ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Exploring the Reconstruction and Development Path of Traditional Choral Teaching Mode in the Intelligent Era

Ying Zhou[1,*] *and Yvwei Lu*[2]

[1] Music College, Zhaoqing University, Zhaoqing, Guangdong, 526061, China
[2] Guangzhou Cigarette Factory, Guangzhou, Guangdong, 510145, China

SUMMARY: *This paper utilizes intelligent technology to assist repertoire teaching and singing instruction, combining with the existing computer music software, to explore the development path of the traditional choral teaching mode in the intelligent era. For the teaching of choral repertoire, a real-time music beat tracking algorithm is proposed, which carries out wavelet transform on the pre-processed music signals, detects the peaks of the resulting detail coefficients, constructs and solves the smooth histogram of the music beats, and obtains the real-time value of the music beats. In addition, in order to improve teachers' guidance to students' singing, deep convolutional networks with powerful dimensionality reduction and feature learning ability are used to embed high-dimensional and time-sequential vocal spectral features into the 3-dimensional timbre embedding space, to realize the characterization and similarity metrics of vocal timbre in the 3-dimensional timbre embedding space, and to build a vocal timbre characterization model. The experimental samples are selected and the experimental group and control group are set up, in which the students in the experimental group have 1-6 different semitones of range broadening with the technical assistance of the algorithms and models in this paper, which verifies the important technical roles of real-time music beat tracking algorithms and vocal coloration characterization models in reconstructing the traditional choral teaching mode.*

KEYWORDS: *music beat tracking; human voice color representation; traditional chorus; teaching mode reconstruction*

1 Introduction

In the context of the information age of the 21st century, the rapid development of artificial intelligence (AI) is affecting our life and work in an unprecedented way. The field of education is also experiencing this wave of change, especially in music teaching [1-3]. The introduction of AI technologies not only promotes the innovation and restructuring of teaching methods, but also provides new opportunities to improve teaching quality and efficiency [4]. The application of these technologies is gradually transforming from the stage of theoretical research to concrete practical application, and driving a series of revolutionary changes in the field of music teaching.

In music teaching, chorus, as one of the main teaching contents, aims to cultivate students' sense of music, make students feel the joy of music and enrich music cognition [5]. With the steady progress of teaching reform, the traditional choral teaching method has exposed many problems, which is not conducive to improving students' comprehensive literacy [6-8]. In order to solve these kinds of problems, teachers should have an innovative consciousness and

*13006888686@163.com
https://doi.org/10.65102/is2026043
Published: 30 April, 2026

constantly try new theories and methods to construct a brand new choral teaching mode according to the requirements of the new standard for choral teaching [9-13]. However, at present, the research of AI in music education mainly focuses on higher education and related professional fields [14], while the exploration of its application in choral teaching is relatively less, which leads to our incomplete understanding of the potential of the application of AI technology in choral teaching [15]. Therefore, by exploring the potential and possible challenges of AI technology in choral teaching in the intelligent era, in order to support the reform and development of middle school music teaching [16-18]. By analyzing how AI technology can optimize the teaching process, improve the quality of teaching, and promote personalized learning, it will provide music teachers with new teaching strategies and tools [19]. The challenges of introducing AI technology, such as teacher professional development and learning assessment standards, will also be discussed [20, 21]. Through comprehensive analysis and discussion, we aim to provide scientific suggestions for the reconstruction and reform of the choral teaching mode and further promote the development of the field.

Based on the overall framework of the real-time music beat tracking algorithm, this paper demonstrates the implementation process of its data preprocessing, wavelet transform, peak check, histogram construction, real-time and beat point location determination in order. Design the structure of vocal color embedded training model, analyze the characterization of input signals and the content of pairwise training method in it. After completing the technical preparations, the training methods of rhythm, vocalization, two-part harmony, three-part harmony, and dissonance and chordal harmony for choral teaching under Cubase 12 software are discussed in turn, and compared with the traditional teaching model methods. The performance of the real-time beat tracking algorithm and the human voice color representation model were examined respectively, and after confirming that they could be put into practical application, the experimental samples were selected and set up with the corresponding teaching methods. By comparing the performance of the samples in terms of range and acoustic parameters, the effectiveness of reconstructing the traditional choral teaching mode based on the algorithms and models in this paper is explored.

2 Real-Time Music Beat Tracking and Vocal Chromatic Representation

This chapter is divided into two main parts: (1) designing real-time music tracking algorithms to provide technical assistance to teachers for student-oriented choral teaching; (2) establishing a model for human voice color representation to assist teachers in precise instruction and correction of students' singing skills and so on.

2.1 Real-time music beat tracking algorithm

The real-time beat tracking algorithm flow is shown in Figure 1.

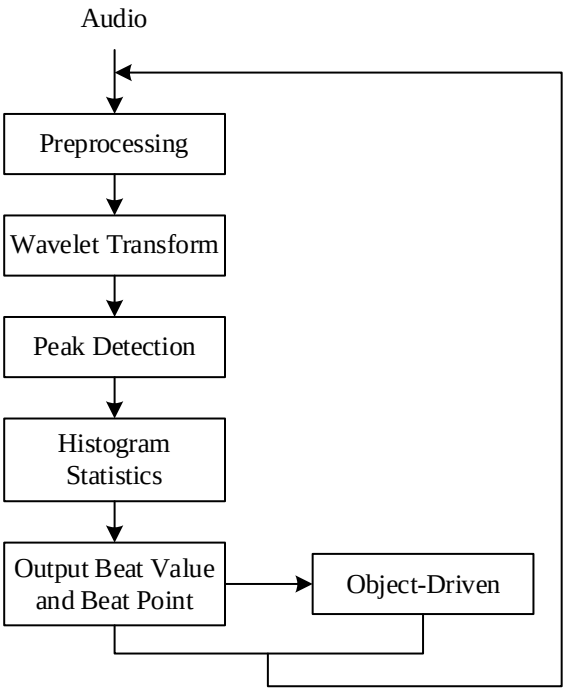

Figure 1: Overall framework of the algorithm

2.1.1 Pre-processing

When converting analog sound to digital sound, the sound signal is discretized in amplitude, and the length of the data at the sampling point is represented by the number of quantization bits. Different quantization bits represent different data ranges of the sampling points. For uniform processing, the acquired audio data is normalized by normalizing the sample point data to $[-1,1]$. Assuming that the sequence of sampling points is $S[]$, the sequence of sampling points after normalization is $N_s[]$, using equation (1):

$$N_s[] = (S[] - S_{\min}) / (S_{\max} - S_{\min}) * (D_{\max} - D_{\min}) + D_{\min} \tag{1}$$

Normalized processing sample points, $S_{\max}$, $S_{\min}$ are the maximum and minimum values in the sequence of sample points, and $D_{\max}$, $D_{\min}$ are the maximum and minimum values expected in the sequence of sample points after normalization, respectively. The following music signals all refer to the normalized signals.

2.1.2 Wavelet transform

A discrete wavelet transform is done on a music signal, where the signal is transformed once to produce an approximate part and a detailed part. The decomposition process is repeated continuously for the newly generated approximate part. The algorithm in this paper requires 4 such wavelet decompositions, which will produce 4 sets of coefficient values. For every wavelet decomposition the signal undergoes, its sampling rate decreases by a factor of two.

2.1.3 Peak detection

The peak detection algorithm is applied to the four sets of detail coefficients generated in the first step as follows:

(1) Full-wave correction: all data in the audio signal are converted to their absolute values.

(2) Windowing for extreme values: Windows must be added for short-time analysis, which should theoretically be done using windows with smooth excess properties at both ends. The full-wave corrected data to add window processing, in the window for the local extremes at the same time sliding window, and constantly for the extremes, based on the local extremes in a window range of local extremes the same number of times to determine the final peak point, if the same number of times is greater than 90% of the width of the window, it is considered to be the peak of the extremes that is.

2.1.4 Constructing histograms

After the peaks of the n group of wavelet coefficients are detected, the time interval IOI between every two neighboring beat points is calculated, assuming that all peaks are beat points.

Transform the transformed interval values to the interval values in the original signal. Assume that the current interval value is the value computed after the nth wavelet transform of the original signal, denoted as $IOI2$, and assume that this value is represented by $IOI1$ in the original signal, then $IOI1 = IOI2 * 2^n$.

Construct a beat interval histogram based on the probability of occurrence of the beat interval. Observe the beat interval with the highest probability of occurrence in the histogram and obtain the beat value by equation (2):

$$B_v = 60/\text{Maximum Probability Beat Interval} \tag{2}$$

Improve the accuracy of beat extraction using smoothed histograms.

Introducing beat interval weights. The weight of a beat interval value is estimated by judging the similarity of the values of other beat intervals close to that beat interval. And if the extreme values are simply selected without considering the distribution of the histogram, the calculated beat values will be inaccurate. To solve the above problem, the histogram is convolved with a Gaussian function to smooth the histogram. The Gaussian function highlights areas of high density. After smoothing, the beat interval value can be determined from the maximum magnitude and the final beat value can be derived using equation (2).

2.1.5 Real-time

Set a counter counter to indicate the current time unit number to be processed, according to the value of counter to determine the operation to be carried out as follows:

(1) ifcounter%1=0: This step is performed on all data units of the audio; the wavelet transform is performed on this unit of data, the resulting detail coefficients are peak detected, and the estimated value of the beat interval obtained after detection is restored to the position in the original signal, and then counted with a histogram for the first layer of results;

(2) ifcounter%2=0: the approximation coefficients obtained from the wavelet decomposition of the previous layer generated by the two neighboring units of data are merged, and wavelet decomposition continues to be done on them, and the estimated value of the beat spacing obtained from the newly generated detail coefficients after the peak detection is restored to the position in the original signal is merged in a histogram for the results of the second layer;

(3) ifcounter%4=0: the process is the same as (2), for the result of the third layer;

(4) ifcounter%8=0: the process is the same as (2), for the result of the fourth layer. Up to this step, a part of the data has been a complete four wavelet transform, you can solve the histogram to derive a beat value; as the amount of raw data continues to increase, the beat value will be updated, and the beat point will be updated, converging on the final result. And so on,

until all the data are processed, the overall process is shown in Figure 2.

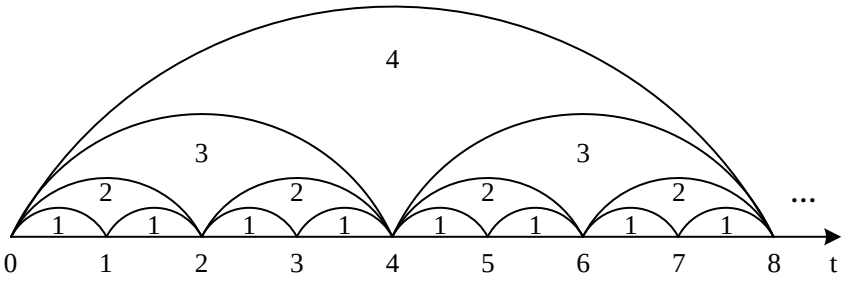

Figure 2: Real-time beat tracking processing

Each time the counter reads in data in multiples of eight, it will be able to complete the calculation of four wavelet decompositions, and it will also be able to calculate the beat value and update the beat point position once, realizing real-time beat tracking.

Assuming that there is an audio clip of length 10 seconds, the data of 0.125 seconds is set as a time unit, and each time 1 second of data is read in the beat tracking algorithm is updated once with the beat value and the position of the beat point according to the beat tracking algorithm.

The variables used are as follows:

static const int TimesForLoop=4

vector<double> part_ca[TimesForLoop]

vector<double> part_cd[TimesForLoop]

vector<vector<double>> ca[TimesForLoop]

vector<double> para_ca[TimesForLoop-1]

The value of counter counter is initialized to 0 and the data is read into the buffer in real time. The value of counter is 2 when the data is read in up to 0.25 seconds, and (1) is executed after the first condition is satisfied, the wavelet transforms the data in that time unit, the peak detection and histogram statistics are performed for the detail coefficients $part\_cd[0]$, and the approximation coefficients $part\_ca[0]$ are added to $ca[0]$ to the results of the first layer.

At the same time the value also satisfies (2), the approximate coefficients after wavelet decomposition of the first layer of two adjacent time units are combined into $para\_ca[0]$, wavelet transform is done on it, and the beats with detail coefficients of $part\_cd[0]$ are detected in real time, and the approximate coefficients $part\_ca[1]$ is appended to $ca[1]$ for the result of the second layer.

The counter value is 4 when enough data is read in for 0.5 seconds, satisfying condition (3). Combine the approximation coefficients from the wavelet decomposition of the second layer of two neighboring time units into $para\_ca[1]$, do wavelet transform on it, detect the beats with detail coefficients of $part\_cd[2]$ in real time, and append the approximation coefficients $part\_ca[2]$ appended to $ca[2]$ for the result of the third layer.

The data in the buffer reaches 0.625 seconds when the counter value becomes 5, which only satisfies the condition (1), and the processing is shown in the processing when the counter is 1. Similarly, the processing when the counter is 6 is the same as the processing when the counter is 2.

The value of the counter of the data read in for 1 second is 8, which satisfies all four

conditions, and the processing of the first three conditions is the same as the processing when the counter is 4. For the fourth condition, the approximate coefficients from the wavelet decomposition of the third layer of the two adjacent time units are combined to $para\_ca[2]$, wavelet transform is done, the beats with detail coefficients of $part\_cd[3]$ are detected in real time, and approximate coefficients of $part\_ca[3]$ are appended to $ca[3]$ for the result of the fourth layer. The processing of the third layer when COUNTER is 8 is shown in Fig. 3, and the processing of the fourth layer is shown in Fig. 4.

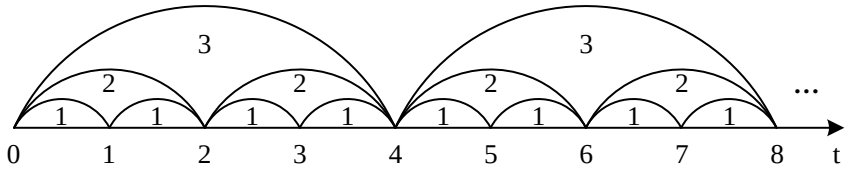

Figure 3: counter is the third-level processing at 8 o 'clock

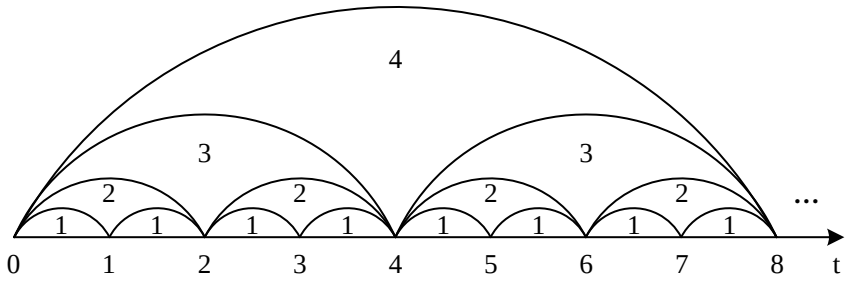

Figure 4: counter is the fourth layer processing at 8 o 'clock

2.1.6 Determination of beat point location

The beat point position is determined by the beat start point and beat value together, the beat start point is the average value of the first possible beat point position obtained by four wavelet transforms, according to the beat start point plus a number of times the beat value to all the beat point positions. If the current beat value is B_v and the beat start point is P_1, then the position of the nth beat point is equation (3):

$$P_n = P_1 + (n-1)B_v \tag{3}$$

2.2 A model for human voice color representation

Although existing timbre-related research can solve some musical instrument classification or instrument recognition problems relatively effectively, it neither addresses what are the essential features of timbre nor has the means to provide a metric mechanism for timbre similarity.

In terms of the characterization of musical instrument timbre, a nonlinear semantic embedding (NLSE) approach is used, where a convolutional neural network is introduced to embed the high-dimensional time-frequency representations of musical instrument samples into a low-dimensional, semantically organized metrizable space. Due to the introduction of a deep learning approach, NLSE is not only capable of dimensionality reduction of high-dimensional feature vectors, but also automatically learns salient acoustic features, merges extrinsic semantic features, and generates an intuitive timbre embedding space.

2.2.1 Deep learning model for human voice color embedding

The vocal color embedding model converts multi-frame high-dimensional CQT features into a low-dimensional vocal color embedding space by feeding them into a deep convolutional neural network. The deep convolutional neural network has five layers, and the specific structure is shown in Fig. 5.The first three layers of the deep convolutional neural network are convolutional layers. The convolutional neural network has translation invariance and can reduce the parameter space to achieve the effect of dimensionality reduction. The last two layers are fully connected affine transform layers and generate the embedding space from the last layer. The activation function of the first four layers uses hyperbolic tangent function, this is because hyperbolic tangent function has better numerical stability. The activation function of the last layer uses a linear function, this is because linear functions are more suitable for finding the appropriate scale of the embedding space.

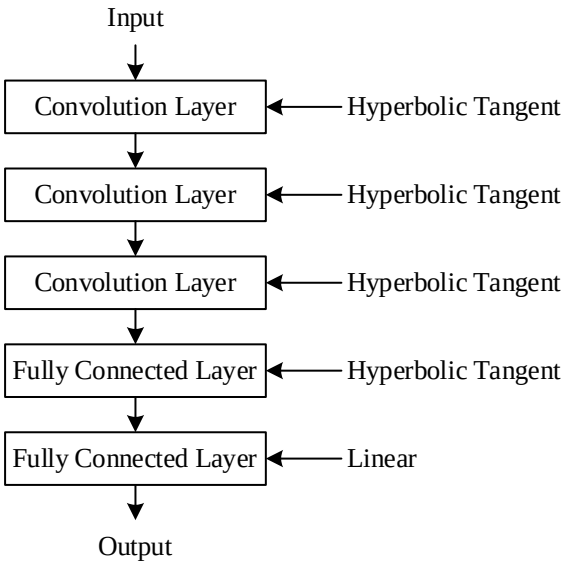

Figure 5: Deep convolutional neural network structure

The human voice color embedding space training model consists of two input data X_1, X_2, two symmetric neural networks G_A and G_B and the computation of loss function. The structure of the vocal color embedding training model is given in Fig. 6. The neural networks G_A and G_B have the same structure and share parameters.

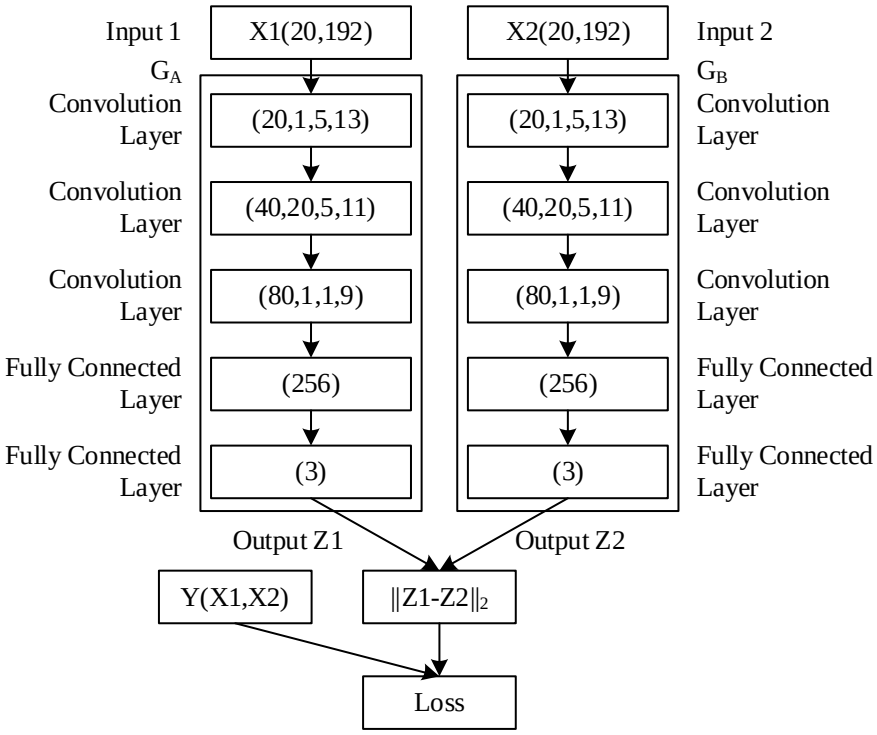

Figure 6: The human voice timbre is embedded in the training model structure

Since sound is a timing-dependent signal, it is taken into account that differences in the length of the input signal may have an effect on the timbre embedding. The number of convolution kernels and the size of convolution kernels in each layer of the neural network are different for different number of frames of input data. Taking the input frame number of 20 frames as an example, the specific settings of this deep convolutional network are as follows: the first layer (convolutional layer) uses 20 convolutional kernels with a size of $(1,5,13)$ and a max-pooling of $(2,2)$. The second layer (convolutional layer) uses 40 convolutional kernels of size $(20,5,11)$ and max-pooling of $(2,2)$. The third layer (convolutional layer) uses 80 convolutional kernels of size $(1,1,9)$ and max-pooling of $(2,2)$. The fourth layer (fully connected layer) has 256 output nodes. The fifth layer (fully connected layer) has 3 output nodes.

2.2.2 Input signals

Since human voice color is affected by many factors, such as changes in the position and morphology of the vocal organs and changes in the shape of the vocal tract. By analyzing the acoustic spectrogram, it can be seen that the changes of all the timbre-influencing factors are reflected in the acoustic spectrogram. Therefore, in this paper, the vocal signal is transformed into a spectral signal, which is used as the input signal of the vocal timbre embedded deep convolutional neural network. Since the CQT spectrum is a logarithmic spectrum, which is more in line with the human auditory system than the STFT spectrum, the CQT spectrum is used as the input for timbre characterization.

The relevant parameters of the CQT filter bank are as follows: all the audio is first downsampled to 16kHz, and b is taken to be 24, which is equivalent to taking the frequencies at quarter-tone intervals, with 24 frequency subscripts for each octave. The frequency range

contains 8 octaves, the minimum center frequency value $f_{\min}$ is set to 27.5Hz, and the maximum center frequency value $f_{\max}$ is set to 7040Hz, so that $K = 8 \times 24 = 192$, i.e., the CQT feature of each frame is 192 dimensions.

2.2.3 Pair training

The deep convolutional neural network for human voice color embedding used in this paper is a supervised learning model, and its learning process requires a pairwise training method. The pairwise training method trains the deep neural network by minimizing the distance between similar data and maximizing the distance between dissimilar data in the embedding space. In this paper, we use Euclidean distance to measure the similarity of timbre segments. In the trained timbre embedding space, sound segments with the same timbre should be clustered together, and sound frequency segments with greater timbre differences should be farther apart. Since there is not yet an accepted method for calculating the similarity of human voice timbre, it is assumed that the voice fragments of the same person have 100% similarity, and the voice fragments of different people have 0 similarity.

Firstly, the two input data X_1, X_2 trained in pairs are fed into two independent neural networks G_A and G_B, respectively, the two neural networks have the same structure and the same parameters, $W_{G_A} = W_{G_B}$, and the outputs of G_A and G_B are Z_1 and Z_2, respectively. There is equation (4):

$$\begin{cases} Z_1 = G_A(X_1 \mid W_{G_A}) \\ Z_2 = G_B(X_2 \mid W_{G_B}) \end{cases} \tag{4}$$

The cost function is defined as the Euclidean distance D between Z_1 and Z_2 as in equation (5):

$$D = \left\| Z_1 - Z_2 \right\|_2 \tag{5}$$

Depending on whether the two input data X_1, X_2 correspond to the same human voice label or not, different loss functions are set as in equation (6):

$$\begin{cases} L_{sim} = \max\left(0, D^2\right) \\ L_{dis} = \max\left(0, M_d - D\right)^2 \end{cases} \tag{6}$$

where D is the Euclidean distance between the two input data X_1, X_2 at the outputs Z_1, Z_2 of the vocal color embedding space. M_d is a constant related to the range of the vocal color embedding space, in this paper, $M_d = 10$.

The loss function L_{sim} when two input data X_1, X_2 correspond to the same vocal label, and the loss function L_{dis} when two input data X_1, X_2 correspond to different vocal labels.

The two different loss functions are merged as in equation (7) by the similarity scores of the human voice labels corresponding to the input data X_1, X_2:

$$Loss = V * L_{sim} + (1 - V) * L_{dis} \tag{7}$$

where the value of V is determined by equation (8):

$$V = \begin{cases} 1 & Label(X_1) = Label(X_2) \\ 0 & Label(X_1) \neq Label(X_2) \end{cases} \tag{8}$$

3 Teaching training of traditional teaching into computerized music software

For the current chorus encountered heavy and difficult problems are mostly concentrated in the melody, rhythm, voice, breathing and starting voice and so on, in the actual operation process is divided into the traditional piano rehearsal group (traditional group), the actual application of Cubase12 software combined with traditional rehearsal (software group).

3.1 Rhythm Training Music

Rhythm precedes melody in choral teaching, and skillful rhythm mastery can reduce the overall difficulty of music learning for students.

(1) Software

Use the Cubase software to change the strength of the notes and use color to differentiate. When the notes are traveling, the ruler passes through the MIDI block and emits matching strength and pitch, which visually and dynamically solves the problem of note strength and weakness for students' understanding.

(2) Traditional Group

Exactly the same score as the software group, with piano or drums as the main sounding instrument, and standardized notes played on the piano keyboard or drums; the tempo is set to the metronome tempo as the standard in the teaching process.

3.2 Vocalization training

(1) Software group

Use Cubase12 software to assist students to deal with song phrasing, breathing and other details, to help students form a basic vocal state, keep their voice stable, and form good singing vocal habits. For special melodies and rhythms, teachers can edit the notes in the editing interface. Conventional rhythms, on the other hand, can be adjusted to the selected notes in a batch of one-key strength, and can also be lifted and lowered at any time.

(2) Conventional group

The practice process of the piano or electric steel as the main vocal instrument, the speed of the metronome-based, students holding a simple score or five-line score to sing.

3.3 Two-part harmony training

The pitch and even the nature of the intervals are the parts of the two-part harmony training that students need to focus on.

(1) Software

With the Cubase 12 software, teachers can play a part individually, adjust the volume and speed according to the teaching progress, and also use the timbre of the software to replace any timbre in order to promote the regular teaching and enhance the students' understanding of the intervallic relationship of the double parts. And in the regular teaching, teachers can also focus on the students' pitch performance, according to the actual situation of the students to modify

and adjust the speed of the practice song, the volume size of the voice, and so on.

(2) Traditional Group

In the training process, the piano or electric piano is used as the main instrument, and the teacher performs sight-reading for each single voice part, and the students sing along. When the voices are combined, the teacher should focus on controlling the volume of the high and low voices, and adjust the teaching method according to the actual situation of the students.

3.4 Three-part harmony training

Polyphonic choral works have complex structural features, such as the unity of opposites between function and color in chord relationships, the stability and instability between the tuning range and chords, etc., all of which need to be taken into account in the work.

(1) Software group

Editing in MIDI form in Cubase 12 software, gradually adjusting the volume level or changing the instrumental timbre during practice, focusing the students' attention on their respective vocal parts and keeping them singing steadily. And you can adjust the volume level of the MIDI voice part when adding new voice parts, realizing the pitch cue for each voice part.

(2) Traditional Group

The traditional group will use the piano or electric piano as the main accompaniment, focusing on intonation and breathing. When the pitch of the root voice is stabilized, the line will be kept on the root voice, and then the sound will be sustained, and finally the pitch will be adjusted to keep the same voice.

3.5 Intonation of dissonant chords

Due to the acoustic “instability” of dissonant chords, it is necessary to strengthen the tendency of the voice to move and to pay attention to the stability of the intonation.

(1) Software

When practicing dissonant chords, it is possible to enhance the contrast of timbre by using the same timbre for the sustained notes and replacing the timbre for the altered notes, so as to differentiate the pitches. Utilize the Cubase 12 software to maintain volume while the voices are held. In polyphonic training, you can adjust the sound retention and volume retention at any time.

(2) Conventional group

Since dissonant chords are oriented and unstable, the difference in spitting can be added to make chord distinctions during pitch practice. During the singing process, the line keeps the tritone part stable, the bass part is slightly lowered, and the pentatonic part is slightly higher, so as to strengthen the chord color.

4 Exploration of Path Reconstruction of Choral Teaching Model

4.1 Performance of models and algorithms

4.1.1 Real-time music beat tracking effects

The public test samples provided by MIREX are selected, and 30 samples are randomly selected for the experiment. Each sample is 30s mono, with a sampling rate of 44.1kHz and a quantization accuracy of 16bit.Considering the real-time and low-consumption, we use (I1) complex frequency-domain difference, (I2) spectral energy, and (I3) high-frequency content of

the three trigger-point detection methods, and (I4) this paper's music beat-tracking method, to carry out the samples of the non-adaptive whitening, adaptive whitening beat detection experimental results are shown in Table 1. The P-scores of the four methods for the detection of samples after adaptive whitening are all improved to different degrees, among which (I4) the music beat tracking method in this paper not only has the highest P-score for the non-adaptive whitening samples (0.4564), but also improves the P-score of the samples after adaptive whitening the most (5.35%).

Table 1: Real-time beat tracking experimental results(P-score)

Method	Non-adaptive whitening	Adaptive whitening
I1	0.4551	0.4614
I2	0.4558	0.4601
I3	0.4537	0.4613
I4	0.4564	0.4808

Using the music beat tracking method in this paper for real-time beat tracking of a single music with large ups and downs, a total of 10 samples are selected, and the results of the non-adaptive whitening and adaptive whitening beat P-score for the samples are shown in Table 2.The real-time tracking effect of this paper's music beat tracking method for music with large ups and downs is in line with the real-time tracking effect of the overall samples, and the results for the adaptive whitening beat P-score can reach up to 0.4875.

Table 2: Real-time beat tracking experiment for music with large fluctuations(P-score)

Sample number	Non-adaptive whitening	Adaptive whitening
5	0.4586	0.4875
18	0.4554	0.4804
19	0.4617	0.4833
23	0.4621	0.4803
27	0.4605	0.4811
32	0.4581	0.4856
35	0.4645	0.4822
42	0.4609	0.4838
47	0.4587	0.4827
50	0.4606	0.4844

Using (I5) static WRR algorithm, (I6) DWPS algorithm as a control, the access request is initiated through the backend server, the initial access request volume is 500, and it is tested for 10 times with a frequency increment of 1000/s. The response time of the two control algorithms is shown in Fig. 7. The response times of the two control algorithms and (I4) music beat tracking method in this paper are shown in Fig. 7. In the case of a small number of accesses, the difference between the algorithms is not obvious. When the number of requests rises gradually, the response time of the three algorithms reaches the highest point at 3500, of which the response time of (I4)this paper's music beat tracking method is the shortest, which is only 2423 s. And the response time decreases the fastest after exceeding 3500 requests, and when the number of requests is 5000, the response time is only 2000s.

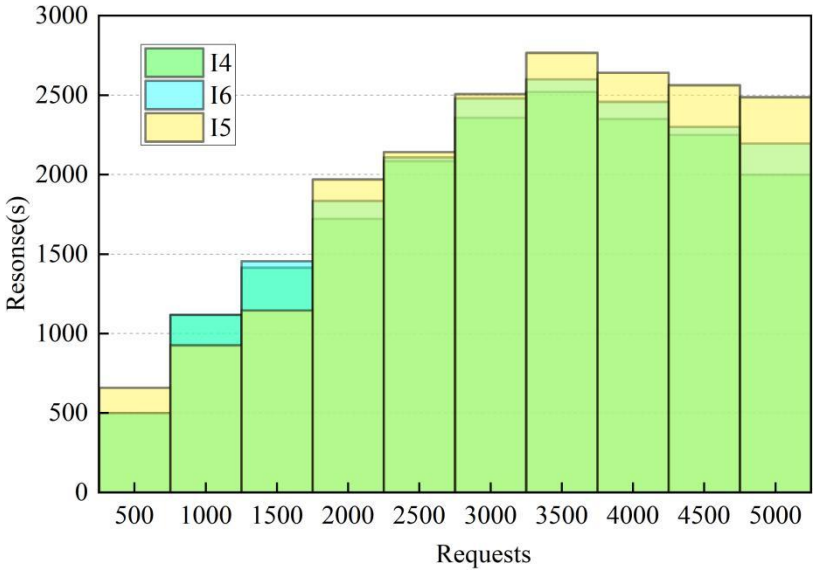

Figure 7: The response times of the three algorithms

4.1.2 Learning effect of transferring human voice color features

The dataset containing 10 categories (numbered J1-J10) of vocal singing was selected as the experimental training set. The confusion matrix of the classification accuracy of the timbre feature Fea1 extracted on the experimental training set using the (K1) vocal color representation model is shown in Fig. 8, and the confusion matrix of the classification accuracy of the timbre feature Fea2 extracted using the same type of (K2) HTim-DCNN model is shown in Fig. 9. Overall the classification accuracies of the 10 human voice features extracted by the two models are higher, but the (K1) vocal color representation model is slightly better than the (K2) HTim-DCNN model. (K2)HTim-DCNN model performance, its classification accuracy is in the interval of (0.6,0.85), while the highest classification accuracy of (K2)HTim-DCNN model is only 0.70.

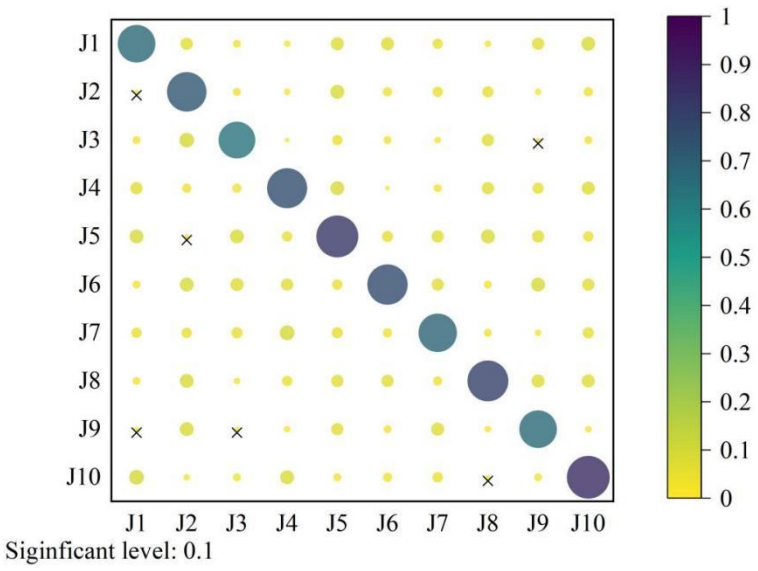

Figure 8: The K1 model timbre feature Fea1 classification accuracy confusion matrix

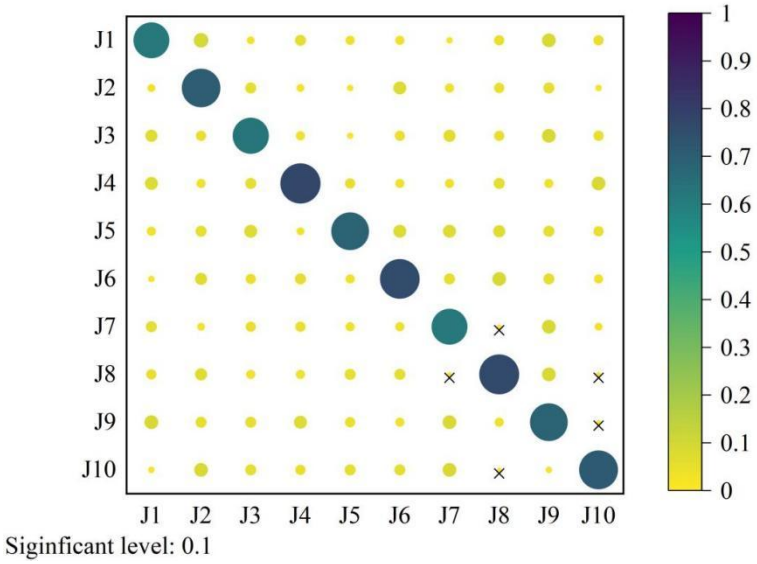

Figure 9: The K2 model timbre feature Fea2 classification accuracy confusion matrix

To further demonstrate the vocal color representation extraction and classification ability of the vocal color representation model, the 10 categories of vocal singing were further subdivided into 15 categories (numbered J1-J15), and the classification accuracy confusion matrix of the vocal color embedded deep learning model on the experimental training set is shown in Figure 10. Due to the more adequate amount of categorized data, the human voice color representation model shows a better and more stable performance compared to the 10-category extraction classification experiment. Its overall classification error is <0.01, and its classification accuracy is stable in the (0.7,0.85) interval.

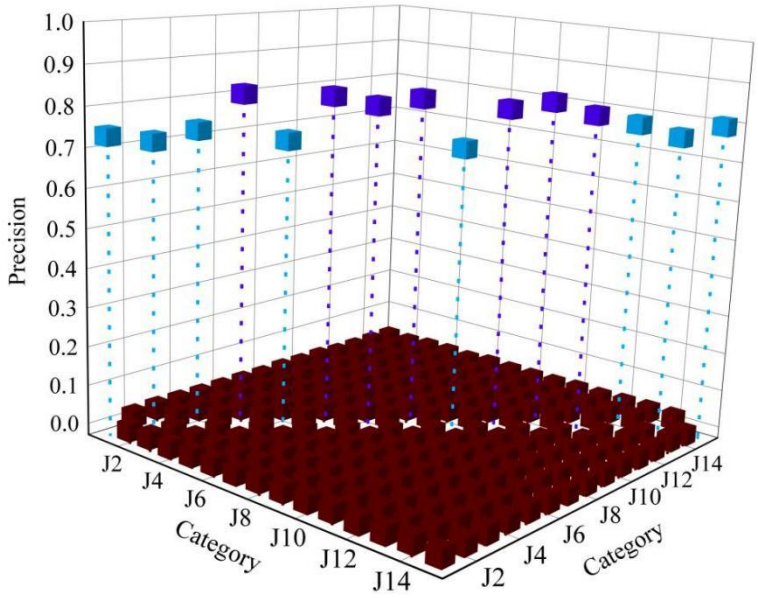

Figure 10: Textual model extracts classification confusion matrices for 15 categories

4.2 Traditional Choral Teaching Mode in the Intelligent Era

Two classes in the choral department of a performing school's vocal music program were selected as experimental subjects, with 20 students in each of the two classes. The class as a unit set (L1) experimental group and (L2) control group, of which (L1) experimental group of

students' teachers need to accept the traditional teaching fusion of computer music software teaching training, auxiliary teaching tools for real-time music beat tracking algorithms, the human voice color representation model, the method used in the following analysis referred to as this paper's method. (L2) The control group was trained to sing music using traditional teaching methods.

4.2.1 Tone range performance

The post-experimental changes in the registers of the students in the (L1) experimental group and the (L2) control group are shown in Fig. 11, numbered according to their degree of register enhancement on a scale of 1-20. With the aid of the methodology in this paper, only one student in the (L1) experimental group had the same range status as before the experiment, while the remaining 19 students each had 1-6 different semitone widening changes. In contrast, nearly half of the students in the (L2) control group had mostly unchanged range states, and five students had narrowed range states.

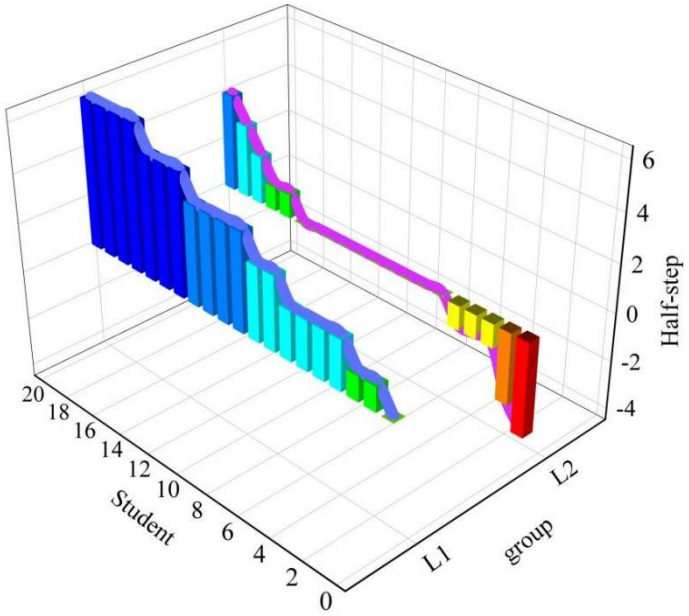

Figure 11: The changes in the students' vocal range after the experiment

4.2.2 Performance of acoustic parameters

Using the vowel /a/ and the vowel /i/ as test objects, the maximum vocalization time (MPT), average sound pressure level (AvgSPL) and harmonization ratio (Hnr) of the students in the post-experimental (L1) experimental group and the (L2) control group are compared in Fig. 12, and all three parameter indicators are in a certain range of the higher the better. Students in the (L1) experimental group were significantly better than students in the (L2) control group in all three vocalization parameters for the vowel /a/ and the vowel /i/, with a difference of up to 6.68 between the average sound pressure level (AvgSPL) for the vowel /a/ and that of students in the (L2) control group.

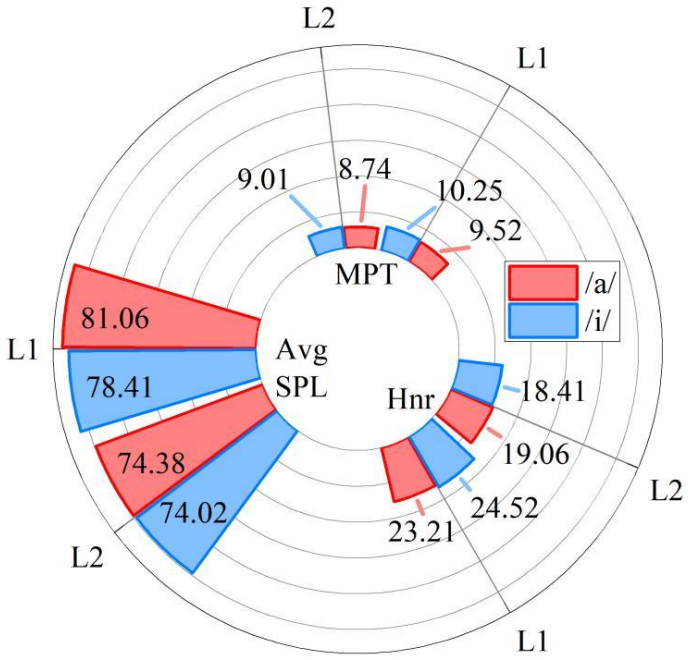

Figure 12: Comparison of the performance of MPT, AvgSPL and Hnr

The fundamental frequency perturbation (jitter), amplitude perturbation (shimmer), hoarseness index (Axqi), and breathiness index (Abi) of the students in the (L1) experimental group and the (L2) control group on the vowel /a/ and the vowel /i/ after combing experiments are shown in Fig. 13, and all four acoustic parameters are smaller and better within a certain range. The (L1) experimental group students still maintained a much better performance than the (L2) control group students in all four parameters, with the hoarseness index (Axqi) of the vowel /a/ and the vowel /i/ being only 0.01 and 0.03. The (L2) control group students not only had higher values for all four parameters, but also had amplitude perturbation (shimmer) (<3.08) and breathiness index (Abi) (<3.44) that exceeded the limits of the (L2) experimental group students' performance in all four parameters. 3.44) were outside the normal range.

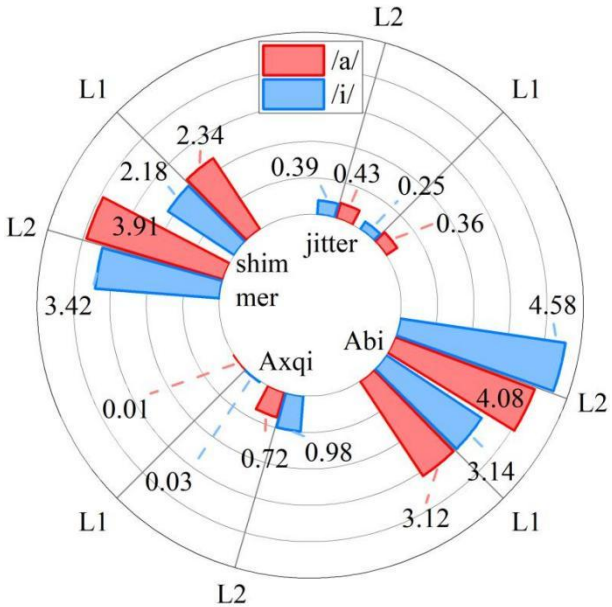

Figure 13: Comparison of the performance of jitter, shimmer, Axqi and Abi

Comprehensively analyzing the above, the method of this paper not only has obvious improvement effects on students' voices, but also can assist in the effective reduction of students' hoarseness, breath and roughness, as well as the significant improvement of voice loudness, and promote the enhancement of students' overall speech sound effects, with reliable practical application effects. That is, the optimization and improvement of teaching choral repertoire to students through real-time music beat tracking, and guiding students' choral vocalization through the human voice color representation model are feasible development paths for the traditional choral teaching mode in the intelligent era.

5 Conclusion

This paper proposes a real-time music beat tracking algorithm based on wavelet transform that can detect the beat value and the specific location of the beat point in real time, and constructs a human voice color representation model based on deep convolutional neural network. The real-time music beat tracking algorithm can reach a maximum of 0.4875 for music beat P-score, and the response time for 3000 accesses is only 2423s, which is better than similar algorithms. The vocal color representation model, on the other hand, has a classification error of <0.01 for 15 categories of vocal singing, and the classification accuracy is stable in the interval of (0.7,0.85). Based on the real-time music beat tracking algorithm and the vocal color representation model to assist the traditional choral teaching, the students performed much better than the control students in nine acoustic parameters, except for the widening of the range of 1-6 different semitones, and the difference between the average sound pressure level (AvgSPL) of the vowel /a/ and that of the control students was as high as 6.68, and the hissiness index of the vowels /a/ and /i/ was only 0.01 (Axqi), which was 0.01 (0.7,0.85). Axqi) was only 0.01 and 0.03. Empowering traditional choral teaching through intelligent era technological tools and software is an effective exploration of reconfiguring the traditional choral teaching path.

About the Author

Ying Zhou was born in Shangrao, Jiangxi, China, in 1980. She received her master's degree from South China Normal University in Guangzhou and is currently a faculty member at the School of Music, Zhaoqing University. Her primary research focuses on choral conducting and choral music education.

Yuwei Lu was born in Yongding, Fujian, China, in 1981. He received his bachelor's degree from Nanchang Hangkong University and is currently employed at a cigarette factory in Guangzhou. His primary research interests lie in data analysis and the application of multimedia technology.

References

[1] Yuan, S. (2020, April). Application and study of musical artificial intelligence in music education field. In Journal of physics: Conference series (Vol. 1533, No. 3, p. 032033). IOP Publishing.

[2] Rizal, H. & Milyartini, R. (2024). Improving the quality of music education through applications based on Artificial Intelligence (AI). In SHS Web of Conferences (Vol. 197, p. 01004). EDP Sciences.

[3] Yang, W. Shen, L. Huang, C. F., Lee, J., & Zhao, X. (2024). Development status, frontier hotspots, and technical evaluations in the field of AI music composition since the 21st century: a systematic review. IEEE Access, 12, 89452-89466.

[4] Li, N. & Wu, D. (2025). The Auxiliary Function and Realization Mechanism of Artificial Intelligence in Cross-Cultural Traditional Music Education. Journal of Cases on Information Technology (JCIT), 27(1), 1-18.

[5] Freer, P. K. (2011). The performance-pedagogy paradox in choral music teaching. Philosophy of Music Education Review, 19(2), 164-178.

[6] Gumm, A. (2004). The effect of choral student learning style and motivation for music on perception of music teaching style. Bulletin of the council for research in music education, 11-22.

[7] Debrot, R. A. (2017). Incorporating popular music and dance: A student-centred approach to middle school chorus. Journal of Popular Music Education, 1(3), 297-316.

[8] Halvasi, B. (2018). Basic Properties of Chorus and Fundamental Approaches to Improve. Educational Policy Analysis and Strategic Research, 13(1), 113-126.

[9] Frizzell, E. Y. & Windsor, L. C. (2021). Effects of teaching experience and culture on choral directors' descriptions of choral tone. Plos one, 16(12), e0256587.

[10] Huang, D. (2020). Research on the Innovative Teaching Method of Chorus Conducting in Colleges and Universities. Frontiers in Art Research, 2020, 2 (9).

[11] Zhou, Z. (2023). Innovative learning environments for choral conducting education. Education and Information Technologies, 28(7), 7827-7843.

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

[13] Zuojun, W. Pattananon, N. & Leangsomboon, W. (2024). Chorus course teaching method in universities in China. Journal of Modern Learning Development, 9(4), 540-548.

[14] Zhang, L. (2025). Compositional tools based on artificial intelligence for choral artistic education: Enhancing creative skills in choral arrangements. Thinking Skills and Creativity, 56, 101768.

[15] Xiaofang, L. Pattananon, N. & Saengthong, T. (2023). The Teaching of Chorus Courses in Chinese Higher Education. Journal of Modern Learning Development, 8(4), 264-272.

[16] Holland, S. (2013). Artificial intelligence in music education: A critical review. Readings in music and artificial intelligence, 239-274.

[17] Wei, J.Karuppiah, M. & Prathik, A. (2022). College music education and teaching based on AI techniques. Computers and Electrical Engineering, 100, 107851.

[18] Wu, Q. (2025). The application of artificial intelligence in music education management: Opportunities and challenges. Journal of Computational Methods in Sciences and

Engineering, 25(3), 2836-2848.

[19] Liu, P. (2025). Modern AI program Chinese choral arts: cognitive training and motivation of college choristers. Interactive Learning Environments, 1-15.

[20] Wu, X. (2025). Singing Syllabi with Virtual Avatars: Enhancing Student Engagement Through AI-Generated Music and Digital Embodiment. arXiv preprint arXiv:2508.11872.

[21] Zheng, H. & Dai, D. (2022). Construction and Optimization of Artificial Intelligence-Assisted Interactive College Music Performance Teaching System. Scientific programming, 2022(1), 3199860.