path recommendation, intelligent assessment and application effect, the system constructed in this paper is demonstrated to be an effective aid in vocal music teaching, providing technical support for the intelligent reform of vocal music teaching.

2. System design

2.1. Needs analysis and modeling of course structure

The first step in constructing a vocal music teaching aid system based on artificial intelligence technology is to analyze the teaching needs and the characteristics of the curriculum system in depth. This involves a detailed study of the syllabus, lesson plans and other related documents, supplemented by in-depth interviews with teachers and questionnaire surveys of students, in order to gain an all-round insight into the distribution pattern of the knowledge points of the professional vocal music courses, the conditions of the prerequisite courses, the requirements of the practical operation, and many other details. On this basis, the ontology modeling technique is applied to express the knowledge of the course, and an ontology model structure consisting of concepts, attributes, examples and other basic units is constructed. At the same time, the complex network theory is used to reveal and analyze the intrinsic connection and knowledge dependency between courses, and then a directional acyclic graph (DAG) course structure model with directional and acyclic characteristics is established. In this DAG model, each node represents a course, and the edges symbolize the logical dependencies between courses. Through the quantitative calculation of network metrics such as degree centrality and meso-centrality of nodes, the core courses of the course network and their key delivery paths can be precisely located, providing strong data base support for subsequent intelligent teaching services.

2.2. Intelligent Learning Path Recommendation

In the educational environment of vocal music majors, reasonable planning and optimization of learning paths play a decisive role in improving the overall learning effectiveness. To this end, a set of teaching assistance system equipped with artificial intelligence technology should fully utilize the advantages of the ontological framework of course knowledge and complex network model to develop an intelligent learning path recommendation function. Specifically, it can combine the long and short-term memory network (LSTM) and the attention mechanism to generate a personalized list of students' learning path recommendations while improving the recommendation accuracy by using the learners' task sequences and behavioral characteristics, capturing the time-dependence through the LSTM, and focusing the attention mechanism on the key tasks.

2.3. Intelligent assessment and feedback systems

Intelligent assessment and feedback is one of the core functions of the artificial intelligence-based teaching aid system, which is of great significance in promoting students' independent learning and optimizing teaching decisions. The system adopts diversified assessment methods, taking into account the multiple dimensions of students' singing, such as pitch, rhythm and timbre, to build a comprehensive and objective evaluation system of singing ability. In terms of pitch, through physical modeling of vocal pitch and spectral analysis, the system tests the frequency of students' vocals at different pitches and compares it with the standard frequency to assess the pitch of students' singing. In terms of rhythm, the system extracts the rhythmic features of the audio sung by the students, such as beat and rhythmic pattern, and compares them with the standard rhythm to judge the accuracy of the rhythm sung by the students. In terms of timbre evaluation, this paper proposes a timbre classification recognition method of harmonic feature extraction combined with support vector machine (SVM), which uses discrete harmonic transform to extract timbre harmonics and sequentially constructs timbre expression spectra. Then the linear prediction cepstrum coefficient (LPCC), Mel frequency cepstrum coefficient (MFCC) and other features are integrated as the input of SVM, and the classification and recognition of students' singing timbre is finally realized. The recognition results are compared with the standard timbre samples to give the timbre evaluation.

3. Key algorithm design

3.1. Intelligent Learning Path Recommendation Model

In order to improve the accuracy of personalized learning path recommendation, the study proposes a learning path recommendation model that combines LSTM and attention mechanism [14]. The model recommends personalized learning task sequences for learners by capturing their behavioral sequences and dynamically identifying key learning behaviors.

The EdNet dataset, a large-scale dataset widely used in vocal education and personalized learning research, contains learning behavior records from a large number of vocal learners. Using these behavioral data, the study constructs time-series models in order to capture learners' behavioral patterns and dynamically weight the importance of each task through an attentional mechanism in order to recommend the most appropriate subsequent learning tasks.

3.1.1. Model architecture. The model architecture consists of two main components, namely the LSTM network for capturing the temporal dependence of the learner's behavior and the attention mechanism for dynamically adjusting the model's attention to different time steps. The overall architecture of the model is shown in Figure 1. Among them, the input layer input is the learner's task sequence. The LSTM layer is used to process the input sequence to generate hidden states. The attention mechanism layer weights the hidden states output from the LSTM and calculates the attention weights. The output layer outputs the final learning task recommendation.

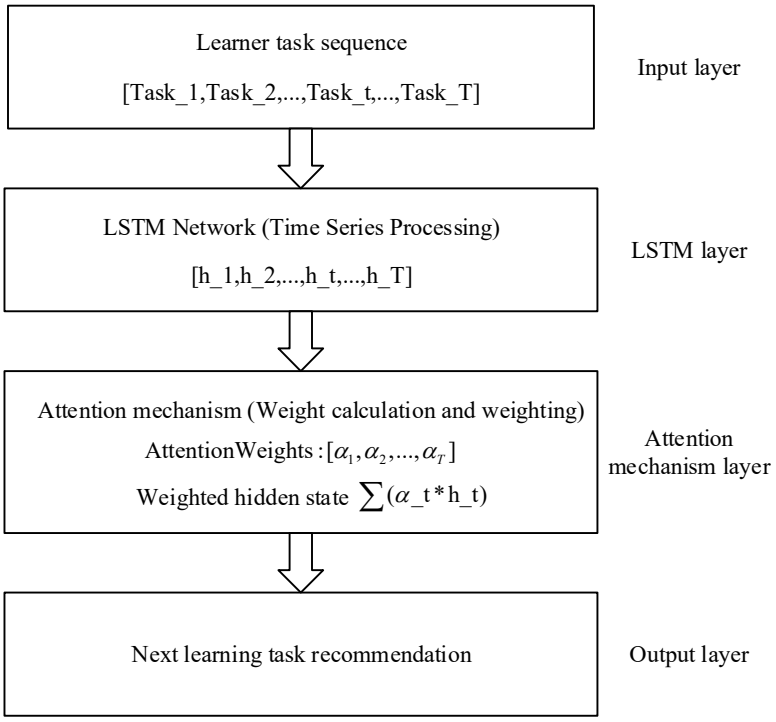


Fig. 1. Overall architecture of the model

3.1.2. Integration of LSTM and Attention Mechanisms. LSTM is a typical model for processing time-series data, which can effectively capture learners' learning behaviors in the time dimension.The input of LSTM is a sequence of learning tasks completed by learners in a chronological order, and the completion of each task (e.g., time of completion, whether it is correct or not) serves as an input to the timesteps.The input sequence is processed by LSTM through Equation (1) and Equation (2):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1}$$

where f_t is the output of the oblivion gate. x_t is the input task at time t. h_{t-1} is the hidden state of the previous time step. W_f is the weight matrix. σ is the Sigmoid function.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2}$$

where C_t is the current memory state that captures the learner behavior pattern at the current moment.

Although LSTM is able to capture the temporal dependence of learner behavior, not every task's behavior has the same impact on the recommendation path. The attention mechanism improves recommendation accuracy by assigning weights to each hidden state, enabling the model to focus on those learning tasks that are most important to the learning path.

First, based on the hidden states h_t output from the LSTM, the score e_t for each time step is computed through the attention layer, as shown in Equation (3):

$$e_t = \tanh(W_e \cdot h_t) \tag{3}$$

where W_e is the weight matrix of the attention layer. e_t is the score for each task.

Then, the scores e_t for all time steps are normalized by Eq. (4) to obtain the attention weights α_t, which reflect the importance of each task for the current learning path recommendation:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)} \tag{4}$$

Finally, the hidden states h_t at each time step are weighted according to the attention weights to obtain the final context vector c_i, which is used to generate personalized learning path recommendations as shown in Equation (5):

$$c_t = \sum_{t=1}^{T} \alpha_t \cdot h_t \tag{5}$$

3.1.3. Model output. In the learning path recommendation task based on the EdNet dataset, the output of the model is the recommendation of learning tasks. Specifically, the model recommends a sequence of subsequent learning tasks for each learner based on the learner's history of task completion. Each learning task may be a singing exercise, a knowledge point module, or a specific test that helps the learner select the appropriate content for subsequent learning. The recommended sequence of these tasks is based on the time dependency of the learner's past tasks and the weighting adjustment for task completion.

3.2. Intelligent assessment

3.2.1. Sound level assessment. The fundamental frequency in the glottal wave determines pitch, so physical modeling of pitch also means exploring the vibrational processes of the vocal folds. The vocal folds are symmetrical structures composed of the vocal ligament, the vocal fold muscle, and the arytenoid muscle. The dual piezoelectric oscillator model of the vocal folds is used, as shown in Fig. 2(a), where one side of the vocal folds is regarded as a dual piezoelectric oscillator composed of two identical rectangular thin-plate piezoelectric wafers bonded together with the length of piezoelectric wafer being L, the width of the piezoelectric wafer being W, and the thickness of the piezoelectric wafer being H, and the polarity of the bonded side being reversed. The dual piezoelectric oscillator is connected to the circuit shown in Fig. 2(b) to simulate a sound band, and the direction of the electric field in the piezoelectric sheet is shown in Fig. 2(c). According to the inverse piezoelectric effect, the left piezoelectric sheet has the same polarity direction as the direction of the external electric field and is stretched in the x-axis direction, and the thickness H increases, and the area decreases, i.e., both the length L and the width W decrease, since the total volume of the piezoelectric sheet remains unchanged. The right piezoelectric sheet has opposite polarity to the external electric field and is compressed in the x-axis direction, the thickness H decreases and the area increases, i.e., the length L and width W both increase. The area of the bonding layer remains unchanged, the area of the left side decreases and the area of the right side increases, so that the double piezoelectric sheet bends positively toward the x-axis (Fig. 2(d)). The same double-pressure oscillator is then used to simulate the left vocal folds, with the upper and lower ends connected, and access to the circuit showing symmetric bending (Fig. 2(e)). When the external electric field is withdrawn, the piezoelectric sheet will vibrate along the x-axis of bending, which is similar to the vibration of the vocal cords.

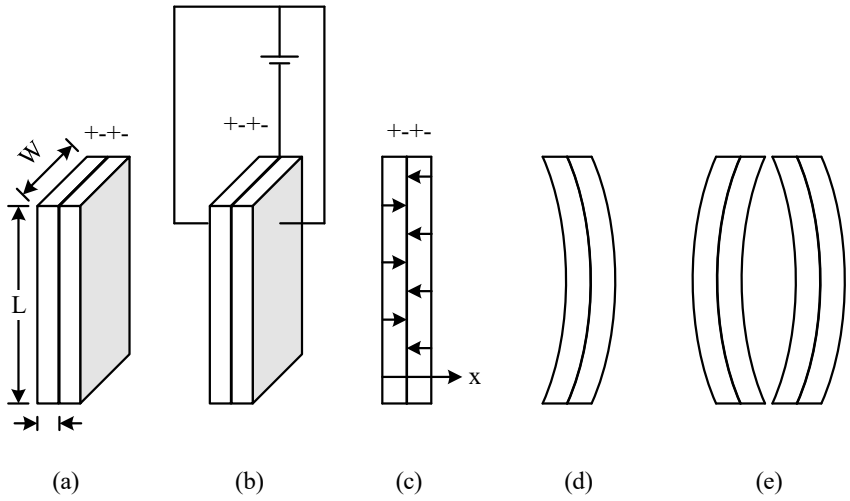


Fig. 2. Double pressure oscillator model of vocal cord

From the knowledge of elastic mechanics, the bending vibration equation of the rectangular cross section of the double compression vibrator is:

$$\frac{\partial^2 \eta}{\partial t^2} - r^2 \frac{\partial^4 \eta}{\partial x^2 \partial t^2} + \frac{Er^2}{\rho} \frac{\partial^4 \eta}{\partial x^4} = 0 \tag{6}$$

where η is the vibrational displacement, ρ is the density of the oscillator, r is the radius of gyration, and E is the Young's modulus of the oscillator Solving Eq. (6), the intrinsic frequency of the bending vibration of the piezoelectric oscillator with rectangular cross-section can be obtained as:

$$f = \frac{H}{\pi} \sqrt{\frac{E}{3\rho(1-\gamma^2)} \left[\frac{P^4(i)}{L^4} + \frac{Q^4(j)}{W^4} \right]} \tag{7}$$

where γ is the Poisson's coefficient of the material and i and j are the vibration orders:

$$P(i) = \frac{2i+1}{4}\pi, Q(j) = \frac{2j+1}{4}\pi \tag{8}$$

When $i = j = 1$, the fundamental frequency is obtained:

$$f_\circ = \frac{9\pi H}{16} \sqrt{\frac{E}{3\rho(1-\gamma^2)} \left(\frac{1}{L^4} + \frac{1}{W^4} \right)} \tag{9}$$

In Eq. (9), E/ρ is the specific modulus of the material, which reflects the load carrying capacity of the material, and the larger the specific modulus, the more rigid the material is. From Eq. (9), the main factors determining the fundamental frequency of the vocal folds are the vocal fold dimensions (length L and width W) and the performance parameter (specific modulus E/ρ). The dimensions of the vocal folds are different for different people, e.g., the length of the soprano vocal folds is 11-19 mm, the length of the mezzo-soprano vocal folds is 18-21 mm, the length of the tenor vocal folds is 14-22 mm, the length of the baritone vocal folds is 22-24 mm, and the length of the baritone vocal folds is 24-25 mm. it is not difficult to see that the vocal folds of males are longer than those of females, and their fundamental frequency is lower. The specific modulus of the vocal folds is determined by the degree of tension in the vocal muscles and ligaments, and under tension the specific modulus is greater and the fundamental frequency of the vocal folds is higher. Vocally trained people can adjust the specific modulus of the vocal folds over a wider range, resulting in a wider range of fundamental frequencies, and thus a wider range of pitches to achieve intonation.

3.2.2. Rhythm assessment. The extraction of singing rhythm features can be understood as the extraction of beat tempo related features in BPM. The autocorrelation function can be obtained by determining the singing audio articulation points, from which the beat period is calculated, and the BPM value is calculated based on the beat period.

1) Note onset detection. The phase variation of the spectrum can be applied to note onset detection. However, using only the phase change to detect the audio onset can be problematic because noise can cause some phase distortion. Based on the above, note onset detection can be effectively performed using a combination of energy and phase, calculated as in equation (10):

$$\tilde{S}_k(m) = \tilde{R}_k(m)e^{j\Phi_k(m)} \tag{10}$$

where, m is the number of the frame, $\tilde{R}_k(m)$ is the amplitude of the previous frame, and $\tilde{\Phi}_k(m)$ is obtained by the phase difference between the previous frame and the one before it, calculated as:

$$\tilde{\Phi}_k(m) = \text{princ arg}[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)] \tag{11}$$

where princ arg is mapping the obtained phase values to the $[-\pi, \pi]$ interval. The actual value of the k rd frequency band is calculated as:

$$(m) = R_k(m)e^{j\varphi_k(m)} \tag{12}$$

where $R_k(m)$ and $\varphi_k(m)$ are the amplitude and phase, respectively, obtained from the current frame after a short-time Fourier transform. The features of each frame are computed as:

$$\Gamma(m) = \sum_{k=1}^{k-K} |S_k(m) - \tilde{S}_k(m)|^2 \tag{13}$$

By calculating the features of all frames of the audio and normalizing them, the note onset detection signal is obtained.

2) Beat Cycle Estimation. The first step of beat period estimation is to compute the autocorrelation function for the start point detection function of each frame. In order to make the autocorrelation function clearer, it is necessary to preprocess the data of each frame by first setting an adaptive moving average threshold, which is calculated as follows:

$$\bar{\Gamma}(m) = \sum_{q=m=0}^{q-m+\theta} \Gamma_i(q) \tag{14}$$

The sliding window size is set to 16 points. Each point of the probe function is then allowed to subtract the corresponding threshold and half-wave rectify the output, calculated as in equation (15):

$$\tilde{\Gamma}_i(m) = \frac{(\Gamma_i(m) - \bar{\Gamma}_i(m)) + |\Gamma_i(m) - \bar{\Gamma}_i(m)|}{2} \tag{15}$$

Finally the autocorrelation function is calculated for the preprocessed signal as in equation (16):

$$A(i) = \frac{\sum_{m=1}^{N} \tilde{\Gamma}_i(m)\tilde{\Gamma}_i(m-i)}{|i-N|} \tag{16}$$

where $i = 1, 2, \cdots, N$ denotes the number of points on the frame and N denotes the frame length. The point τ in the autocorrelation domain can be mapped to the beat rate with the mapping relationship as in equation (17):

$$\mathrm{BPM} = \frac{60}{\tau_i * 0.0116} \tag{17}$$

After obtaining the autocorrelation function for each frame, it needs to be weighted. This is because if the same weight is given to each point representing a beat period, it will give too many degrees of freedom to the possible beat periods and thus lead to inconsistency in the output results. The weighting is done by using a function based on the Rayleigh distribution, calculated as follows:

$$L_w[i] = \frac{i}{b^2} e^{\frac{-i2}{2^2}} \tag{18}$$

where i denotes each point corresponding to a value in the range [1,128]. The value of b is 50, which indicates that the weights are maximized at the 50th point, and according to Eq. (17), the BPM of this point is about 103.

In order to calculate that the beat rate t_b of the current frame can depend on the beat rate t_{b-1} estimated from the previous frame, a Hidden Markov Model is constructed in this paper. Each column of the state transfer matrix of this model consists of a Gaussian distribution with standard deviation 8. The state transfer matrix is calculated as in equation (19):

$$A(t_i, t_j) = P(t_b = t_j \mid t_{b-1} = t_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(t_i - t_j)^2}{2\sigma^2}\right) \tag{19}$$

The initial probability distribution is a Rayleigh distribution, where σ is fixed to 8 and t_i, t_j ranges from 0 to 127. The state probability vector of the current frame is calculated by multiplying the state probability vector of the previous frame by the corresponding vector of the corresponding state transfer matrix and taking the maximum value. The calculation method is:

$$\Delta_b(t_j) = \max\left(\sum_{t_j=0}^{t_j-127} A(t_i, t_j)\Delta_{b-1}(t_j)\right) \tag{20}$$

The state probability for that frame is then multiplied by the autocorrelation function of the corresponding point in that frame, i.e:

$$\Delta_b(t_j) = \Delta_b(t_j) * A_b(t_j) \tag{21}$$

The velocity of the current frame is indexed by the maximum value of the state probability vector though:

$$t_b = \arg\max(\Delta_b(t_j)) \tag{22}$$

The mapping of points to beat periods can be determined by equation (17).

3) Beat tracking. Beat tracking can use a Hidden Markov Model to predict the exact point in time when a beat occurs based on the start point detection function. The beat tracking algorithm first assigns a state Φ to each point of the start point detection function, representing the distance of the point at that location from the previous beat point. This distance is expressed in points, e.g., if point t is a beat point, then its state Φ_t is 0 and the state Φ_{t-1} of the point at which it is written has a value of 1. Each state produces an observation O, and the sequence of observations is the note onset detection signal.

3.2.3. Tone assessment:

1) Tone signal preprocessing. The acquired student singing timbre signal is preprocessed and unified quantization. Among them, the timbre signal is saved in Wav format, with a precision of 32bit and a

uniform sampling rate of 45.32kHz. The timbre signal preprocessing includes removing mute segments, pre-emphasizing, frame-splitting and windowing, etc.

2) Harmonic structure-based timbre expression spectrum construction. In the construction of the harmonic structure of the timbre expression spectrum, the harmonic structure is transformed based on the fundamental [15]. Therefore, the fundamental is detected first. In this study, the autocorrelation function is used to analyze the fundamental. The specific principle is as follows:

Let $x(n)$ be the time sequence of the signal, i denote the number of frames, L be the length of each frame, and p be the amount of time delay, then the mathematical expression of the short-time autocorrelation function of the i th frame of the speech signal $x_i(l)$ obtained by the windowed frame-splitting process is:

$$R_i(p)=\sum_{i=1}^{L-l}x_i(l)x_i(l+p) \tag{23}$$

According to the definition of autocorrelation function, it can be seen from Eq. (23) that when the period of signal $x_i(l)$ is T, the correspondence between signal $R_i(p)$ and signal $x_i(l)$ under the same period condition is:

$$R_i(P)=R_i(P+T) \tag{24}$$

$R_i(p)=R_i(-p)$ when the short-time autocorrelation function is an even function.

Where the value of the autocorrelation function of a periodic signal can be maximized when $p=0$, i.e., when the delay is an integer multiple of the period ahead or behind.

The mathematical expression for the coefficients of the commonly used autocorrelation function is:

$$ACF_i(p)=R_i(p)/R_i(0) \tag{25}$$

Since the amount of delay is also 0 when $p=0$, the autocorrelation function, i.e. $R_i(0)$, has the largest value. Therefore, the value of $ACF_i(p)$ is always less than or equal to $R_i(0)$, i.e. less than or equal to 1.

In order to make the extracted harmonic structure of the timbre more accurate, the audio feature extraction method of discrete harmonic transform is proposed on the basis of Fourier transform, taking into account the sparse transformation of harmonic frequencies. If the highest harmonic number is m, the set of fundamental frequency vectors of the harmonic spectrum of the student's singing is $H=(f_0,f_1,\cdots,f_m)$. The simulated center frequency of the harmonic spectrum corresponding to the M th spectral line is:

$$f_m=f_0\cdot m \tag{26}$$

Therefore, the bandwidth of the interval between the harmonic spectrum and the harmonic spectrum is constant:

$$B_m=f_{m+1}-f_m=f_0 \tag{27}$$

The ratio p_m of the analog center frequency to the interval bandwidth is then:

$$p_m=f_m/B_m=m \tag{28}$$

Let f_s, be the sampling frequency of the signal and the corresponding gene period of the harmonic spectrum fundamental frequency f_0 be T_o, then:

$$T_0=f_s/f_0 \tag{29}$$

When $f_s \geq 2f_{\max}$ ($f_{\max}$ is the highest frequency component of the signal) is satisfied, the signal does not produce aliasing distortion. And the folding frequency ($f_s/2$) is in the harmonic spectrum and can analyze the highest frequency of the analog signal, therefore:

$$f_{\max} \geq M \cdot f_0 \tag{30}$$

obtained by rounding down:

$$M = \left\lfloor \frac{f_{\max}}{f_0} \right\rfloor = \left\lfloor \frac{\frac{f_s}{2}}{\frac{f_s}{T_0}} \right\rfloor = \left\lfloor \frac{T_0}{2} \right\rfloor \tag{31}$$

For a signal sequence with a window length of N, the spacing of the digital frequencies of the discrete spectrum must satisfy $B_m \cdot N \leq f_s$, so the expression for N can be obtained by rounding down:

$$N = \left\lfloor \frac{f_s}{B_m} \right\rfloor = \left\lfloor \frac{f_s}{f_0} \right\rfloor = \left\lfloor \frac{f_s}{f_s / T_0} \right\rfloor = \lfloor T_0 \rfloor \tag{32}$$

For $B_m \cdot N \leq f$, taking the equal sign and combining the ratio of the center frequency to the bandwidth shows that the digital frequency at spectral line m in the harmonic spectrum is:

$$F_m = \frac{f_m}{f_s} = \frac{m}{N} \tag{33}$$

Let $x(n)$ be a finite-length sequence of tone signals, N be the length of the window corresponding to the time-frequency transform of the harmonic spectrum, and the expression for the m rd spectral component of the discrete harmonic transform is:

$$X^{ks}(m) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\,\omega_N(n) e^{-j\frac{2mw}{N}n} \tag{34}$$

where $m = 1, 2, \cdots, M$ and M are the highest harmonic numbers, and $\omega_N(n)$ is a window function of length N.

Let the highest harmonic number of the constructed timbre feature be M, when a frame of musical signal is $x(n)$, the fundamental period of the singing signal is T_0. The harmonic information of the musical signal is constructed according to the expression of the discrete harmonic transform, which can be obtained by assuming that the length of the window of the harmonic transform is $N = |T_0|$:

$$DHT(m) = \frac{2}{N}\left|\sum_{n=0}^{N-1} x(n) w_N(n) e^{-\frac{2\pi \mathrm{n}}{N}}\right| \tag{35}$$

where $m = 1, 2, \cdots, M$, and $M \leq \left|\frac{T_0}{2}\right|$.

Let DHTC(m) denote the result after normalization of DHT(m), which is the constructed timbre feature. According to HTC(m), the first-order differential timbre feature $\Delta DHTC(m)$ and the second-order differential timbre feature $\Delta(\Delta DHTC(m))$ are constructed.

3) Tone classification. In order to make the classification of students' singing timbre more accurate, support vector machine (SVM) is used to classify and recognize the timbre of students' singing [16]. Therefore there is an objective function:

$$\min \frac{1}{2} \square w \square^2 + C \sum_{i=1}^{N} \xi_i \tag{36}$$

$$s.t. y_i (w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \cdots, N \tag{37}$$

Introducing Eq. (36) into the Lagrange operator gives:

$$\min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j (x_i \cdot x_j) - \varepsilon \sum_{i=1}^{N} \xi_i \tag{38}$$

$$s.t. \sum_{i=1}^{N} a_i y_i = 0, 0 \leq a_i \leq C \tag{39}$$

Eqs. (38) and (39) show that C controls the model fit, ε controls the generalization ability and error pass approximation ability.

3.2.4. Comprehensive scoring. The purpose of vocal scoring is to give an objective assessment of the student's grasp of the melody of a piece of music, mainly including pitch, rhythm and timbre. Higher scores indicate that the student's interpretation of the piece is more accurate, while lower scores indicate that the student's grasp of the melody of the piece is less accurate. The scoring formula is as follows:

$$score = k_1 \frac{100}{1 + a_1 (\min disv)^{b_1}} + k_2 \frac{100}{1 + a_2 (\min disp)^{b_2}} + k_3 \frac{100}{1 + a_3 (st)^{b_1}} \tag{40}$$

where $a_1, a_2, a_3, b_1, b_2, b_3 > 0; k_1 + k_2 + k_3 = 1, k_1, k_2, k_3$ is the weight of each scoring parameter in the scoring mechanism. $\min disv$ and $\min disp$ are the pitch and rhythm parameters, respectively. st represents the timbre smoothness parameter. The selection of weights can be adjusted according to different requirements or different scoring priorities. In order to make the computer better simulate expert scoring, the weights can be trained to find out the best mapping relationship between computer and manual scoring.

4. System application analysis

In order to test the auxiliary role of a series of artificial intelligence algorithms proposed in this paper in the teaching of vocal music, the second-year students of vocal music majors of a university are selected as the experimental research objects, and the system proposed in this paper is applied to conduct a series of experiments to analyze the effectiveness of this paper's system in practical application. The following is the analysis of the experimental results of the core functions of this system and the analysis of the effectiveness of the practical application of this system.

4.1. Learning path recommendation analysis

4.1.1. Data sets. In order to verify the effectiveness of the learning path recommendation algorithm proposed in this paper, a series of experiments were conducted. The experimental data include not only the data of learning resources, but also the records of learners' access to resources. The existing EdNet open dataset provides up to 20 attributes related to vocal music teaching, including course information, learner base data, learner type data and learner behavior data. In order to verify the generality of the algorithm, the dataset in this section adds 15 additional experimental colleges and universities' voice teaching course information as well as student behavior information to the Edx dataset. The final dataset contains 28 vocal music teaching course information and 21,963 students' learning behavior information.

4.1.2. Experimental results. In previous studies, classical recommendation algorithms are mainly used for learning recommendation, such as Convolutional Neural Network (CNN) based recommendation algorithm, Collaborative Filtering based recommendation algorithm (CF) and Content Based Recommendation (CB). In this paper, these algorithms are used in experiments for learning path recommendation and compared with the algorithm based on the combination of LSTM and attention mechanism (LSTM-A) proposed in this paper.

1) Comparison of accuracy between models. Figure 3 shows the calculation results of the mean absolute error (MAE) and root mean square error (RMSE) of each model in the learning path recommendation process, and the smaller values of MAE and RMSE indicate the higher accuracy of the model. For the dataset of this experiment, the recommendation accuracy based on deep learning recommendation algorithms (CNN, LSTM-A) is higher than that of traditional recommendation algorithms (CB, CF), which indicates that deep learning algorithms are more advantageous when dealing with a large number of datasets with complex structures. And compared with CNN, LSTM-A can again have higher accuracy. Through experiments, it can be found that LSTM-A has a very good performance in learning path recommendation.

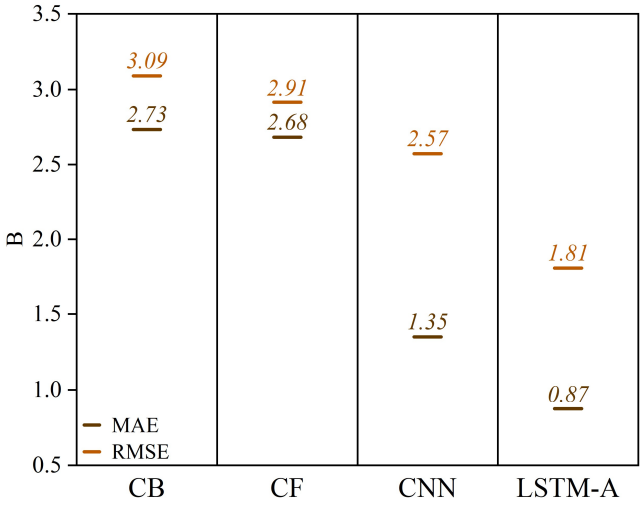
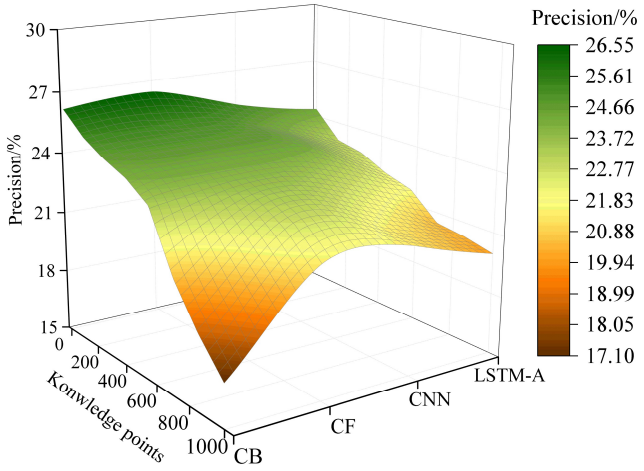
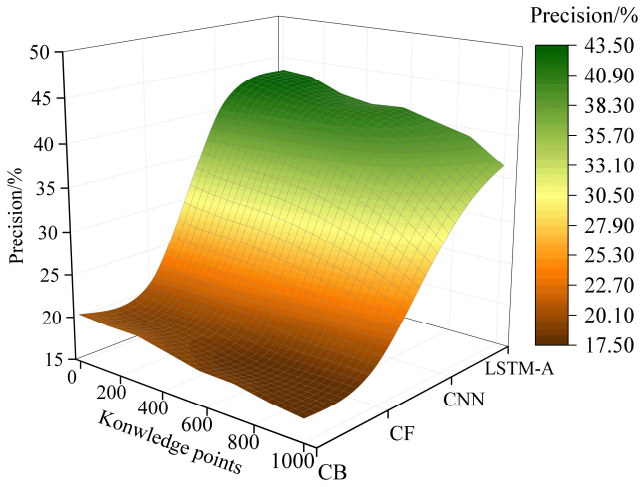


Fig. 3. Model accuracy comparison results

2) Recommendation precision and recall comparison experiments. Figure 4 shows the precision of four different recommendation algorithms based on different numbers of knowledge points for learning results and prediction, for different numbers of learners and learning features, a comparison is made between the two cases, where r indicates the number of learner features, l indicates the number of learners, and (a) and (b) are the recommendation precision of the model at $r=10, l=5000$ and $r=20, l=10000$, respectively. From Fig. 4(a), it can be seen that the prediction accuracies of all four algorithms perform generally when both the number of learners and features are relatively small, and the prediction accuracies of all four algorithms show a decreasing trend when the number of knowledge points rises. After increasing the number of learners and features in Fig. 4(b), the collaborative filtering algorithms and content-based recommendation algorithms are significantly less capable, which illustrates that the deep learning algorithms are more advantageous when dealing with large datasets. For the content-based recommendation algorithm, the fluctuation is particularly pronounced as the knowledge point rises, probably due to the fact that the algorithm is unable to tap into the potential interests of the students and thus lacks diversity in recommendations.



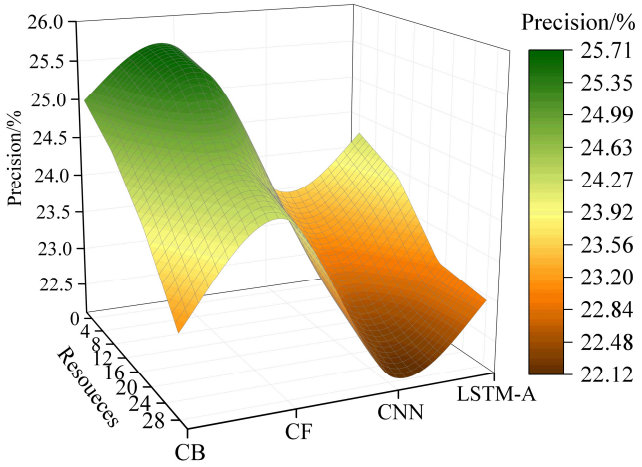
(a) $r = 10, l = 5000$



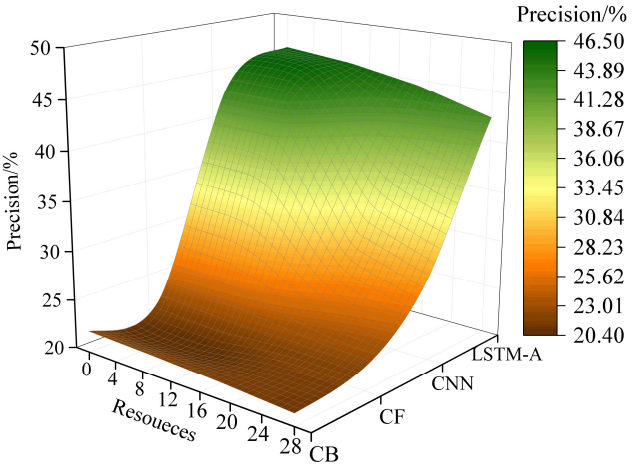
(b) $r = 20, l = 10000$

Fig. 4. Comparison results of different scenarios based on the quantity of knowledge

Fig. 5 shows the prediction accuracy of the four algorithms on the learning effect based on different number of resources, (a) and (b) indicate the results when the knowledge point $kp = 500、1000$ and the number of learners $l = 5000、10000$, respectively. The figure shows that when the number of learners is 5000, the increase of resources has no significant effect. If the number of learners reaches 10000, there is a significant effect with the increase of resources. It can be seen from Figure 5(b) kind that when the number of resources is increasing, the accuracy of CCN prediction of the results fluctuates more dramatically, while LSTM-A performs stably. This situation may be due to the existence of the gradient vanishing problem. The LSTM-A network is able to have a good solution for gradient vanishing and gradient explosion due to the existence of the mechanism of self-loop and gate control and the introduction of the attention mechanism.



(a) $kp = 500, l = 5000$



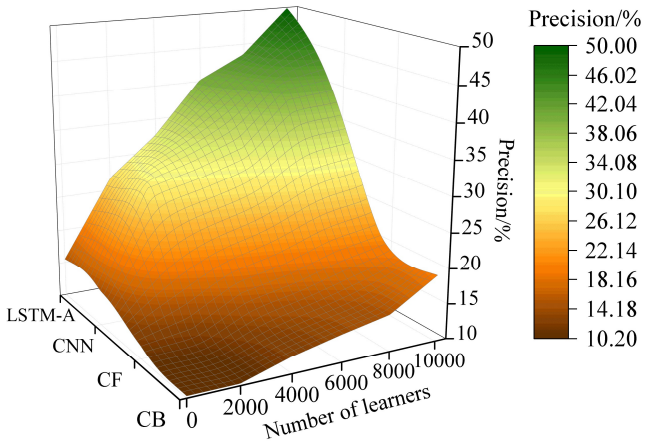
(b) $kp = 1000, l = 10000$

Fig. 5. Comparison results of different scenarios based on the number of resources

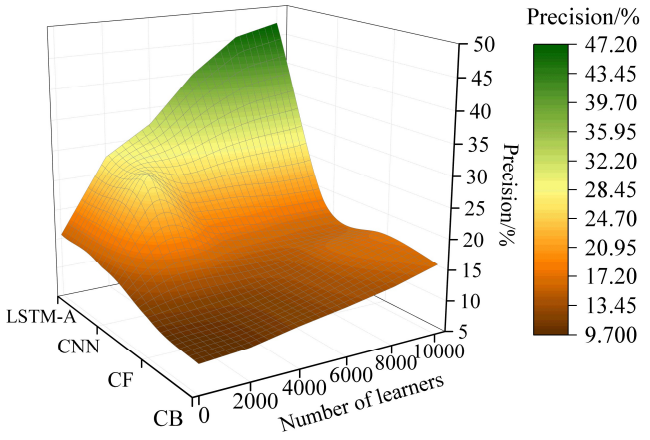
This paper compares three classical methods with the method proposed in this paper. A comparative study was conducted with two different parameter settings for the knowledge points, resources and number of learners in the dataset. As shown in Fig. 4, when the number of learners is 5000, similar to the situation of the school program, the learning effect of the classical method is slightly higher than that of the deep learning method. However, when the number of learners increases to 10,000, the deep learning method performs better and the performance improves by about 20%. In addition, as the number of knowledge points increases, the accuracy of all algorithms decreases to varying degrees, and thus more challenges are encountered when making recommendations. As can be seen in Figure 5, the increase in learning resources has a limited impact on the algorithms. Overall, the accuracy keeps decreasing as the resources increase. For the larger dataset ($l = 10000$), such an effect is not significant. However, for the smaller dataset ($l = 5000$), such an effect is observed.

Figures 6 and 7 show the comparison results of precision and recall under different scenarios based on the number of learners, (a) and (b) for $kp = 500, r = 10$ and $kp = 1000, r = 20$, respectively. The information in the figure shows that the number of learners correlates with the system performance, i.e., the accuracy of the algorithm increases with the increase of the dataset, especially for the algorithm proposed in this paper. For the traditional methods, there is no significant increase in the accuracy. As can be seen from the figure, the accuracy is increasing with the number of learners while

the recall is decreasing, and the traditional recommendation algorithms are still not capable enough when dealing with more knowledge points as well as more learner characteristics. Compared with the classical deep learning recommendation algorithm CNN, LSTM-A network based on the fusion of LSTM and attention mechanism has a higher and more stable prediction accuracy. Experiments demonstrate that the LSTM-A model performs better than other learning path recommendation algorithms when dealing with large datasets with vocal music teaching course information and learner behavioral characteristics.

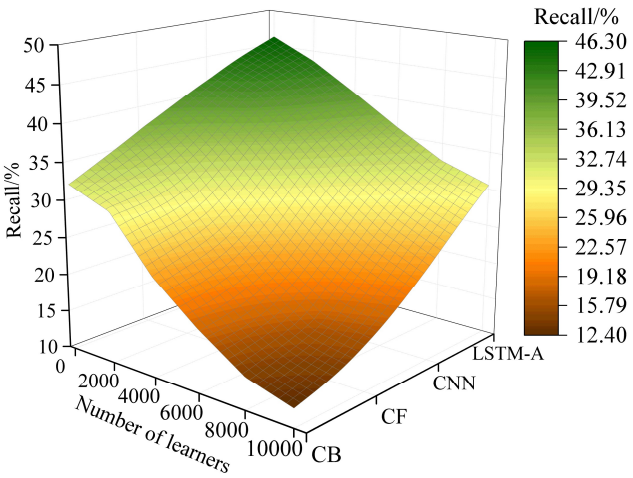


(a) $kp = 500, r = 10$

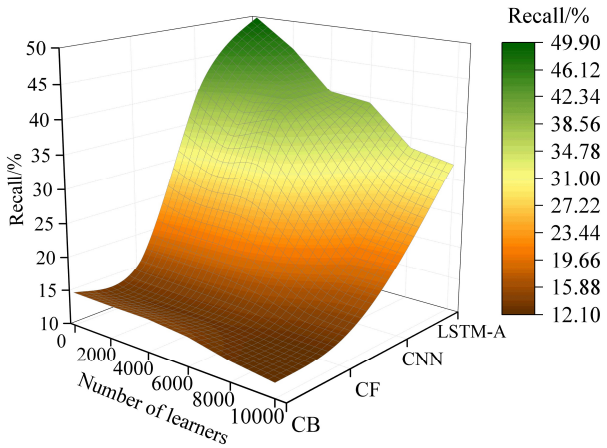


(b) $kp = 1000, r = 20$

Fig. 6. Comparison results of different scenarios based on the number of learners



(a) $kp = 500, r = 10$



(b) $kp = 1000, r = 20$

Fig. 7. The recall rate compares results

4.2. Intelligent Assessment Analysis

4.2.1. Sound level assessment. In order to verify the feasibility of the pitch assessment method in this paper, a student was randomly selected from the experimental sample to sing 30 repertoires, including Song of the Setting Sun and Tibetan Plateau. The pitch of the song sung by the student was calculated using the physical model of human voice pitch constructed in this paper and compared with the standard pitch of the 30 repertoire, and the results are shown in Fig. 8, where the red vertical line indicates that the singing pitch is higher than the standard pitch, and the black vertical line indicates that the singing pitch is lower than the standard pitch. Overall, there is a certain gap between the singing pitch of the randomly selected students and the standard pitch of the 30 songs, in which the largest difference between the singing pitch and the standard pitch is track 5, with an absolute value of 0.115, and the smallest difference is track 27, with an absolute value of 0.003. It can be seen from the data in the figure that there is still room for improvement in the student's singing pitch for the 30 songs, and further learning paths can be recommended by the system. The learning path recommended by the system can be further practiced to narrow the gap between the singing pitch and the standard pitch.

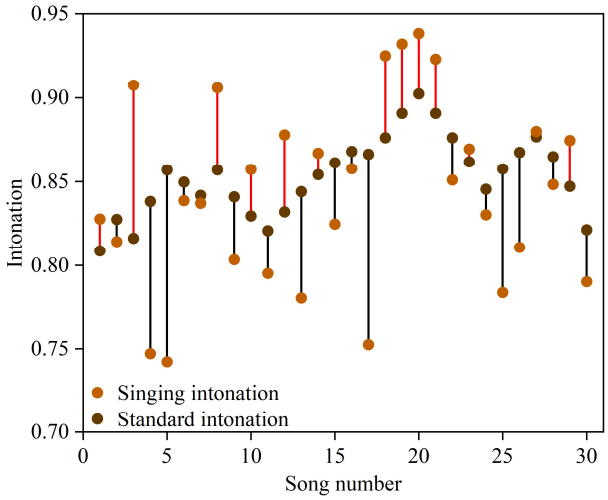
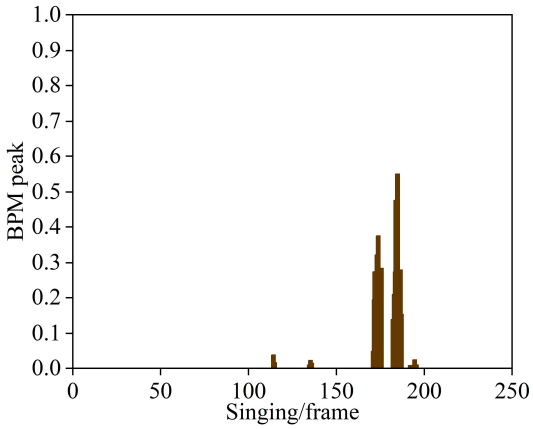


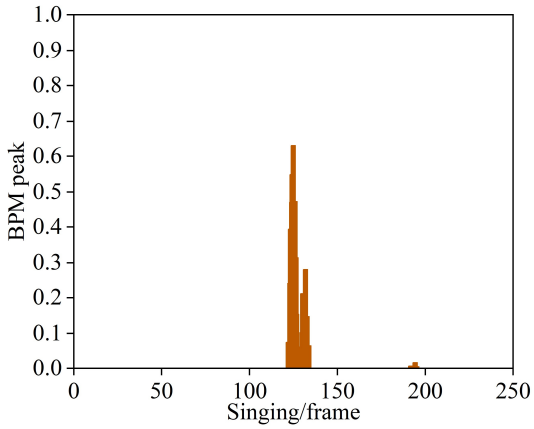
Fig. 8. Intonation assessment results

4.2.2. Rhythm assessment:

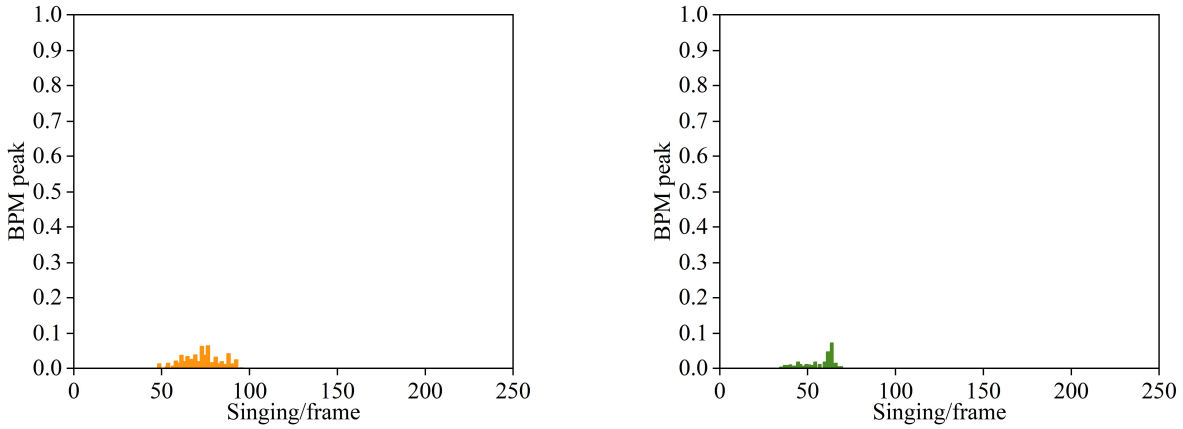
1) Rhythm feature extraction analysis. In order to verify the usability of the automatic extraction method of rhythmic features of students' singing designed in this paper, the frequency segments of songs sung with four emotions, namely, anger, excitement, relaxation, and sadness, were selected in 2 songs, “Don't Listen to Slow Songs if You're Sad” and “The Law of Everything Blooming”, respectively, and 50 segments of each emotion were sung with each emotion. The method of this paper was used to extract the rhythmic features of 200 song singing segments, obtain the histograms of the BPM distribution of different emotional singing, and describe the differences of different emotional singing. The histograms of the distribution of the rhythmic features of singing with different emotions are shown in Fig. 9, with (a)-(d) indicating the four emotions of anger, excitement, relaxation and sadness, respectively. From the figure, it can be seen that the BPM peaks of anger and excitement are higher in the singing segments of different emotions, which indicates that the singing rhythmic features are more significant and easy to be extracted when there are anger and excitement-inducing segments in the song singing. The sound effects and background music production of the anger and excitement emotional segments mostly used percussion instruments. In contrast, the BPM peaks of the relaxed and sad mood clips are lower, indicating that the singing rhythmic features are not obvious and not easy to be extracted in these clips. For such clips, orchestral instruments were used more often.



(a) Anger



(b) Excitation



(c) Ease (d) Sad

Fig. 9. Different mood singing rhythm characteristics distribution histogram

This paper's method (Method 1), the feature extraction method based on spectral energy distribution (Method 2) and the feature extraction method based on tone-related fundamental frequency (Method 3) were used to classify the emotions of the selected 250 singing frequency bands, and the obtained results are shown in Tables 1-Table 3, respectively. According to the statistical results of emotion classification in the table, the accuracy rate of this paper's method for extracting students' singing rhythmic features for emotion classification is higher than that of the singing feature extraction method based on spectral energy distribution and the feature extraction method based on intonation-related fundamental frequency. The method based on tone-related fundamental frequency is better than the method based on spectral energy distribution for the classification of anger and excitement, indicating that the method is more effective for the classification of singing with higher BPM peaks. The differences in the classification detection indexes for the four emotions are relatively smooth, and the classification effect is also better for the singing of relaxed and sad emotions with lower BPM peaks, indicating that the method in this paper can accurately extract the rhythmic features of the students' singing, which is conducive to the further assessment of the students' singing rhythms.

Table 1. Method 1 emotional classification results

Emotions	Anger	Excitation	Ease	Sad
Anger	48	1	0	1
Excitation	2	47	1	0
Ease	2	4	42	2
Sad	3	1	2	44

Table 2. Method 2 emotional classification results

Emotions	Anger	Excitation	Ease	Sad
Anger	37	6	4	3
Excitation	8	35	1	6
Ease	11	7	28	4
Sad	1	9	16	24

Table 3. Method 3 emotional classification results

Emotions	Anger	Excitation	Ease	Sad

Anger	33	9	5	3
Excitation	12	30	1	7
Ease	4	6	23	17
Sad	7	9	13	21

2) Rhythm evaluation analysis. The same 30 songs as in the previous section were selected and sung by the same randomly selected student, and the rhythmic features of his singing segments were extracted using the method of this paper and compared with the standard rhythm, and the results are shown in Figure 10, where the red vertical line indicates that the singing rhythmic BPM peak is higher than the standard rhythm, and the black vertical line indicates that the singing rhythmic BPM peak is lower than the standard rhythm. As can be seen from the graph, the student's rhythmic control for the 30 songs was significantly better than his control of pitch. The absolute value of the difference between the singing rhythm BPM peak and the standard rhythm BPM peak is 0.022, and the minimum value is almost 0. According to the results of the systematic evaluation, this student should focus more on the practice of pitch in the process of vocal teaching, and for the rhythm of the songs, he/she only needs to focus on the practice of the songs with poorer rhythms.

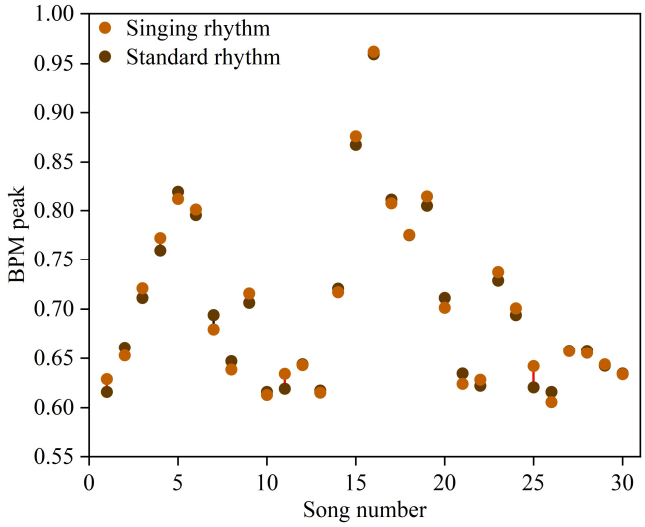


Fig. 10. Rhythm assessment result

4.2.3. Tone assessment. From the experimental samples in this paper, 100 students of different genders were selected to perform six different song singing timbres, such as soprano, mezzo-soprano, alto, tenor, baritone, and bass, respectively. The timbre identification and classification method constructed based on the harmonic structure of this paper is utilized to identify the timbre of the students' singing segments, and the results are shown in Fig. 11. The data in the figure firstly reflect that the timbre classification and recognition method of this paper can effectively realize the recognition of different genders of men and women, and there is no case of identifying female timbre as male timbre or male timbre as female timbre. At the same time, this paper's method for male and female students singing timbre recognition accuracy is high, the lowest accuracy rate is as high as 91%. Although the method is less effective in recognizing soprano and tenor compared with other tones, the experimental results can still verify that the classification and recognition method of students' singing tones based on the harmonic structure proposed in this paper has significant effectiveness.

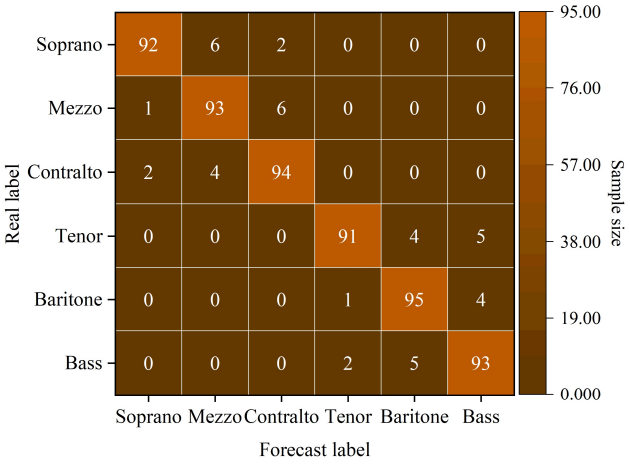


Fig. 11. Sound color identification results

According to the identification results of timbre classification, a student was randomly selected from the tenors to sing 30 songs including "Dare to Ask Where the Road is" and "The Great Wall of China Never Falls", etc., which contain the treble part, and the timbre amplitude when the student sang was calculated and compared with the standard timbre amplitude, and the results are shown in Fig. 12, in which the red vertical line indicates that the student sang with timbre amplitude greater than the standard timbre amplitude, and the black vertical line indicates that the student sang with timbre amplitude less than the standard timbre amplitude. The red vertical line indicates that the students' singing tone amplitude is greater than the standard tone amplitude, and the black vertical line indicates that the students' singing tone amplitude is smaller than the standard tone amplitude. From the figure, it can be seen that the timbre amplitude of the selected students' singing for the 30 songs is almost identical to the standard timbre amplitude. This further demonstrates the accuracy of this paper's timbre classification and identification method for students' singing timbres, and also shows that this paper's method can effectively realize the evaluation of students' singing timbres.

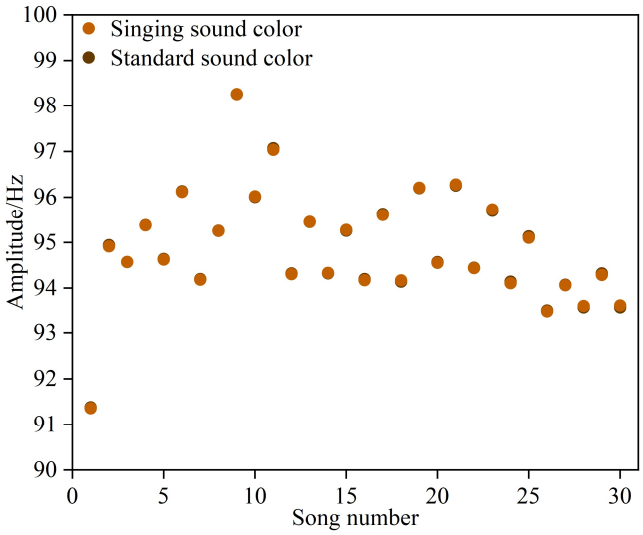


Fig. 12. Sound color assessment result

4.3. Analysis of application effects

In the application effect analysis, the results of the implementation of the vocal teaching aid system constructed based on artificial intelligence algorithms in this paper were evaluated exhaustively. Through quantitative and qualitative methods, the assessment of students' vocal skills and singing

ability found that students participating in the study significantly improved their tone control, pitch and rhythm, and were able to use breath and resonance more freely. During the teaching process, students' expression of musical emotions was more delicate and rich, and their expressiveness was enhanced. At the same time, students' overall musical literacy and stage performance also improved, and through in-depth study of the systematic recommended vocal theory courses and techniques, students demonstrated a deeper understanding and recognition of vocal culture. Utilizing questionnaires, interviews, and comprehensive scoring of singing ability, the data collected indicate that the system in this paper effectively promotes students' ability to apply themselves in a variety of musical styles, laying a solid foundation for their future vocal career development. This system is of great reference significance in enhancing the comprehensive quality of students majoring in vocal music, showing its important auxiliary role and application value in vocal music teaching.

5. Conclusion

This paper designs and implements a vocal music teaching assistance system based on artificial intelligence algorithms, which contains functions such as personalized learning path recommendation and intelligent evaluation of students' singing practice. The results of the experimental study show that:

1) The learning path recommendation method of long and short-term memory network (LSTM) fused with attention mechanism adopted by the system in this paper has better performance than the recommendation algorithms based on convolutional neural network, collaborative filtering and content. It is still able to maintain high recommendation accuracy and recall under large-scale datasets. It fully demonstrates that the personalized learning path recommendation method included in the system of this paper can provide effective personalized guidance for students to learn vocal knowledge or practice vocal skills.

2) In the assessment of students' pitch, the randomly selected students' pitch performance is poor when they sing 30 pieces, and the absolute value of the difference between their singing pitch and the standard pitch can be up to 0.115, whereas the rhythm assessment of their singing performs better relative to the pitch assessment, and the absolute value of the difference between their singing pitch and the standard rhythmic BPM peak is only 0.022, and the minimum is almost 0. The results of the tone assessment also verify that the intelligent learning path recommendation method designed in this paper can provide effective personalized guidance for students to learn vocal knowledge or practice vocal skills. The results also verify the effectiveness of the intelligent evaluation method designed in this paper.

Combined with the experimental results and the application effect of the system, it can be concluded that the vocal teaching assistance system designed based on the artificial intelligence algorithm in this paper can give full play to its auxiliary role in the process of vocal teaching. It has far-reaching significance for promoting the intelligent development of vocal music teaching and improving students' vocal music professionalism.

References

[1] Yuan, Y. (2024). Influencing Factors and Modeling Methods of Vocal Music Teaching Quality Supported by Artificial Intelligence Technology. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 19(1), 1-16.

[2] Han, B. (2024). Information intelligent teaching method and its influence on vocal music teaching effect. International Journal of Embedded Systems, 17(1-2), 98-112.

[3] Zhang, T. (2024). Research on Online Vocal Music Smart Classroom-Assisted Teaching Based on Wireless Network Combined With Artificial Intelligence. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 19(1), 1-19.

[4] Mi, H. (2024). Application of Technological Means and Innovative Teaching Methods in Vocal Music Education. Journal of Modern Educational Theory and Practice, 1(1).

[5] Wu, R. (2023). A Hybrid Intelligence-based Integrated Smart Evaluation Model for Vocal Music Teaching. IEEE Access.

[6] Cui, X., & Chen, M. (2024). A novel learning framework for vocal music education: an exploration of convolutional neural networks and pluralistic learning approaches. Soft Computing, 28(4), 3533-3553.

[7] Chen, W. (2022). Design of music teaching system based on artificial intelligence. Mathematical Problems in Engineering, 2022(1), 2627395.

[8] Liu, C., & Li, S. (2023, June). Design of Interactive Teaching Music Intelligent System Based on AI and Big Data Analysis. In International Conference on Computational Finance and Business Analytics (pp. 333-341). Cham: Springer Nature Switzerland.

[9] Liu, M., & Huang, J. (2021). Piano playing teaching system based on artificial intelligence–design and research. Journal of Intelligent & Fuzzy Systems, 40(2), 3525-3533.

[10] Wang, X. (2022). Design of vocal music teaching system platform for music majors based on artificial intelligence. Wireless Communications and Mobile Computing, 2022(1), 5503834.

[11] Xuelian, H. (2023). Development of music teaching software based on neural network algorithm and user analysis. Soft Computing, 1-9.

[12] Ma, X. (2021). Analysis on the Application of Multimedia-Assisted Music Teaching Based on AI Technology. Advances in Multimedia, 2021(1), 5728595.

[13] Yang, F. (2020, November). Artificial intelligence in music education. In 2020 International Conference on Robots & Intelligent System (ICRIS) (pp. 483-484). IEEE.

[14] Guoli Xu & Cora Un In Wong. (2024). Deep Learning-Based Educational Image Content Understanding and Personalized Learning Path Recommendation. Traitement du Signal(1).

[15] Jea-Yul YOON,Chai-Jong SONG & Hochong PARK. (2013). Predominant Melody Extraction from Polyphonic Music Signals Based on Harmonic Structure. IEICE Transactions on Information and Systems(11),2504-2507.

[16] Lhoucine Bahatti,Omar Bouattane,My Elhoussine Echhibat & Mohamed Hicham Zaggaf. (2016). An Efficient Audio Classification Approach Based on Support Vector Machines. International Journal of Advanced Computer Science and Applications(5).