International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5832-5846

https://doi.org/10.70517/ijhsa463502

# Design of a personalized teaching system for vocal music based on data fusion algorithm

**Jialu Qin**[1,*]
[1] College of Education, Hubei Business College, Wuhan, Hubei, 123456, China
Corresponding authors: (e-mail: qinjialu2025@163.com).

**Abstract** With the rapid development of educational informatization, traditional vocal music teaching faces problems such as lack of personalized teaching resources and single learning evaluation. Through data fusion technology, creating a personalized teaching system for vocal music that integrates audio and text features can effectively improve the teaching effect and learning experience. In this study, the MFCC method is used to extract audio features, the TF-IDF function is used to extract text features, and the EFFC multimodal fusion algorithm based on HMM constraints is designed to realize the effective fusion of the two modalities. The experimental results show that the accuracy of MFCC audio feature extraction reaches 0.972, which is significantly better than other feature extraction methods; the accuracy of HMM-EFFC fusion algorithm is 0.9635, and the F1 value reaches 0.9676, which is better than the comparative algorithm; in the system application test, the accuracy of the judgment of learning interest reaches 98.37%, and the CPU occupancy rate is only 27.53%. The study proves that the personalized teaching system for vocal music based on data fusion algorithm can effectively improve the teaching effect and learning experience, and provides a new idea for the innovation of vocal music teaching.

**Index Terms** Data fusion algorithm, vocal music teaching, personalized teaching, MFCC feature extraction, HMM constraint, EFFC fusion algorithm

## I. Introduction

Vocal music teaching is characterized by the transmission of skills, and it is a teaching aiming at cultivating students' mastery of certain singing skills and singing ability. In recent years, due to the expansion of enrollment scale, the number of students majoring in vocal music has been increasing, resulting in a large individual difference in the student population, with varying degrees of professional ability, and the expansion of the student population has also led to a serious shortage of teachers of vocal music [1], [2]. In order to solve the problem of teacher shortage, institutions have set up group lessons, large classes, diversified forms of teaching, but the individual small classes that can best reflect the characteristics of vocal music teaching, “tailored to the needs of the students”, are less and less. At the same time, the teaching process of vocal music is basically teacher-driven, teachers are overly concerned about the core indicators of vocal music, such as pitch and diction, and empirical teaching is dominant, which are not fully consistent with the basic characteristics and laws of vocal music teaching, and also limit the quality of vocal music talent training [3]-[7].

From the students' point of view, although everyone's vocal organs have something in common, the fact is that each student's differences in physiology, gender, and psychology exist objectively, such as the length of the vocal cords, thickness and other subtle differences that determine each person's different voice color. In general, boys' vocal cords are longer, wider and thicker in pitch, while girls' vocal cords are shorter, thinner and narrower in pitch, and psychological quality affects the quality of vocal performance [8]-[10]. Therefore, according to the individual differences of students, following the educational concept of “people-oriented”, how to implement individualized teaching reform in the vocal music course, establish a relatively scientific evaluation system, help students really feel their own progress and improvement in the learning process, stimulate students' interest and enthusiasm in singing and performing, and contribute to the formation and development of good emotional attitudes of students. The formation and development of good emotional attitudes is of great practical significance for cultivating the professional skills of music majors and improving their professional core competitiveness [11]-[14].

At present, with the rapid development of Internet technology and artificial intelligence, education informatization has become an important direction of education reform. In the field of vocal music teaching, the traditional teaching mode is difficult to meet the needs of students' personalized development, which is mainly manifested in the lack of rich teaching resources, a single learning evaluation system, and a lack of pertinence in teaching methods. At the same time, the rapid growth of a large amount of audio data and text data provides a data basis for vocal

personalized teaching, and it becomes possible to build an intelligent vocal personalized teaching system by using advanced feature extraction and multimodal data fusion technology. Currently, audio feature extraction in vocal music teaching mainly uses linear prediction coefficients (LPC), energy features and chroma features (Chroma), but there are still limitations in feature expressiveness and robustness; Text feature extraction is commonly used in Word2Vec and other methods, but the effect is not ideal in short text processing; multimodal data fusion is commonly based on CNN, RNN, SVM and other methods, but there is still room for improvement in fusion efficiency and accuracy. Therefore, it is of great significance to study and design a personalized teaching system for vocal music based on multimodal data fusion algorithm. In this study, the MFCC method is firstly used for audio feature extraction, which can very accurately reflect the features of audio in multiple dimensions, including preprocessing, fast Fourier transform, Mel filter bank processing, discrete cosine transform and differential cepstrum coefficient computation; Then the TF-IDF function is used for text feature extraction, and the digital conversion of text information is realized by calculating the word frequency (TF) and the inverse text frequency (IDF); then the EFFC multimodal fusion algorithm based on HMM constraints is designed, and the effective fusion of the two modalities, audio and text, is realized by the three steps of feature cascade, feature degradation and matrix fusion; Finally based on Django web development framework, using MVC architecture model and MySQL database, combined with front-end technology (HTML5, JavaScript) and back-end management framework (Xadmin) for system implementation. The system is mainly divided into six functional modules: registration module, login module, backstage management module, multimodal data acquisition module, vocal course resource recommendation module and course evaluation module. In order to verify the system performance, this study selects 10 pieces of music with different beat types for audio feature extraction experiments and compares different feature extraction methods; 400 text feature words are selected for text feature extraction experiments to compare the effects of different dimensional features; 254 vocal music teaching videos are selected for system testing to analyze the performance of the system in terms of judgment of learning interest, CPU occupancy, evaluation of learning experience, efficiency of listening and acceptance. The results of the study are of reference value for improving the effect of vocal music teaching, promoting education informatization and personalized development.

## II. Vocal Music Teaching System Combined with Data Fusion Algorithm

### *II. A.MFCC-based audio feature extraction*

Because MFCC can accurately reflect the characteristics of audio in many dimensions, its research in music feature extraction and similarity has made great progress and become one of the most important features in the field of audio analysis [15]. Here, the correspondence between the Mel frequency and the actual frequency is given first:

$$Mel(f) = 2595\lg\left(1+\frac{f}{700}\right) \tag{1}$$

(1) Pre-processing

Before the extraction of the audio signal, a series of preliminary work needs to be performed on the audio signal, called audio signal preprocessing. The formula for the first order high pass filter is given here:

$$H(z) = 1-\lambda z^{-1} \tag{2}$$

where $\lambda$ represents the pre-emphasis coefficient, taking a value between 0.9 and 1.0.

In order to maintain the frames in a continuous state so as to ensure the integrity of the audio signal without distortion, it is necessary to carry out windowing and frame splitting processing. Here, the Hamming window is selected, and the formula is as follows:

$$x(l) = \sum_{-\infty}^{+\infty} T[y(i)] * w(l-i) \tag{3}$$

Among them:

$$w(l) = \begin{cases} 0.54-0.46\cos\left(2\pi l / L-1\right), & 0 \le l \le L-1 \\ 0£\neg, & Other \end{cases} \tag{4}$$

(2) Fast Fourier Transform

After preprocessing each frame of the signal needs to do the Fast Fourier Transform (FFT), through the FFT, the signal is converted to the frequency domain, at this time the distribution of the energy changes in the frequency spectrum will become more and more clear. The FFT of the transformation formula is as follows:

$$X(j)=\sum_{l=0}^{L-1}x(l)W^{lj},\ j=0,1,2,\cdots,L-1 \tag{5}$$

(3) Mel filter bank

The Meier filter bank consists of $M$ triangular bandpass filters with the transfer function equation shown below:

$$H_i(j)=\begin{cases}\dfrac{f(i+1)-j}{f(i+1)-f(i)}, & f(i)<j<f(i+1)\\ \dfrac{j-f(i-1)}{f(i)-f(i-1)}, & f(i-1)<j<f(i)\\ 0, & \text{Other}\end{cases} \tag{6}$$

The effect of harmonics can be minimized by passing the filter banks through the Meier filter bank. The distribution of the Mel filter bank is given in Fig. 1. Also, the output signal needs to be converted into logarithmic form to enhance the robustness of the spectral estimation as shown in the following equation:

$$Y(i)=\ln\left[\sum_{k=0}^{L-1}\left|X(j)\right|^2H_i(j)\right],\ 0\le i\le M \tag{7}$$

(4) Discrete Cosine Transform

Discrete Cosine Transform (DCT) is performed on $Y(i)$ and after passing through the DCT, the signal is transformed to the cepstrum domain. The formula is shown below:

$$C_i=\sum_{i=1}^{M-1}Y(i)\cos\left(\frac{l\pi(i+0.5)}{M}\right),\ 0\le i<M \tag{8}$$

(5) Differential cepstrum coefficients

In order to make the description of the speech signal can become more dynamic, a linear prediction operation is required. The coefficients obtained from the linear prediction can be further derived from the cepstrum coefficients to calculate the differential cepstrum coefficients. The formula is shown below:

$$\Delta D_i(t)=\frac{\sum_{\eta=-i}^{M}D_j(t+\eta)}{\sum_{\eta=-i}^{M}\eta^2} \tag{9}$$

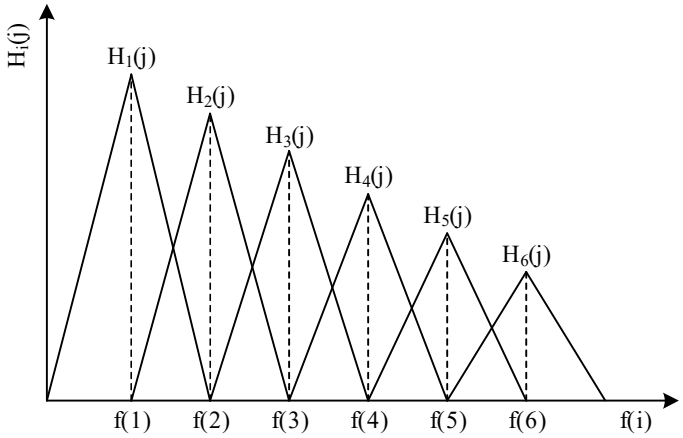



Figure 1: MEL filter bank distribution

### *II. B.Text Feature Extraction*

#### II. B. 1) Text Feature Extraction Overview

For the lyrics of a piece of music, as a kind of text form, it makes the computer can't process it directly, here we need to borrow the TF-IDF function to digitize and convert the lyrics information. For a document containing $n$ feature words, the TF-IDF function can represent the document $d$ as a vector of $d=d\left(t_1,w_1;t_2,w_2;\cdots;t_n,w_n\right)$, where $t_i$ represents the feature word, and $w_i$ represents the weight of the feature word $t_i$ in document $d$, which is used to indicate the importance of the feature word $t_i$ in document $d$, and is usually also abbreviated as $d=d\left(w_1,w_2,\cdots,w_n\right)$.

#### II. B. 2) TF-IDF function

(1) Word Frequency TF

Word frequency TF, also known as in-text frequency, i.e., the frequency of occurrence of a feature word within a document, is an index for evaluating the importance of a particular feature word for that document [16]. Its formula is as follows:

$$TF=\frac{N_{ij}}{N_{*j}} \tag{10}$$

where $N_{ij}$ denotes the number of times the feature word $t_i$ appears in the document $d_j$ and $N_{*j}$ denotes the number of all feature words in the document $d_j$.

(2) Inverse Text Frequency IDF

The inverse text frequency IDF reacts to the frequency of a feature word appearing in all documents, if the number of documents containing the feature word is less, the larger the value of IDF is, then it means that the feature word has a good ability to distinguish. The formula is as follows:

$$IDF=\log\left(\frac{D}{D_i+1}\right) \tag{11}$$

where $D$ denotes the total number of documents, $D_i$ denotes the number of documents containing the feature term $t_i$, and the denominator is treated as plus one to avoid the denominator being zero.

Since the TF-IDF function combines the characteristics of TF and IDF to indicate the importance of the feature terms, the formula of the TF-IDF function is given here as follows:

$$TF-IDF_{i,j}=\frac{N_{ij}}{N_{*j}}\times\log\left(\frac{D}{D_i+1}\right) \tag{12}$$

TF-IDF indicates that the document is realized by VSM. For document $D=D\left(t_1,w_1;t_2,w_2;\cdots;t_n,w_n\right)$, it has been mentioned above that $w_i$ stands for the weights of the feature word $t_i$ in the document $D$, and if here $w_i$ uses the TF-IDF function, then a lyrics document can be represented as a vector by the TF-IDF function, which realizes the extraction of lyrics features.

### *II. C.Multimodal fusion algorithm*

#### II. C. 1) HMM

A Hidden Markov Model (HMM) describes the process of randomly generating a random sequence of unobservable states from a hidden Markov chain, and then generating a random sequence of observations by generating an observation from each state [17]-[19]. For the EFFC multimodal fusion method based on HMM constraints proposed in this paper, the HMM is required to provide the ratio of the two modes accounted for:

$$\frac{p\left(Like\middle|Audio\right)}{p\left(Like\middle|Lyrics\right)} \tag{13}$$

It is equivalent to finding two different $P\left(O\middle|\lambda\right)$, which corresponds to the parameter learning problem in HMM, assuming that the given data contains only $S$ sequences of observations $\left\{O_1,O_2,\cdots,O_S\right\}$ of length $T$, to estimate

the parameters of the Hidden Markov Model $\lambda=(A,B,\pi)$, the HMM can be expressed as if the state sequences are considered as unobservable hidden data $I$:

$$P(O\mid\lambda)=\sum_{I}P(O\mid I,\lambda)P(I\mid\lambda) \tag{14}$$

The first step is to determine the log-likelihood function for the complete data

All observed data are written as $O=(o_1,o_2,\cdots,o_T)$, all hidden data are written as $I=(i_1,i_2,\cdots,i_T)$, and the complete data are $(O,I)=o_1o_2\cdots o_Ti_1i_2\cdots i_T$. The log-likelihood function for complete data is $\log[P(O,I\mid\lambda)]$.

Step 2, the $E$ step of the EM algorithm: find the $Q$ function $Q(\lambda,\bar{\lambda})$:

$$Q(\lambda,\bar{\lambda})=\sum_{I}\log\left[P(O,I\mid\lambda)P(O,I\mid\bar{\lambda})\right] \tag{15}$$

where $\bar{\lambda}$ is the current estimate of the HMM parameter and $\lambda$ is the HMM parameter to be maximized. I.e:

$$P(O,I\mid\lambda)=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T) \tag{16}$$

Thus function $Q(\lambda,\bar{\lambda})$ can be written as:

$$\begin{aligned}Q(\lambda,\bar{\lambda})=&\sum_{I}\log\left[\pi_{i_1}P(O,I\mid\bar{\lambda})\right]\\&+\sum_{I}\left(\sum_{t=1}^{T-1}\log[a_{i_ti_{t+1}}]\right)P(O,I\mid\bar{\lambda})\\&+\sum_{I}\left(\sum_{t=1}^{T}\log\left[b_{i_t}(o_t)\right]\right)P(O,I\mid\bar{\lambda})\end{aligned} \tag{17}$$

Step 3, the $M$ step of the EM algorithm: maximize the $Q$ function $Q(\lambda,\bar{\lambda})$ to find the model $A,B,\pi$.

Since the parameters to be maximized appear individually in three terms in Eq. (17), it is only necessary to maximize each term separately. The first term in Eq. (17) can be written as:

$$\sum_{I}\log\left[\pi_{i_1}P(O,I\mid\bar{\lambda})\right]=\sum_{i=1}^{N}\log\left[\pi_iP(O,i_1=i\mid\bar{\lambda})\right] \tag{18}$$

where $\pi_i$ satisfies the constraint $\sum_{i=1}^{N}\pi_i=1$, and the Lagrange multiplier method is used to take the partial derivatives of $\pi_i$ and make the result zero:

$$\frac{\partial}{\partial\pi_i}\left\{\sum_{i=1}^{N}\log\left[\pi_iP(O,i_1=i\mid\bar{\lambda})\right]+\gamma\left(\sum_{i=1}^{N}\pi_i-1\right)\right\}=0 \tag{19}$$

Gotta:

$$P(O,i_1=i\mid\bar{\lambda})+\gamma\pi_i=0 \tag{20}$$

Summing over $i$ gives $\gamma$:

$$\gamma=-P(O\mid\bar{\lambda}) \tag{21}$$

The result is obtained by substituting into equation (18):

$$\pi_i=\frac{P(O,i_1=i\mid\bar{\lambda})}{P(O\mid\bar{\lambda})} \tag{22}$$

The 2nd term in Eq. (17) can be written as:

$$\sum_{I}\left(\sum_{t=1}^{T-1}\log[a_{i_t i_{t+1}}]\right)P\left(O,I\mid\bar{\lambda}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1}\log\left[a_{ij}P\left(O,i_t=i,i_{t+1}=j\mid\bar{\lambda}\right)\right] \tag{23}$$

Similarly to item 1, it can be derived by applying the Lagrange multiplier method with constraint $\sum_{j=1}^{N}a_{ij}=1$:

$$a_{ij}=\frac{\sum_{t=1}^{T-1}P\left(O,i_t=i,i_{t+1}=j\mid\bar{\lambda}\right)}{\sum_{t=1}^{T-1}P\left(O,i_t=i\mid\bar{\lambda}\right)} \tag{24}$$

The 3rd term in Eq. (17) is:

$$\sum_{I}\left(\sum_{t=1}^{T}\log\left[b_{i_t}(o_t)\right]\right)P\left(O,I\mid\bar{\lambda}\right)=\sum_{j=1}^{N}\sum_{t=1}^{T}\log\left[b_j(o_t)P\left(O,i_t=j\mid\bar{\lambda}\right)\right] \tag{25}$$

Again using the Lagrange multiplier method with a constraint of $\sum_{k=1}^{M}b_j(k)=1$. Derived:

$$b_j(k)=\frac{\sum_{t=1}^{T}P\left(O,i_t=j\mid\bar{\lambda}\right)I\left(o_t=v_k\right)}{\sum_{t=1}^{T}P\left(O,i_t=j\mid\bar{\lambda}\right)} \tag{26}$$

By expressing each probability in Eq. (24)~Eq. (26) in terms of $\gamma_t(i)$ and $\xi_t(i,j)$ respectively, the corresponding equations can be expressed as:

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{27}$$

$$b_j(k)=\frac{\sum_{i=1,a_i-v_i}^{T}\gamma_i(j)}{\sum_{i=1}^{T}\gamma_i(j)} \tag{28}$$

$$\pi_i=\gamma_1(i) \tag{29}$$

Among them:

$$\gamma_t(i)=P\left(i_t=q_i\mid O,\lambda\right)=\frac{P\left(i_t=q_i,O\mid\lambda\right)}{P(O\mid\lambda)} \tag{30}$$

$$\begin{aligned}\xi_t(i,j)&=\frac{P\left(i_t=q_i,i_{t+1}=q_j,O\mid\lambda\right)}{P(O\mid\lambda)}\\&=\frac{P\left(i_t=q_i,i_{t+1}=q_j,O\mid\lambda\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N}P\left(i_t=q_i,i_{t+1}=q_j,O\mid\lambda\right)}\end{aligned} \tag{31}$$

For the ratios mentioned above as needing to be found, it is simply a matter of finding the $B$ matrix by $b_j(k)$, which is equivalent to finding two different $P(O\mid\lambda)$, which can be derived:

$$\frac{p\left(Like\middle|Audio\right)}{p\left(Like\middle|Lyrics\right)} \tag{32}$$

**II. C. 2) EFFC multimodal fusion method based on HMM constraints**

(1) Feature cascade. Cascade the $m\times n$-dimensional matrix $A$ proportionally to the $m\times(n\times d)$-dimensional matrix $A'$, multiply the $A'$-matrix with the mapping matrix (IIO) to get a $m\times(n\times d\times 2)$-dimensional matrix $A''$, and intercept the first $q$ columns of this matrix to get a $m\times q$-dimensional matrix (AIO), where the size of the $q$ can be set by yourself but need to ensure $q\times(n\times d\times 2)$.

(2) Feature dimensionality reduction. Reduce the $m\times p$-dimensional matrix $B$ to $m\times\left[\frac{p}{d}\right]$-dimensional matrix $B'$ by Principal Component Analysis (PCA), and multiply the $B'$ matrix with the mapping matrix (OII) to get a $m\times\left[\frac{p}{d}\times 2\right]$-dimensional matrix $B''$, which is supplemented with all 0 elements to obtain a $m\times q$-dimensional matrix (OIB).

(3) Matrix fusion. Merge the $m\times q$-dimensional matrix (AIO) with the $m\times q$-dimensional matrix (OIB), and finally obtain a $m\times q$-dimensional song-multimodal feature matrix containing two musical features.

### *II. D.Individualized Vocal Music Teaching System Design*

The vocal personalization system based on multimodal data fusion algorithm proposed in this paper makes use of two different modalities, audio features and lyrics features of a song, in the hope of solving the two problems mentioned above through the multimodal features of the song. In order to better combine the two features, a new multimodal fusion method is similarly proposed here. In this section, the overall framework of vocal personalization system based on multimodal data fusion algorithm is described, which includes requirement analysis, outline design, and system implementation. The design flow of the vocal personalization system is as follows:

**II. D. 1) Demand analysis**

This paper based on data fusion algorithm of vocal music personalized teaching system is to use software engineering development ideas, in turn, requirements analysis, outline design, system implementation, in which the requirements analysis is the first and essential part of the whole system development process, which is mainly divided into two main categories. Functional requirements analysis and non-functional requirements analysis. First of all, clarify the function of the system, as well as specific functional modules, and then conduct feasibility analysis for each module, and finally confirm its feasibility. From the demand analysis of the personalized teaching system for vocal music, the system users are mainly divided into teachers and students.

(1) Functional Requirements Analysis

The main purpose of the functional demand analysis is to divide the functional modules of the vocal music personalized teaching system, and then to meet the demand of the online part of the vocal music personalized teaching in the blended teaching, firstly to analyze the traditional classroom teaching and the existing network education, to determine the main functions that should be in line with the personalized teaching system oriented to online learning in the current educational background, and to fully respond to the call for educational The call for teaching reform is fully responded to. According to the demand analysis, the vocal personalized teaching system is divided into the following six modules: registration module, login module, backstage management module, multimodal data collection module, vocal course resource recommendation module, course evaluation module, and each module corresponds to different sub-modules. Functional analysis is conducted for each functional module in turn, using use case diagrams to illustrate the permissions of the two roles of teachers and students, and use case statute description tables to introduce the interaction between the system and users in each module.

(2) Non-functional Requirements Analysis

Nowadays, each school is scrambling to establish its own network teaching platform, the endless emergence of major network platforms make users dazzled, no choice, while the user demand for online personalized teaching is also increasing, single face problems should not be ignored: expensive costs, user information leakage, system data is not centralized, resources can not be completely shared, system stability is not strong. Therefore, to develop an online personalized learning system that is really suitable for students, non-functional requirements should be considered:

(a) Technical feasibility of the system: the vocal personalized teaching system mainly adopts Django web architecture, uses python language for programming, and adopts Xadmin classic front-end backend management framework.

(b) system security: vocal personalized teaching system needle can ensure the security of the system, customer data, its working principle is for different types of user knowledge according to their common browsing habits, user behavior characteristics, etc. to determine the user's identity, the use of the system's unique information processing environment to establish a specific user information barriers, to protect the security of user's personal information, the use of a specific, complex, standardized code structures The platform is built to stop hackers from illegally intervening and prevent lawbreakers from stealing user account information.

(c) Data concentration of the system: the development of the personalized teaching system for vocal music is only for students learning python courses, and it concentrates and categorizes the excellent videos, courseware and other materials of the whole network as well as individual teachers.

(d) Convenient operation of the system: the vocal personalized teaching system is developed for school teachers and students, fully corresponding to the call for education and teaching reform, to meet the needs of students' personalized learning, the system has a simple interface, clear modules, and easy to operate.

(e) Maintainability of the system: the R&D technology of the vocal personalized teaching system is mature and reliable, the code is readable, the notes are clear, and the maintenance is convenient.

#### II. D. 2) Outline design

After the requirement analysis of vocal personalized teaching system, the technical solution selects the Django web development framework based on MVC constructs, i.e., model, view, and controller. Firstly, the controller waits for user input, when the user inputs, the controller passes the user input instruction to the business instruction, then the model makes business logic judgment and database access, selects different views according to the business logic, renders the page, and finally feeds back the result to the user, who gets the feedback and then proceeds to the next operation. The front-end design mainly uses HTML5, JavaScript language, using a lightweight front-end management framework based on layui Xadmin, simple, good compatibility. And used MySQL database as a data-driven source to store a large amount of data in the system.

#### II. D. 3) System implementation

Data fusion algorithms based on vocal personalized teaching system development tools used with a full set of tools that can help users improve the efficiency of language development PyCharm editor, can provide users with better display, and use python3.6 and Django2.0 version of the web framework construction, and with virtualenv virtual environment, compatibility testing were conducted in IE, 360 Safe Browser, Goolge Chrome, etc. Django is a web development framework that follows the MTV development model, the development of which starts with the creation of a project and the establishment of the corresponding app in this project, followed by the need to change the configuration of the project's setting.py file, including the following MySQL database, apps, CSS styles and templates and other file settings, and then the development of the data model is mainly in the model.py file, model.py file defines a variety of data models and database interaction, in the definition of a good data model, view.py file defines how to operate the model.py data objects and interface implementation, and then by writing the front-end display. After defining the data model, the view.py file defines how to manipulate the data objects in model.py and how to realize the interface, and then the methods in view.py file are called by writing the front-end display. The module files in the templates folder represent the view, which is responsible for displaying the data content, and the front-end files such as html are stored in the templates. The last is to add URL mapping, urls.py defines a variety of url access entry. Django-based web application development activities are mainly concentrated in the model.py, view.py, templates folder, so Django's development model is also known as the MTV development model. With the above software support technology, the design work of vocal personalized teaching system is finally achieved.

## III. Algorithm verification and system simulation analysis

### *III. A. Algorithm Validation Analysis*

#### III. A. 1) Validation of MFCC Audio Feature Extraction Methods

(1) Experimental data

In order to accurately grasp the position of each starting point in the music. The music in this experiment uses the music score through the scoring software Overture to generate music with multiple voices, different styles and speeds. The experimental music data sampling rate is 44.1KHz, quantization precision is 32bit, mono wav format file. The duration of each song varies from 20s to 80s, and a total of 10 pieces of music with different beat types are selected, as shown in Table 1, as the experimental data for audio feature extraction. A total of 200 music bars, totaling 360 accented music clips were intercepted from it.

Table 1: The music selected for note start point detection

| Name of the Music Piece | Beat type |
|---|---|
| A | 2/4 |
| B | 2/4 |
| C | 2/4 |
| D | 2/4 |
| E | 3/4 |
| F | 3/4 |
| G | 4/4 |
| H | 4/4 |
| I | 4/4 |
| J | 4/4 |

(2) Evaluation Criteria

For the evaluation of audio feature extraction, the correct rate and the misrecognition rate are used as the evaluation criteria. Among them, the correctly recognized note onset is the note onset identified by the system, which is the real note onset position of the music, (the error is less than 40ms), and the misrecognized music onset indicates that the recognized note onset is not the real note onset in the music.

Sample results of accent onset feature extraction detection using different feature extraction methods are shown in Figures 2 to 5, which represent LPC (linear prediction coefficient) based features, Chroma based features, energy based, and MFCC based features, respectively. The results of the analysis of the piece of music 1 with a length of 10s. Different audio features are extracted for the experimental data, respectively, and the time corresponding to the current frame is considered to be the start position of an accented note if the distance value of the current frame is greater than the threshold value. The accuracy of the recognition, as shown in Table 2, Table 3. From the above results, it can be seen that the method of using energy is not ideal for accent note onset feature extraction, and when the MFCC method is used its correct rate is 0.966 (0.972), 0.028 (0.024), which is much better than the other audio feature extraction methods (Chroma features, LPC features, and energy features), which verifies the validity of the MFCC audio feature extraction method for the The subsequent design of vocal personalized teaching system based on multimodal data fusion algorithm provides accurate audio feature guarantee, so that its system performance and application effect can be improved.

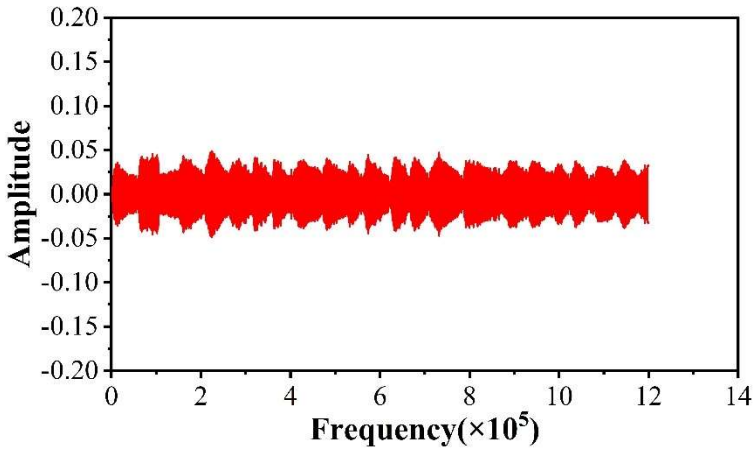



Figure 2: Based on the LPC features

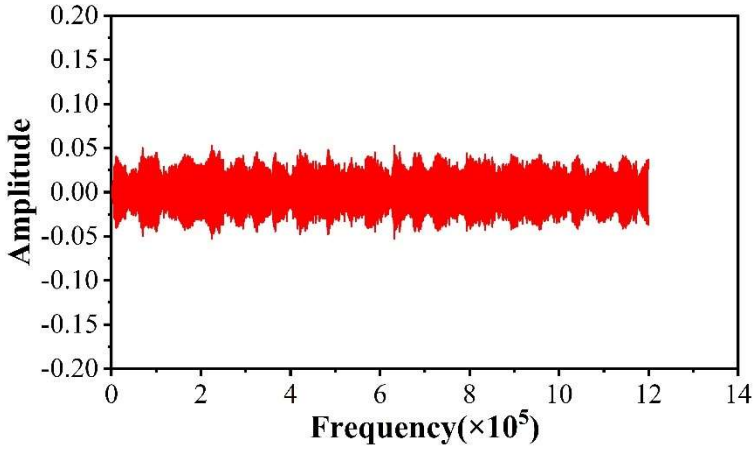



Figure 3: Based on Chroma features

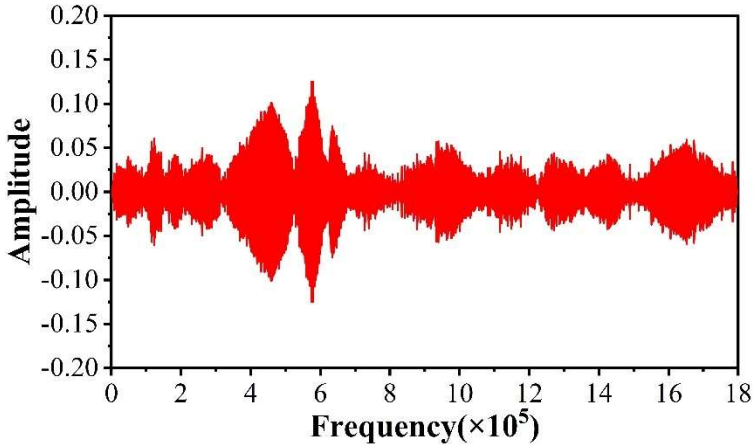



Figure 4: Based on energy

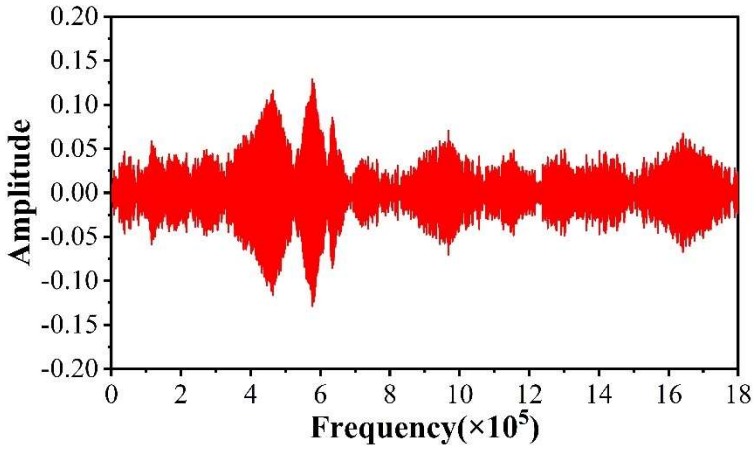



Figure 5: Based on the MFCC features

Table 2: Accuracy rate comparison (threshold 1.1)

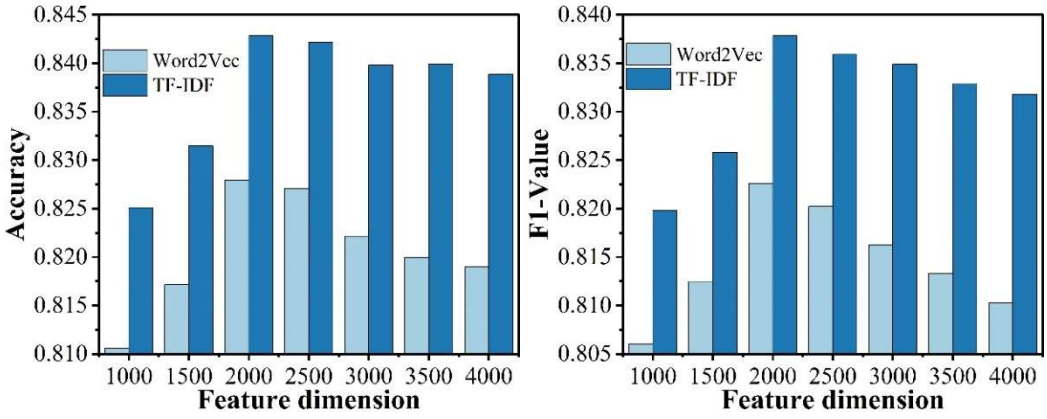
| Selection features | Starting point number | Correct number | False recognition number | Rejection number | Accuracy rate | False recognition rate |
|---|---|---|---|---|---|---|
| Energy | 500 | 353 | 5 | 5 | 0.706 | 0.284 |
| Chroma | 500 | 406 | 8 | 8 | 0.812 | 0.172 |
| LPC | 500 | 437 | 7 | 7 | 0.874 | 0.112 |
| MFCC | 500 | 483 | 3 | 3 | 0.966 | 0.028 |

Table 3: Accuracy rate comparison (threshold 0.5)

| Selection features | Starting point number | Correct number | False recognition number | Rejection number | Accuracy rate | False recognition rate |
|---|---|---|---|---|---|---|
| Energy | 500 | 356 | 11 | 11 | 0.712 | 0.266 |
| Chroma | 500 | 411 | 9 | 9 | 0.822 | 0.16 |
| LPC | 500 | 429 | 5 | 5 | 0.858 | 0.132 |
| MFCC | 500 | 486 | 2 | 2 | 0.972 | 0.024 |

**III. A. 2) Verification and Analysis of Vocal Text Feature Extraction Methods**

Since music lyrics are generally short and many words and phrases appear repeatedly, the extracted text features are not many, totaling 400 words. As with “audio words”, the effects of different numbers of text features on the experimental results were compared in the experiment, and the validation analysis of the vocal text feature extraction method is shown in Figure 6. Where (a)~(b) are the accuracy rate and F1 value, respectively. Based on the data performance in the figure, it can be seen that when the music lyrics features are increased from 1000 to 2000 dimensions, the Word2Vec, TF-IDF accuracy, F1 value values are improved, and as the feature dimensions continue to increase, the accuracy, F1 value values of Word2Vec and TF-IDF show a decreasing trend, and in general, the TF-IDF method is better than the Word2Vec, verifying the priority of TF-IDF method in vocal text feature extraction.




(a)Accuracy (b)F1

Figure 6: Verification of Vocal text Feature extraction methods

**III. A. 3) Validation analysis of multimodal data fusion algorithms**

The above extracted vocal audio and text features are integrated to various classification algorithms for vocal multimodal fusion analysis, and the multimodal data fusion algorithm validation analysis is shown in Fig. 7, in which CNN-EFFC, RNN-EFFC, SVM-EFFC, and LSTM-EFFC are the control multimodal data fusion algorithms, and the performance evaluation indexes of the different multimodal data fusion algorithms are evaluated by comparing the performance values (Accuracy, Recall, F1 value) to verify the effectiveness of the multimodal data fusion algorithm based on HMM-EFFC. Based on the magnitude of Accuracy, Recall, and F1 values, it can be seen that HMM-EFFC (Accuracy: 0.9635, Recall: 0.9718, F1 value: 0.9676) is better than CNN-EFFC (Accuracy: 0.7531, Recall: 0.7644, F1 value: 0.7887), RNN-EFFC (Accuracy: 0.7762, Recall: 0.8099, F1 value: 0.7927), SVM-EFFC (Accuracy: 0.8134, Recall: 0.8675, F1 value: 0.8396), LSTM-EFFC (Accuracy: 0.8527, Recall: 0.8841, F1 value: 0.8681), proving that the practical application performance of the multimodal data fusion algorithm based on HMM-EFFC makes the

designed vocal personalized teaching system more suitable for the music teaching classroom in colleges and universities, and then promotes the innovative development of vocal teaching in colleges and universities.

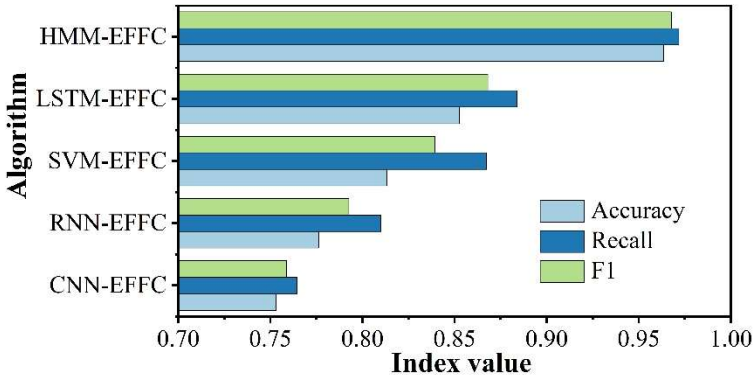



Figure 7: Verification and analysis of multimodal data fusion algorithms

### *III. B. System Simulation Analysis*

The study conducted experiments on a dataset used to validate the effectiveness of the application of the HMM-EFFC-based personalized teaching system for vocal music. In this dataset, there are a total of 254 lessons of vocal music teaching video data in different environments, and the teaching locations contain various regions. Since the amount of data in the dataset is limited, the study divides them into two groups according to the ratio of 1:4, which are used for the training and testing of the system respectively. In order to verify the superiority of the fusion system, the study simultaneously conducts experiments on the vocal personalized teaching system based on CNN-EFFC, the vocal personalized teaching system based on RNN-EFFC, and, the vocal personalized teaching system based on SVM-EFFC, and the interest judgement experiments of the four systems are shown in Fig. 8, and the CPU occupancy rates of the four systems are shown in Fig. 9. As seen in Fig. 8, the accuracy of the system's judgment of students' learning interest gradually increases. When the running time of the system is 64s, the accuracy rate of learning interest judgment of the four systems reaches the maximum. At this time, the accuracy rate of this paper's system is the highest, 98.37%. It indicates that the system in this paper has the best performance in judging students' vocal learning interest in the same classroom. As seen in Figure 9, the CPU occupancy of the proposed system is positively correlated with the running time. In this environment, the average CPU occupancy of HMM-EFFC, CNN-EFFC, RNN-EFF and SVM-EFFC systems are 27.53%, 35.78%, 37.54 and 42.28%, respectively. It indicates that for vocal music teaching, the study proposes that the system in this paper has the highest adaptability.

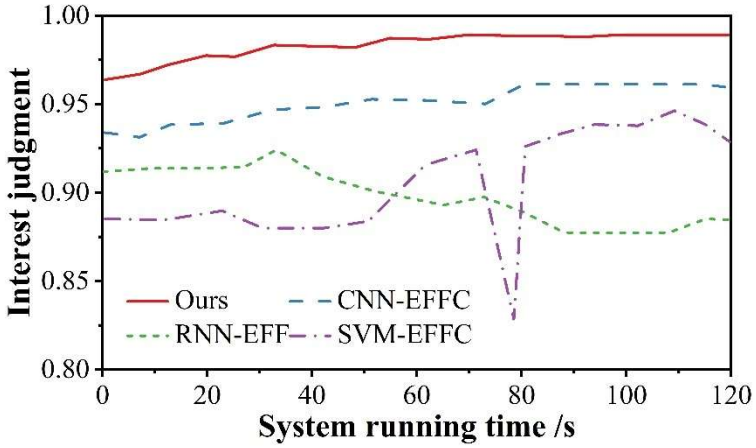



Figure 8: Interest judgment experiments of four systems

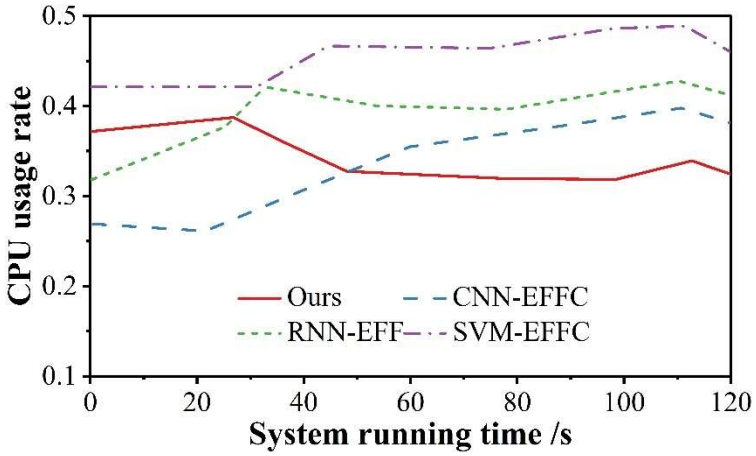



Figure 9: The TPU occupancy rates of the four systems

The study focused on the differences in learning experience brought about by different teaching systems by controlling variables to ensure consistency in environment, content, time and difficulty as much as possible, and the results of the learning experience evaluation are shown in Fig. 10, in which X1~X6 denote the interestingness of learning, visual effect, acceptance, equipment operation, interaction design, and comprehensive evaluation, respectively. As can be seen in Figure 10, the differences in the comprehensive evaluation of the four systems are not significant. However, the interestingness of learning varies greatly among different systems, and the system in this paper is significantly higher than the CNN-EFFC, RNN-EFF and SVM-EFFC systems, with the difference amounting to 0.5172, 0.9154 and 0.6684, and there is no significant difference (P>0.05), and there is a homogeneous phenomenon in the rest of the perceptual effect, acceptance, device operation, interaction design, and comprehensive evaluation.

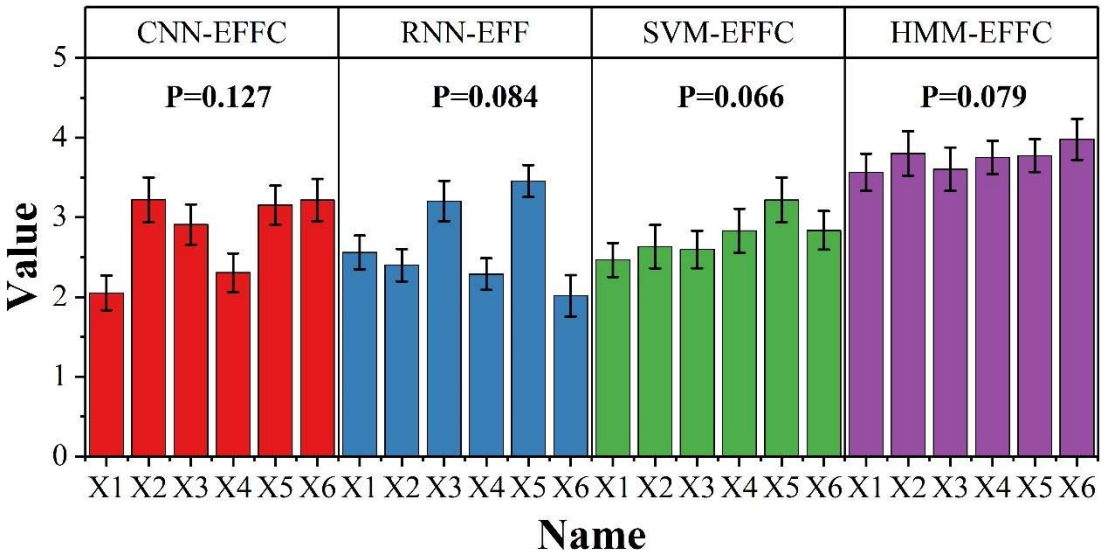



Figure 10: Evaluation results of learning experience

The students' vocal music listening efficiency and acceptance of the four systems were compared, and the experimental data were recorded and the images were plotted as shown in Fig. 11 to Fig. 12. As can be seen from Figure 11, with the increase of the number of listeners in vocal music teaching, the listening efficiency of students in the classroom shows fluctuating changes. And the system of this paper has the highest value in it, which is 2.711, when the experimental values of the other three systems are 2.294, 1.879 and 1.405 respectively, which indicates that the fusion system proposed by the research is widely used in students' listening efficiency. As can be seen in Fig. 12, the four systems show an increasing trend in the performance of student acceptance as the teaching speed rises.The student acceptance of the HMM-EFFC, CNN-EFFC, RNN-EFF and SVM-EFFC systems are 99.38%, 97.14%, 90.25% and 93.33%, respectively, of which the system of this paper has the optimal performance, which shows its wide range of student acceptance, suitable for obtaining applications in vocal music teaching.

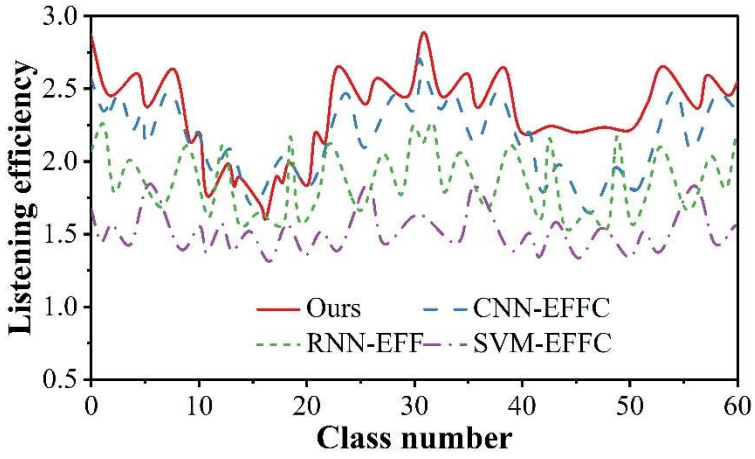



Figure 11: Experiments on the listening efficiency of four systems

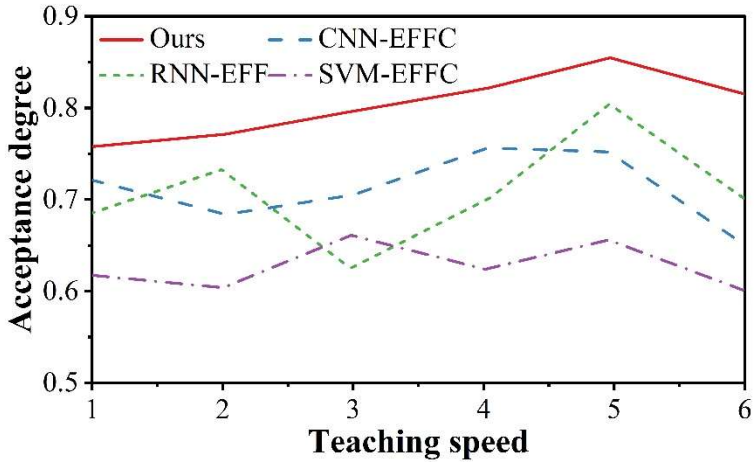



Figure 12: The student acceptance level of this system

## IV. Conclusion

In this paper, we designed a personalized teaching system for vocal music based on data fusion algorithm, and realized the personalized recommendation and evaluation of vocal music teaching through the effective combination of MFCC audio feature extraction, TF-IDF text feature extraction and HMM-EFFC multimodal fusion algorithm.

The experimental results show that the audio feature extraction accuracy of the MFCC method reaches 0.966 (threshold 1.1) and 0.972 (threshold 0.5), which is much higher than that of the energy features (0.706 and 0.712), chromaticity features (0.812 and 0.822), and LPC features (0.874 and 0.858); as for the text features, the TF-IDF method performs best when the feature dimensions are 2000 performs best and is overall better than the Word2Vec method; the HMM-EFFC multimodal fusion algorithm has an Accuracy of 0.9635, a Recall of 0.9718, and an F1 value of 0.9676, which is significantly better than the comparative algorithms such as CNN-EFFC, RNN-EFFC, SVM-EFFC, and LSTM-EFFC; In the system application test, the accuracy of learning interest judgment is up to 98.37%, the average CPU occupancy is only 27.53%, and the acceptance degree of students is 99.38%; in the evaluation of learning experience, this system outperforms the comparative systems in the aspects of learning interest, visual effect, acceptance, device operation, and interaction design.

The study proves that the personalized teaching system for vocal music based on data fusion algorithm can effectively improve the teaching effect and learning experience, and provides new ideas and technical support for vocal music teaching innovation. Future research will further optimize the fusion algorithm, expand the system functions, improve the model generalization ability, and promote the intelligent and personalized development of vocal music teaching.

## References


[1] Okada, B. M., & Slevc, L. R. (2018). Individual differences in musical training and executive functions: A latent variable approach. Memory & Cognition, 46, 1076-1092.

[2] Hash, P. M. (2021). Supply and Demand: Music Teacher Shortage in the United States. Research and Issues in Music Education, 16(1), 3.

[3] Rong, G. (2021). A study on diversified teaching methods of vocal music teaching in colleges. Advances in Vocational and Technical Education, 3(3), 24-31.

[4] Fu, L. (2020). Research on the reform and innovation of vocal music teaching in colleges. Region-Educational Research and Reviews, 2(4), 37-40.

[5] Bo, L. (2021). Analysis of Problems in Vocal Music Singing and Performance Teaching in China's Colleges and Universities and the Corresponding Countermeasures. Journal of Frontiers in Educational Research, 1(4), 90-93.

[6] Salvador, K. (2019). Assessment and individualized instruction in elementary general music: A case study. Research Studies in Music Education, 41(1), 18-42.

[7] Haitao, Y., & Hirunrux, S. (2023). Problems and Solutions of Vocal Music Teaching in Hezhou University in China. Journal of Modern Learning Development, 8(6), 194-200.

[8] Jiang, M., & Chao-Jung, W. U. (2022). Investigation and reflections on gender differences in vocal music education. Revista de Cercetare si Interventie Sociala, 79, 59.

[9] Duan, G., Zeng, J., & Ji, H. (2021). The Importance of Psychological Analysis to Vocal Music Singing Teaching. Psychiatria Danubina, 33(suppl 7), 315-317.

[10] Zarachi, A., Lianou, A., Liontos, A., Garefis, K., Litsou, E., Tafiadis, D., ... & Exarchakos, G. (2021). Stroboscopic Evaluation of the Vocal Folds' Morphometric Characteristics, between the Different Voice Categories on Professional Singers. International Journal of Otolaryngology and Head & Neck Surgery, 10(5), 345-358.

[11] Yin, W. (2024). Innovations and Practical Exploration of Vocal Music Teaching Models in Vocational Colleges. Journal of Modern Educational Theory and Practice, 1(2).

[12] Wang, Y., & Gan, Y. (2025). Research on optimization of vocal music teaching mode and design of personalized learning path based on AI algorithm. J. COMBIN. MATH. COMBIN. COMPUT, 127, 5669-5690.

[13] Chen, S. (2024). The Application of Big Data and Fuzzy Decision Support Systems in the Innovation of Personalized Music Teaching in Universities. International Journal of Computational Intelligence Systems, 17(1), 215.

[14] Pavlenko, O., Rastruba, T., Huseinova, L., Khomenko, L., Pelekh, K., & Koval, O. (2024). The Role of Current Computer Technologies in Developing Professional Competence of Music Teachers: A Model of a Personalized Educational Environment. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15(1), 325-340.

[15] Zuoliang Wang,Qimin Xu & Zehua Chen. (2025). Automatic evaluation method for vehicle audio warning system using MFCC-polynomial hybrid feature. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering,239(5),1577-1586.

[16] Bhutia Sonam Lhamu,Borah Samarjeet,Pradhan Ratika & Sharma Bhushan. (2020). An Experiment on Parameter Selection for Landslide Susceptibility Mapping using TF-IDF. Journal of Physics: Conference Series,1712(1),012029-.

[17] Chen Yunli & Zheng Haiyang. (2023). The application of HMM algorithm based music note feature recognition teaching in universities. Intelligent Systems with Applications,20,

[18] Nan Yang,Cheuk Hang Leung & Xing Yan. (2024). A novel HMM distance measure with state alignment. Pattern Recognition Letters,186,314-321.

[19] Gia-Nhu Nguyen & Trung-Nghia Phung. (2017). Reducing over-smoothness in HMM-based speech synthesis using exemplar-based voice conversion. EURASIP Journal on Audio, Speech, and Music Processing,2017(1),1-7.