International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5383-5393

https://doi.org/10.70517/ijhsa463464

Research on the System Design and Teaching Effect of Interactive Multimedia-Assisted Vocal Music Teaching

Linhao Qin[1] and Meitian Zhao[1,*]
[1] Taiyuan Normal University, Taiyuan, Shanxi, 030619, China
Corresponding authors: (e-mail: lunatian05@163.com).

Abstract This paper introduces a hybrid recommendation algorithm based on collaborative filtering of vocal resources and content in the design of interactive multimedia-assisted vocal teaching system. Based on historical learning data and so on, the sparse matrix between students' attributes and resources is constructed, and the similarity is calculated to complete the accurate recommendation of vocal music learning resources. The speech emotion recognition module uses a convolutional neural network (CNN) model based on multilevel residual improvement. The multilevel residual structure reduces the loss rate of vocal singing voice features, and at the same time reduces the amount of model computation to ensure that students' voices are accurately recognized. The results show that: the resource similarity range of this paper's hybrid recommendation algorithm is [0.748,0.894], the resource coverage are greater than 95%, at the same time, the AUC area is greater than 0.9. The recognition rate of the model based on the improved CNN is stable greater than 0.95 for about 45 iterations, and the loss value is less than 0.4. The introduction of RMSProp algorithm has the optimization of 0.03 and 0.15, respectively. Effect. The mean value of the system's effect on vocal music teaching reaches more than 4.5, and the standard deviations are all less than 0.10.

Index Terms speech emotion recognition, vocal music teaching system, hybrid recommendation, improved CNN model, multilevel residuals

I. Introduction

Vocal music teaching is a course that cultivates students' singing skills and enables them to master the ability to analyze, understand and express works of art [1]. At present, the teaching of practical lessons of vocal music in colleges and universities adopts the teaching mode of one-to-one, one piano for two people, and oral transmission, which has been widely used in the teaching of practical lessons of vocal music in colleges and universities since it was introduced into China at the beginning of the 20th century. After the continuous reform and the rise of multimedia technology in recent years, computer-assisted teaching has been introduced in many courses except for vocal technique courses, and obvious teaching effects have been achieved [2]. On the one hand, multimedia technology can provide an interactive teaching environment so that students can participate in the learning process more actively [3]. Through multimedia technology, students can interact with the teaching content, for example, by clicking, dragging, and selecting to manipulate and explore the learning content [4]-[6]. This interactivity not only improves students' learning interest and participation, but also promotes students' understanding and memorization of knowledge [7]-[9]. On the other hand, multimedia technology can present abstract concepts and theories in an intuitive way, making it easier for students to understand and master them [10]. Audio and video are the artistic language of multimedia technology; they can vividly demonstrate vocal techniques and performances, providing students with models to imitate and learn from [11], [12]. This pedagogical intuition not only reduces the difficulty of learning, but also improves students' comprehension and learning outcomes [13], [14]. Therefore, how to introduce the network, hypertext system, multimedia technology into the teaching of vocal technique class to overcome the shortcomings of the traditional teaching methods and improve the classroom efficiency and students' interactive participation rate is an issue that should be considered by the majority of vocal music educators.

Satisfying students' personalized vocal learning needs is a necessary path for the digital development of vocal teaching. In this paper, we use a hybrid recommendation algorithm based on collaborative filtering and content to comprehensively construct a sparse matrix containing student attributes, operational information, and resource attributes. Based on the vector weights, the similarity between resources to be recommended and between students and resources is calculated to improve the accuracy of vocal music resource recommendation. Reduce the feature extraction loss of students' vocal singing speech through a convolutional neural network improved based on multilevel residuals, combining multiple convolutional pooling layers with a multilevel residual structure. The

RMSProp algorithm is introduced as a loss function optimization algorithm to reduce the amount of loss in the model recognition process. The designed and optimized system is applied to vocal music teaching in colleges and universities, and the practical auxiliary teaching influence ability of the system is mined from the aspect of student evaluation data.

II. Technical analysis of the interactive multimedia-assisted vocal teaching system

This part describes the design and function realization process of the interactive multimedia-assisted vocal music teaching system from the perspective of technical analysis.

II. A. The role of interactive multimedia teaching system in vocal music teaching

II. A. 1) Utilizing rich vocal materials for student instruction

In the process of vocal music teaching, teachers must bring the role of interactive multimedia teaching system into effective play, and instruct students with rich vocal music materials. With the support of interactive multimedia technology, teachers can collect a lot of teaching materials that are related to vocal music teaching, and after teachers organize these materials effectively, they can apply them to classroom teaching, and students can get different gains under the display of intuitive teaching materials. For example, the content of a class needs to introduce students to certain musical instruments, and students have no previous contact with these instruments, in this case, even if the teacher can be detailed, colorful explanation, students can not really feel these instruments on the vocal performance of how to help and role. At this time, if the teacher can display vivid and rich pictures of the instruments and the information of the instrument's performance for the students through the interactive multimedia teaching equipment, then the students can have a more profound understanding of the instruments. After doing so, not only greatly enriched the content of classroom teaching, and students in the vocal subject learning enthusiasm can also be effectively stimulated, the quality of vocal teaching can also be improved.

II. A. 2) Utilization of multimedia technology to assist in classroom instruction

In the process of formal vocal classroom teaching, it is especially important for colleges and universities to make good use of interactive multimedia technology. Teachers can record the students' vocal practice by using the recording function of interactive multimedia teaching equipment in the classroom, and with the help of some software in the interactive multimedia teaching equipment, they can play the practice scenes of the students to help the students to recognize their own deficiencies. In addition, video technology in vocal teaching also has a broad application prospects, because many students in the classroom training performance is very good, once the formal performance of the work will be nervous, unable to play the normal level of the situation, at this time, interactive multimedia video technology can play a key role. Teachers in the teaching process using video technology can be recorded in the formal performance of the scene, and directly applied to the classroom teaching, pointing out the students in the formal singing deficiencies, targeted practice. Video technology can also record the students' singing, so that the students themselves to watch, so that students find their own problems in singing and timely correction. The use of video technology in vocal teaching can also increase students' self-confidence in themselves, no longer afraid of formal singing.

II. B. Vocal Resource Recommendation Model

In order to improve the personalization and student experience of the interactive multimedia-assisted vocal teaching system, the home page of the system contains a recommendation column, which can be derived by the algorithm and then recommended to the current students who may like the resources related to vocal teaching. This process is based on the student's history information, student information, content labels and other data on the one hand, and on the other hand, based on the popular information such as the recent popular vocal songs on the Internet obtained by the Scrapy crawler, and the data to be processed are mostly textual data types with these two aspects as the data support. Considering the factors such as low computing power consumption, processing of pictorial labels of text type, complexity of algorithm implementation, and timeliness requirement of rapid growth of songs, the type of algorithm that best suits the needs under this requirement i.e., Hybrid Recommendation Algorithm based on Collaborative Filtering and Content is selected.

II. B. 1) Hybrid Recommendation Algorithm Based on Collaborative Filtering of Vocal Resources and Content

On the one hand, the reason for choosing the content-based recommendation algorithm is that the use of this algorithm can reduce the interference of all kinds of external factors on the algorithm itself, and at the same time, the recommendation results derived from this algorithm can show certain characteristics of the matched students and have better interpretability. In terms of constructing the feature representation of labeled resources, since most of the labeled resources in the recommendation system belong to text-based data, the TF-IDF algorithm is used to

transform the textual information into feature vectors and compute them to obtain the weights of each keyword in the sparse matrix. On the other hand, the collaborative filtering algorithm model based on vocal resources has strong generalization ability, which can recommend similar vocal resources to students' previous favorites, and can filter out the more complex expression information to avoid the ambiguity of content analysis. The term “vocal resources” covers the individual objects that the recommendation algorithm computationally targets that students may manipulate in the teaching platform, including but not limited to music, teaching videos, instructors, etc., which are also referred to as “labeled resources” in the following. The whole recommendation algorithm is divided into four steps:

1) Generate a sparse matrix of feature weights of the vocal resources by TF-IDF computation;
2) Calculate the similarity between labeled resources by collaborative filtering algorithm based on vocal resources;
3) Calculating the similarity between students and vocal resources;
4) Sorting the results from high to low and recommending them to students using TOP-N analysis.

II. B. 2) Construct sparse matrices based on student attributes with operational information and vocal resource attributes

The TF-IDF algorithm is a text data weighting technique, i.e., it calculates the weights of text type objects based on keywords and generates a weight sparse matrix, which represents the weight ratio of the keywords in a document.TF, i.e., the frequency with which a particular word occurs in a particular document, is calculated by the formula:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

In Equation (1), the numerator $n_{i,j}$ denotes the number of times the keyword appears in document j; while the denominator denotes the sum of all keyword occurrences in document j.

IDF is the inverse document frequency, which represents a measure of the general importance of a keyword. If a word appears in only a small number of documents, then the value it contains is huge, and the opposite reduces the distinction of the keyword to the document. Its calculation formula is as follows:

$$IDF_i = \log_{10} \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right| + 1} \tag{2}$$

In Eq. (2), D denotes the total number of documents in the corpus; $\left|\left\{j : t_i \in d_j\right\}\right|$ denotes the number of documents containing the word t_i, and $+1$ is used to prevent the denominator from zero due to the absence of keywords in the corpus. Then the TF-IDF is calculated as:

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

From equation (3), the size of the keyword weights is directly proportional to the number of occurrences of the word in the document and inversely proportional to the number of occurrences of the word in the entire collection of documents. The sparse matrix of weights of vocal resources can be derived from its calculation.

II. B. 3) Calculating similarity between vocal resources based on vector weights

Before calculating the similarity between vocal resources, it is necessary to filter the results of similarity calculation to exclude the two vocal resources with zero similarity in all dimensions and not to calculate them to reduce the time complexity of the algorithm. The similarity between vocal resources is calculated using the cosine similarity algorithm, and according to the different angles of the cosine value of the value of the domain of $[-1,1]$. The cosine value between two vectors determines the direction of the two vectors, and the similarity is inversely proportional to the distance. Its specific calculation formula is as follows:

$$W(i,j) = \frac{\left|N(i) \cap N(j)\right|}{|N(i)| \times |N(j)|} \tag{4}$$

In Eq. (4), $N(i)$ denotes the n-dimensional vector matrix of vocal resource i, $N(j)$ denotes the n-dimensional vector matrix of vocal resource j, and $W(i,j)$ denotes the similarity between vocal resource i and

vocal resource j, which ranges from $[0,1]$, and the closer it is to 1, the higher the similarity of the two vocal resources.

II. B. 4) Calculating similarity between students and vocal resources

In this step, the labeled resources that students have had records of liking, repeated browsing and other operations are used to calculate the similarity between vocal resources by the above formula, to find out the set of vocal resources that have the highest similarity with them, and the similarity formula of the vocal resources are compared one by one by the students to come up with a new recommendation set. The formula for the degree of interest of student u in vocal resource j is as follows:

$$P(u,j)=\sum_{i\in N(u)\cap M(j,k)} W_{ji}R_{ui} \tag{5}$$

In Eq. (5), $M(j,k)$ denotes the set of k vocal resources that have the highest similarity to vocal resource j; $N(u)$ denotes the set of student's favorite vocal resources; W_{ji} denotes the degree of similarity between the vocal resources; and R_{ui} is the degree of interest of student u in vocal resource i. In particular, let $R_{ui}=1$ when the student's history operation if he/she has performed an active operation on the vocal resource.

In addition, the TOP-N analysis arranges the result set obtained after performing the similarity computation between $P(u,j)$ students and vocal resources in reverse order, and takes the vocal resources in the top N items to recommend them to students.

II. C.Improved CNN-based speech emotion recognition model for vocal singing

In this study, we improve the phenomenon that students' vocal singing speech features are easily lost during the extraction process, and propose a convolutional neural network (CNN) based on multilevel residual improvement to make up for the lost features and improve the recognition rate. Figure 1 shows the improved convolutional neural network structure.

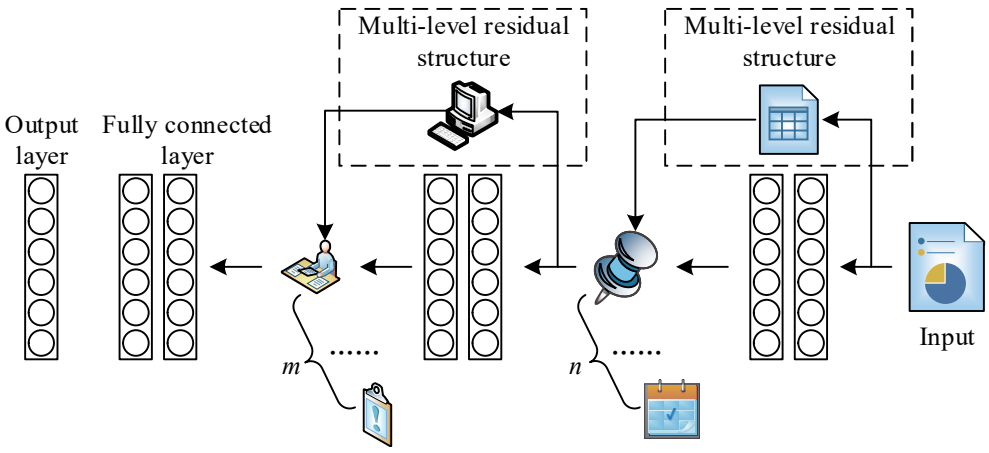


Figure 1: Improved convolutional neural network

In Fig. 1, the improved convolutional neural network contains multiple convolutional pooling layers with multilevel residual structures, in which the multilevel residual structures are able to transmit the original information features across multiple pooling layers, which can compensate for the lost features. The square box indicates the multilevel residual structure involved in the study, and n and m represent the number of convolutional layers across which the residual structure propagates, respectively. The improved multilevel residual convolutional neural network in this study is able to further reduce the computational amount and thus improve the model recognition efficiency on the basis of solving the drawbacks of the traditional CNN structure. The multilevel residual structure proposed in the study improves the convolutional neural network based on the residual structure. The multilevel residual structure maximizes the retention of the original information by interconnecting the original information of the first n convolutional layer with the convolutional layer where it is located, and regulates the dimensions of the input original features by adding control coefficients, which ultimately improves the recognition efficiency of the model and the

convergence speed of the system effectively. If the input is set as x_i when the residual structure is induced, and the output is x_{i+n} after the residual structure is introduced, then the corresponding output of the multilevel residual structure is shown in equation (6).

$$\begin{cases} x_{i+n} = \sigma(w_{i+n}F(x_{i+n-1}) + b_{i+n} + \alpha x_i) \\ F(x_{i+n-1}) = \begin{cases} \sigma(w_{i+n}F(x_{i+n-1}) + b_{i+n}), n \neq 1 \\ \beta x_i, n = 1 \end{cases} \end{cases} \tag{6}$$

In Eq. (6), α, β denote the control coefficients, which are used to limit the dimensionality of the input features. If the loss function is C, the corresponding obtained expression for the weights of the backpropagation is updated to Eq. (7).

$$\begin{aligned} \frac{\partial C}{\partial x_i} &= (\partial C / \partial x_{i+n})(\partial x_{i+n} \partial x_i) \\ &= \left[\left(\beta \prod_{k=1}^{n} w_{i+k} + \alpha \right) x_i + T(w,b) \right] \left(\frac{\partial C}{\partial x_{i+n}} \right) \end{aligned} \tag{7}$$

In Eq. (7), $\left(\beta \prod_{k=1}^{n} w_{i+k} + \alpha \right) x_i$ denotes the lost feature term corresponding to the supplement of the multilevel residual structure; and $T(w,b)$ denotes the constant term of weights and bias. When solving the convolutional neural network layer by layer, the weight w will gradually decrease until it approaches 0, which leads to the gradient update of the back-passing approaching 0, and ultimately causes the phenomenon of feature loss. After adding the multilevel residual structure, the original feature information of the first n convolutional layers can be introduced, so that the features can be maximized and complementary, and at the same time, setting the control parameters α and β, the feature dimensions can be reduced, and ultimately accelerate the training speed and improve the training effect. The training of speech emotion recognition during vocal singing can be regarded as using a set of estimated parameters to portray the differences between the actual data and the predicted data, and by constantly adjusting these differences, the differences between the two are realized to be minimized. On this basis, a new function that can be used to guide the whole system, the loss function, is proposed. Due to the existence of multi-category features in the process of speech emotion recognition, cross entropy is then introduced into speech emotion recognition, and its corresponding expression is shown in equation (8).

$$C = -\sum_{j=1}^{k} \hat{y}_j \log(y_j) \tag{8}$$

In Eq. (8), $\hat{y}_j$ denotes the true label of the j th sample; y_j denotes the predicted output of the j th sample in the constructed model. The problem described by the cross entropy is the degree of difference between the output results and the true value, if more samples are correctly categorized in the output results, the corresponding obtained cross entropy value is smaller, i.e., it means that the less confusing the output results are, the better performance of the model obtained in the end.

In addition, it is also necessary to select an appropriate optimization algorithm to minimize the loss function value of the neural network. The most common method is to use the gradient descent method to minimize the loss function, in this regard, it is proposed to use the RMSProp algorithm (with full parameter adaptive characteristics), the specific expression is shown in equation (9).

$$\begin{cases} r := \eta r + (1-\eta) \left(\frac{\partial C}{\partial w} \right)^2 \\ w := w - \alpha \frac{\partial C / \partial w}{\sqrt{r} + \varepsilon} \end{cases} \tag{9}$$

In Eq. (9), r denotes the sliding rate of the squared value of the gradient; w denotes the decay rate; α denotes the learning rate; ε denotes the constant term that prevents the denominator from being zero; and η denotes the hyperparameter (which is a constant). In addition to this, the regularization method is used in order to

avoid the phenomenon of overfitting. This method is an effective training strategy capable of randomly ignoring neurons in the neural network results.

III. Analysis of system optimization design effects and application results

This section analyzes the design effectiveness of interactive multimedia teaching aid systems from two aspects, namely, resource recommendation and singing voice recognition, and investigates their use in actual teaching aids.

III. A. Learning data collection and organization

College A was chosen as the research object. This university is a well-known institution in the field of vocal music, and has a relatively complete system of vocal music professional training. The scale of students majoring in vocal music is huge, and the annual enrollment is stable at about 300 students.

In terms of teaching digital vocal music courses, the university has initially established a relevant teaching resource library and introduced interactive multimedia teaching equipment to enhance the teaching effect. However, the current systematic teaching resource management and recommendation methods are still relatively traditional, and students need to spend a lot of time searching and screening the required resources. Meanwhile, due to the special nature of the vocal teaching curriculum, students need to practice singing and simulated performances frequently, but the existing experimental equipment and venues are difficult to meet this demand. In summary, although the school has made some progress in teaching digital voice courses, there are still many deficiencies. The student learning data were collected as the sample data for this experiment, and the data were collated to obtain the student vocal learning sample data in Table 1. From the preliminary results of the collation, due to the difficulty of current resource retrieval and screening, the average number of students' browsing per day is not more than 20 items, and the number of online interactions is not more than 40 times. It indicates the urgency of improving the design of the interactive multimedia-assisted vocal music teaching system.

Table 1: Sample data of students' vocal music learning

Student Number	Online learning days/d	Average page views/(items·d^{-1})	Number of online interactions/times
1	26	16	31
2	24	11	26
3	29	19	36
4	22	12	21
5	31	19	38
6	26	17	33
7	23	15	27
8	27	19	37
9	23	13	21
10	27	18	36

III. B. Comparison of the Effectiveness of Hybrid Recommendation Algorithms Based on Collaborative Filtering and Content

III. B. 1) Resource Recommendation Similarity Comparison

To ensure that the method proposed in this paper has resource recommendation performance advantages, the system parameters are set up after collecting and organizing online student learning data. At the same time, rich teaching resources are prepared to test the similarity, coverage and AUC value of resource recommendation of the comparative recommendation system. The prepared teaching resources are uploaded to the cloud computing platform and reasonably categorized and labeled so that the system can accurately identify and recommend relevant resources. Simulate students' learning behaviors, including browsing resources and interacting, in order to trigger the operation of the recommendation system.

On this basis, this paper introduces the recommendation method based on knowledge graph, the recommendation method based on students' behavioral characteristics, which is used as a control with the hybrid recommendation algorithm based on collaborative filtering and content in this paper. At the same time, the three methods are applied to carry out the automatic recommendation of teaching resources for digital voice courses, collect the recommendation results generated by the recommendation system and compare and analyze them with the actual learning needs of students to assess the similarity of the recommendations.

The similarity between recommended resources and learning needs is the key to measuring the accuracy of the recommendation system. This indicator is derived by calculating the degree of matching between recommended resources and students' learning needs, interest preferences, learning progress and other factors. The higher the

similarity, the more accurately the recommender system can capture students' needs and thus provide more targeted learning resources. In practical applications, the similarity can be indirectly evaluated through students' feedback data such as click rate, browsing length, and satisfaction with the recommended resources. Figure 2 compares the similarity between the vocal teaching resources recommended by the three algorithms and the students' learning needs. The similarity between the resources recommended by this paper's method and the resources of students' learning needs ranges from 0.748 to 0.894. While the similarity of the recommendation method based on knowledge graph is between 0.387-0.602, the similarity of the recommendation method based on students' behavioral characteristics is between 0.364-0.562. From the similarity comparison, the resources recommended by the methods in this paper are more in line with students' needs.

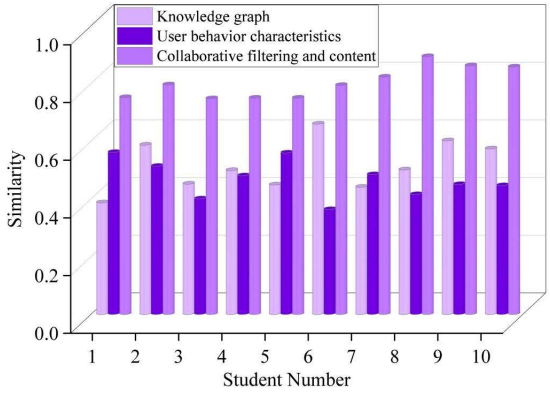


Figure 2: The similarity comparison results of the 3 algorithms

III. B. 2) Resource Recommendation Coverage Comparison

Table 2 shows the resource recommendation coverage of different methods. The resource recommendation coverage of the hybrid recommendation algorithm based on collaborative filtering and content proposed in this paper reaches more than 95% in all 10 iterations, with a maximum of 100%. Among the comparison algorithms, the coverage rate of the knowledge graph-based recommendation method is between 80% and 85%, while the coverage rate of the recommendation method based on student behavioral characteristics is between 75% and 79%. The resource coverage rate of this paper's method is much higher than that of the comparison methods, indicating that the use of a hybrid recommendation algorithm based on collaborative filtering and content for resource recommendation in interactive multimedia-assisted vocal music teaching system can meet students' vocal music learning needs to a large extent.

Table 2: Resource recommendation coverage rates of different methods

Number of iterations	Coverage rate (%)		
	Collaborative filtering and content	Knowledge graph	User behavior characteristics
1	99	81	77
2	98	85	75
3	98	84	78
4	100	81	77
5	100	80	75
6	100	81	78
7	99	85	75
8	99	83	77
9	96	82	79
10	97	81	76

III. B. 3) Comparison of AUC test results

Figure 3 compares the AUC test results of the 3 different methods. AUC is the area under the ROC curve, which is usually taken in the interval of [0.5,1], and the higher the value of AUC, the higher the students' satisfaction with the resource recommendation. As can be seen from the AUC area situation in the figure, the AUC area of this paper's method is greater than 0.9, higher than that of the other 2 methods which is around 0.8 and 0.7, indicating that this paper's method obtains higher student satisfaction in the case of vocal music resource recommendation. This also

reflects that the introduction of the hybrid recommendation algorithm based on collaborative filtering and content in the interactive multimedia-assisted vocal music teaching system can effectively mine students' learning interests and resource needs from their basic learning behavior data, and provide resource recommendation services for students in an accurate and targeted way.

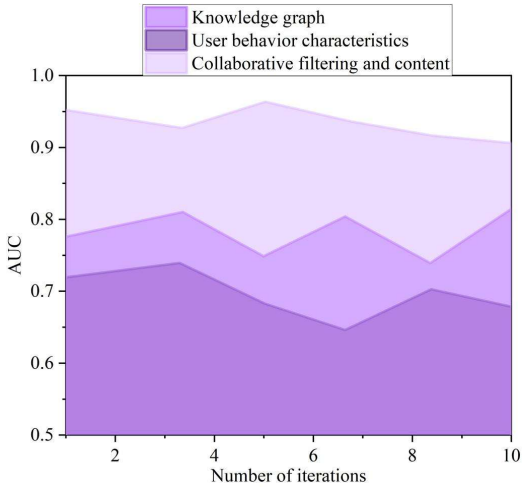


Figure 3: Comparison of AUC test results

III. C. Analysis of the effect of improved CNN-based speech emotion recognition model for vocal singing

III. C. 1) Analysis of Model Recognition Effectiveness

In order to verify the recognition effect of the proposed improved CNN-based vocal singing speech emotion recognition model in the interactive multimedia-assisted vocal teaching system, and to determine whether it can accurately recognize the characteristics of students' vocal singing repertoire. In this paper, the student vocal singing samples are divided into 2 groups, high arousal group and low arousal group, and the model is utilized for vocal feature recognition. Taking the recognition results of note-beat transformations and average velocity transformations of the student vocal singing samples as an example, the violin analysis is plotted with a box-and-line plot inside and a kernel density plot wrapped around the outside.

Figure 4 shows the note-beat transformations of different emotion categories obtained from the recognition. Comparing this note beat map, the beat transformations of the low arousal groups (Q3, Q4) have more dynamics compared to those of the high arousal groups (Q1, Q2), with a total number of beats reaching 300 in both cases, whereas the note beat data distribution of the high arousal groups is more centralized (below 150). There are also some differences in the beat transformations within the same arousal quadrant, e.g., the number of beats in the Q2 quadrant is slightly larger than that in the Q1 quadrant, and the number of beats in the Q4 quadrant is larger than that in the Q3 quadrant.

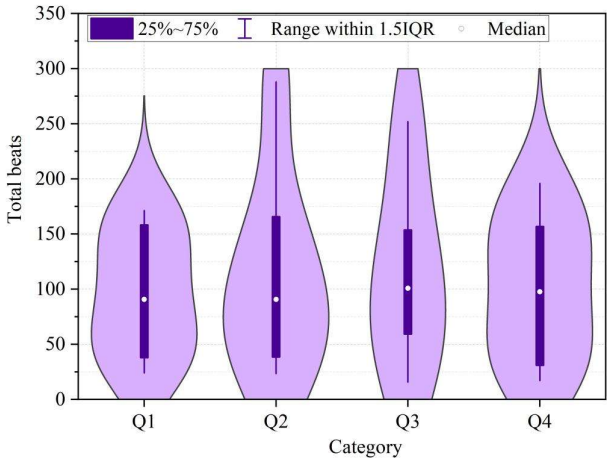


Figure 4: The note beats of different emotional categories change

Figure 5 shows the results of the average velocity transformations of the notes for the different emotion categories obtained from the recognition. The high arousal groups (Q1, Q2) have larger average velocities compared to the low arousal groups (Q3, Q4), with a maximum average velocity of close to 250 in Q1, and a more even distribution of velocities in the low arousal groups compared to those in the high arousal groups. Similarly, some differences can be observed in the average velocity within the same wake-up quadrant, such as a more uniform velocity distribution in the Q1 quadrant compared to the Q2 quadrant, a larger average velocity and a more centralized distribution of average velocity values in the Q4 quadrant compared to the Q3 quadrant, but the dynamics of the velocity distribution in the Q3 quadrant is stronger than that in the Q4 quadrant.

In terms of the actual situation of note recognition by the model, the application of the model in the vocal music teaching aid system can accurately recognize the music sung by the students in practice or formal competitions. The accurate recognition also helps students or teachers to review their vocal performance and find directions for improvement after the performance.

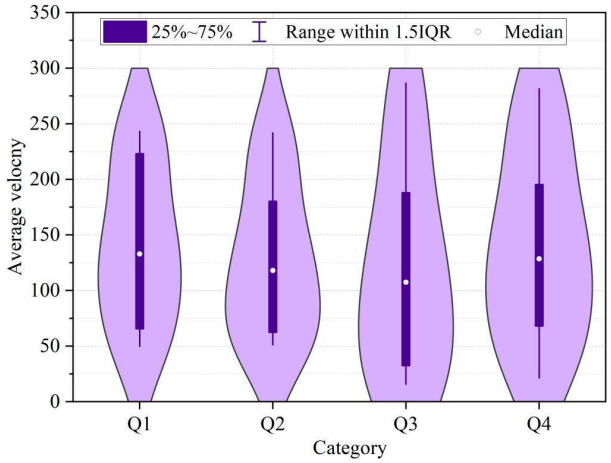
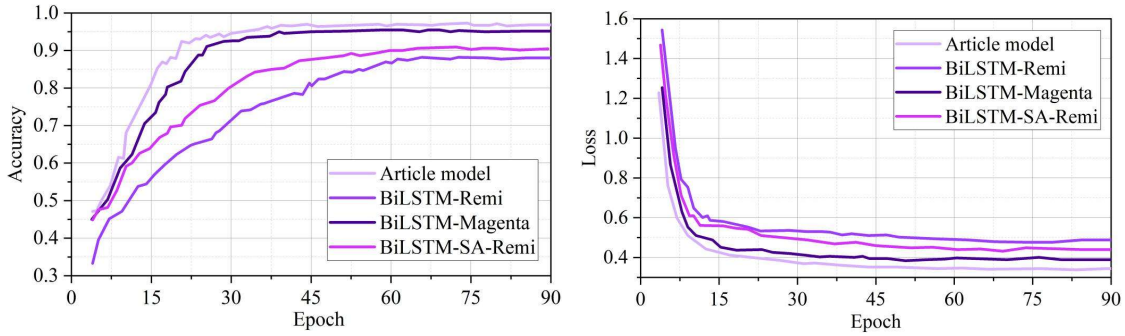


Figure 5: The average speed changes of notes in different emotional categories

III. C. 2) Comparison of model recognition accuracy and experimental loss values

On the premise of analyzing the effect of model identification, further comparison experiments are designed to judge the effect of specific model improvement. Fig. 6 demonstrates the comparison of Acc and Loss changes of multiple multilevel residual-based models. Among them, (a) is the change process of vocal singing emotion recognition accuracy of each multilevel residual-based model with the improved multilevel residual CNN model in this paper. (b) is the change process of their experimental loss values. Before the RMSProp algorithm is introduced to minimize the loss function, this paper's multilevel residual-based CNN model has better performance than the comparison algorithms in both the recognition accuracy and the comparison of experimental loss values. In the recognition accuracy, this paper's model is always higher than the comparison model, and after 45 iterations, it is basically stable greater than 0.95. In the comparison of experimental loss value, this paper's model is also around 45 iterations, i.e., it is stabilized at about 0.4, which is smaller than the comparison algorithm.



(a) Change in recognition accuracy (b) Change of the experimental loss value

Figure 6: Comparison of the changes in model Acc and Loss

The comparison of the recognition accuracy and loss values of the base multilevel residual CNN model and the multilevel residual CNN model after the introduction of the RMSProp algorithm to optimize the loss values of the function is continued. Figure 7 shows the comparison results. After introducing the RMSProp algorithm to optimize the function loss value, the recognition accuracy of vocal singing speech emotion of the improved CNN model in this paper is improved by 0.03 on average, while the experimental loss value decreases by 0.15 on average, which has a more excellent recognition performance.

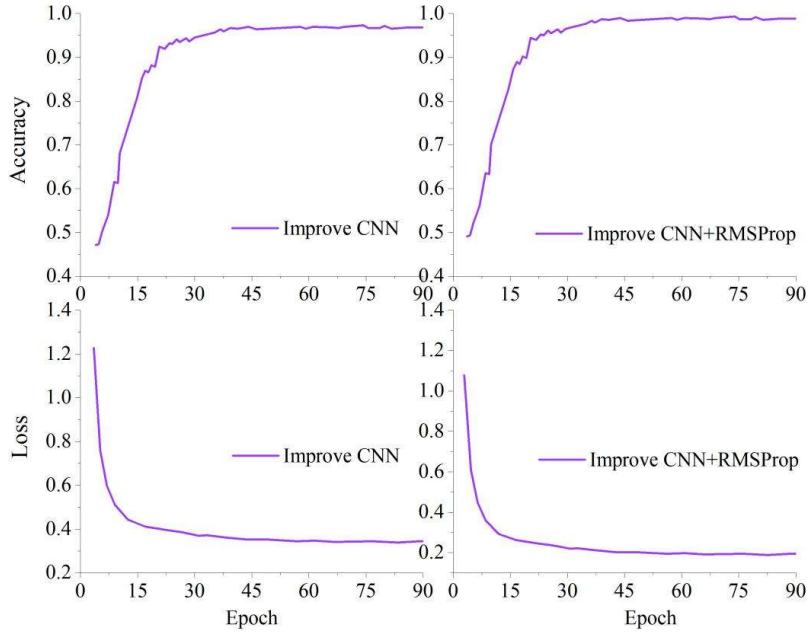


Figure 7: Comparison result

III. D. Analysis of the Impact of Vocal Music Teaching Aid Systems

The optimized interactive multimedia-assisted vocal music teaching system was applied to the freshman class of vocal music majors in the experimental university for one academic year. At the end of the experiment, a questionnaire survey was conducted to collect students' and teachers' evaluations of the impact of the system on vocal teaching assistance. The questionnaire was scored on a six-point TEIP scale, “1 = not at all, 2 = not at all, 3 = less at all, 4 = more at all, 5 = at all, 6 = at all”.

Table 3 provides a general descriptive analysis of the questionnaire results. In the evaluation of the impact of the system on the four aspects of the teaching body, learning resources, learning environment and learning effect, the average value reached more than 4.5, and the median is more than 4.30, while the standard deviation is less than 0.10. It can be judged accordingly that the optimized interactive multimedia-assisted vocal teaching system of this paper is able to facilitate the students' vocal learning from multiple perspectives in the actual teaching of vocal music in colleges and universities and enhance the students' vocal learning effect, and therefore gain the students' support. Vocal music learning effect, so it is recognized by students.

Table 3: Overall description analysis

	Influence of the system on the teaching subject	Influence of the system on learning resources	Influence of the system on the learning environment	Influence of the system on learning outcomes
Sample size	300	300	300	300
Average value	4.52	4.69	4.57	4.69
Median value	4.34	4.85	4.59	4.66
Standard deviation	0.03	0.09	0.09	0.08
Maximum value	6	6	6	6
Minimum value	1	1	1	1

IV. Conclusion

In this paper, we synthesize a hybrid recommendation algorithm based on collaborative filtering of vocal resources and content, as well as a voice emotion recognition model for vocal singing based on improved CNN, to achieve the optimized design of an interactive multimedia-assisted vocal music teaching system. The similarity, coverage and AUC values between the vocal music teaching resources recommendation and students' learning needs of the three algorithms are compared. The similarity of the hybrid recommendation algorithms in this paper ranges from 0.748 to 0.894; the coverage is over 95%, and can reach up to 100%; the AUC area is greater than 0.9. After introducing the RMSProp algorithm to optimize the loss value of the function, the recognition accuracy of the model in this paper reaches 0.95+0.03, and the loss value is 0.4-0.15. In the evaluation of the impact of the system on vocal assisted teaching, the mean of the impacts of the four aspects are more than 4.5, and the median is over 4.30, and the standard deviations are all less than 0.10. Using this paper to design and optimize the system to assist the teaching of vocal music in colleges and universities, good results can be obtained. In the future, a real-time feedback mechanism can be introduced into the system to improve the system's ability to respond instantly to students' learning needs.

References

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

[2] Zheng, D., & Wang, Y. (2022). The application of computer-aided system in the digital teaching of music skills. Computer‐Aided Design & Applications, 19, S7.

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

[4] Guo, S. (2025). Design of Vocal Music Performance Teaching System Based on Multimedia Intelligent Platform. International Journal of High Speed Electronics and Systems, 2540415.

[5] Qin, L., & Zhao, M. (2025). A Multimedia Technology-Driven Vocal Music Teaching System Architecture and Effectiveness Assessment Model Based on Fuzzy Comprehensive Evaluation. J. COMBIN. MATH. COMBIN. COMPUT, 127, 1215-1233.

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

[7] Long, Z., Yap, J. H., & Koning, S. I. (2024). Harmonising Tradition and Technology: A Review of Multimedia Integration in Guizhou's Vocal Music. Asian Pendidikan, 4(1), 14-22.

[8] Zhihong, L. (2019). Exploration on the Application of Multimedia Technology in Vocal Music Teaching. In 1st International Education Technology and Research Conference (pp. 628-631).

[9] CAO, J. (2024). Enhancing student engagement and learning outcomes through interactive approaches in college vocal music instruction. Pacific International Journal, 7(1), 131-136.

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

[11] Li, L., & Han, Z. (2023). Design and innovation of audio iot technology using music teaching intelligent mode. Neural Computing and Applications, 35(6), 4383-4396.

[12] Bautista, A., Tan, C., Wong, J., & Conway, C. (2019). The role of classroom video in music teacher research: A review of the literature. Music Education Research, 21(4), 331-343.

[13] Jiang, Y. (2022). MOOC vocal music blended teaching mode and digital resource platform design in the information age. In 2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City: Volume 2 (pp. 701-707). Springer Singapore.

[14] Xu, C., & Zhai, Y. (2022). Design of a computer aided system for self-learning vocal music singing with the help of mobile streaming media technology. Computer-Aided Design and Applications, 19(S3), 119-129.