International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5685-5698

https://doi.org/10.70517/ijhsa464408

Research on Emotional Expression Recognition and Enhancement Strategies in Vocal Performance

Jinwei Zhang[1,*]
[1] Taizhou College, Taizhou, Jiangsu, 250200, China
Corresponding authors: (e-mail: temporubato11088@163.com).

Abstract Emotional expression analysis in vocal performance is of great significance for enhancing artistic expression, and the rapid development of artificial intelligence technology provides a new solution path for music emotion recognition. This paper proposes a strategy for analyzing and enhancing emotional expression in vocal performance based on pattern recognition, and constructs a multimodal music emotion classification model based on optimized residual network. Methodologically, Mel spectrum is used for audio signal preprocessing, GloVe word vectors are utilized to represent the lyrics text, the model performance is enhanced by teacher-student modeling and transfer learning, the ResNet50 network structure is improved and an improved Center-Softmax classifier is introduced. Experimental results on the classical piano dataset show that the proposed algorithm achieves 88.34% emotion recognition accuracy, which is a 2.43% improvement compared to the XGBoost algorithm, and the recall and F1 values reach 83.34% and 86.52%, respectively. In the Chinese folk song multi-emotion recognition experiment, the recognition accuracy of the multimodal fusion model reaches 85.62%, which is 6.15% and 4.19% higher than the unimodal model, respectively. The vocal performance visualization analysis verifies the effectiveness of the model in ancient poetic art songs such as Guan Ju. The experiments proved that the method can effectively recognize multiple emotional states in vocal performance, providing scientific technical support for vocal teaching and performance evaluation.

Index Terms Pattern Recognition, Vocal Performance, Emotional Expression, Multimodal, Residual Networks, Classification Models

I. Introduction

Vocal performance art is a performing art with voice as the medium and emotion as the core [1]. It not only requires the singer to have exquisite vocal skills, but also requires the singer to be able to present the artistic charm of the work in front of the audience with full and sincere emotions and vivid and evocative performances [2], [3]. Vocal performance, role modeling and emotional expression are two core elements, which are interdependent and complement each other, and together they build up a magnificent temple of vocal art [4], [5].

Emotion is the soul of music and art that can be celebrated for thousands of years, and it is also the source of life of vocal performance [6]. Music relies on sound to expand the dissemination of this art can not rely solely on the composer, lyricist can be completed, but also need to singers through the second degree of creativity to find the heart of the lyrical point of view, will be this emotion, art conveyed to the listener [7], [8]. As the world-renowned musician Beethoven said, the carrier of music is not the score but the emotion between the scores, which guides the singers to express their emotions and shape their image in the singing process. On the one hand, characterization is the basis and premise of emotional expression, the singer only accurately grasp the character image portrayed in the work, in order to really dig out the emotional connotation of the work, so that the emotional expression is more sincere and moving [9]. On the other hand, emotional expression is the soul and life of the characterization, and only when the singer devotes himself to the character can he make the character more vivid and full of vitality [10]-[12]. Therefore, it is of great significance to explore the internal rules and practical paths of characterization and emotional expression in vocal performance to enhance the artistic infectivity and aesthetic value of vocal performance [13], [14].

The core of this study is to construct a framework for analyzing and enhancing strategies of emotional expression in vocal performance based on pattern recognition. Firstly, to address the multimodal characteristics of vocal performance, a feature extraction method is designed to fuse audio signal and lyrics text, using Mel spectrogram to represent audio features and word vector technique to process lyrics text information. Secondly, based on the advantages of residual network, an optimized deep learning model is constructed to improve the accuracy of sentiment classification by improving the network structure and loss function. Finally, through the teacher-student

model and migration learning technology, the problems of insufficient labeled data and limited model generalization ability are solved to realize the migration from coarse-grained to fine-grained emotion recognition, and to provide scientific analysis tools and technical support for the expression of emotion in vocal performance.

II. Multimodal music sentiment classification model based on optimized residual network

In order to realize the analysis and enhancement of pattern recognition-based emotion expression in vocal performance, this paper preprocesses and enhances music audio and constructs a multimodal music emotion classification model based on optimized residual network.

II. A. Data pre-processing

II. A. 1) Audio signal representation and pre-processing

Mel spectrum is a commonly used signal representation in audio classification tasks, which retains the characteristics of music signals more completely than other advanced audio signal representations, and at the same time, Mel spectrum is more in line with human auditory characteristics, so in this paper, Mel spectrum is selected as the input data for music audio analysis.

Mute detection (VAD), that is, to detect the presence of mute frames in the music signal, the presence of mute frames in these parts affect the results of the recognition.

The music audio signal representation and preprocessing flow is shown in Figure 1.

Figure 1: Music audio signal representation and preprocessing process

II. A. 2) Word Vector Pre-Representation of Lyric Texts

The lyrics of the songs in the dataset studied in this paper contain not only Chinese but also other foreign languages, in order to be compatible with lyrics of different languages, this paper cuts the foreign lyrics by special characters and the Chinese lyrics by words, which avoids the complex Chinese word-cutting algorithm. The word vectors are represented in a non-static way, with the pre-trained GloVe word vectors as the initialized lyrics word vectors, and the word vectors are adjusted during the model training process. In this paper, the $word\_dim$ of the word vector is set to 128, and the maximum length of the lyrics max_length is set to 180. The word vector pre-representation process of the lyrics text is shown in Fig. 2.

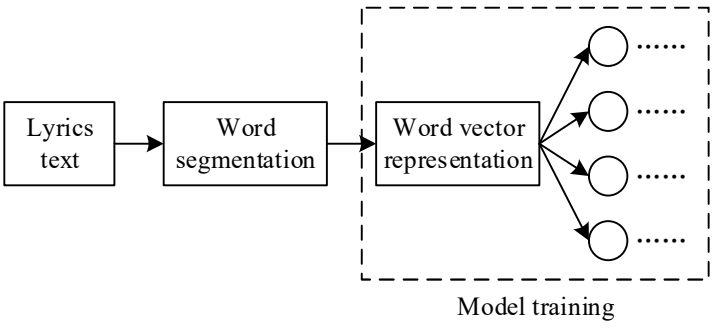


Figure 2: Pre-representation of lyrics word vectors

II. B. Teacher-Student Model and Transfer Learning Approach

II. B. 1) Teacher-student model

In the case of small and uneven labeled data, this paper uses the teacher-student model to improve the accuracy of music emotion recognition. The teacher-student model adopts the existing network architecture for music style recognition, and adopts the network parameters at different stages of style recognition to do migration learning respectively. At the same time, following the idea that the reasoning performance of the teacher network is usually better than that of the student network, the teacher model in this paper utilizes the parameters of the tune network better than those of the student network.

Migration learning is generally used in two ways for knowledge transfer. One, the pre-trained curvilinear network

structure is used as a feature extractor for emotion recognition, and only the part of the emotion recognition network added after training. The second, the trained tune network structure is involved in the training together with the added networks. In this paper, the 2nd approach is adopted.

The teacher network does not participate in the backpropagation of the neural network, and the parameter (W) of the teacher model is obtained from the parameter (w) of the student model by exponential sliding averaging (EMA), and the expression for the parameter of the teacher model at moment t is as follows:

$$W_t = \alpha \times W_{t-1} + (1-\alpha) \times w_t \tag{1}$$

where, α represents the decay smoothing rate, and w_t represents the parameters of the student model at moment t.

The loss function f_{Loss} of the teacher-student model consists of two components, the relative entropy, i.e., the KL dispersion and the categorical cross-entropy (CE), and λ controls the importance of the relative entropy:

$$f_{Loss} = H(\ ) + \lambda KL(\ ) \tag{2}$$

For the same sample X, the output s is first obtained through the student model, and then the output t is obtained through the teacher model, the KL loss is calculated using t and s, the gradient is passed back to the student model, the CE loss is calculated for the data in s, and based on the two parts of the loss function, the parameters of the student model are updated, and the parameters of the teacher model are obtained through the exponential sliding averaging of the student model parameters.

In each iteration of the training process, the input samples X contain both samples with sentiment labels and samples without labels, and at the same time these samples are augmented to improve the generalization ability of the sentiment data.

II. B. 2) Data enhancement

(1) Gaussian noise

Songs are accompanied by various noises during recording and propagation. In this paper, in the audio preprocessing section, Gaussian noise is used for audio noise enhancement for all the audio.

(2) Audio Cropping

It is assumed that based on different song clips, humans can still judge the emotion of the song. In this paper, the average length of audio cropping is 20s.

(3) Audio Mixing

Random mixing of different audio. Assuming a pair of audio samples (X_a, X_b), the generated mixing samples are X_{com}, and the generated X_{com} is used to compute the KL loss only, and the mixing of the sentiment labeling dimension is not considered. Mixing of label dimensions is also done to generate mixed label data Y_{com} and used to compute CE loss:

$$\begin{aligned} X_{com} &= (1-\gamma) \times X_a + \gamma \times X_b \\ Y_{com} &= (1-\gamma) \times Y_a + \gamma \times Y_b \end{aligned} \tag{3}$$

where, γ is the sample mixing coefficient.

II. C. Optimized residual network model

II. C. 1) Deep residual networks

In response to the problem that deep network training will have the problem of gradient vanishing or gradient dispersion, related scholars have proposed a residual-based deep learning framework, DeepResNet [15], which is able to improve the accuracy rate by increasing the depth of the network, and at the same time, it adopts the residual blocks for jumping connection to construct the network structure to solve the problem of degradation of the performance, and the basic idea is that when constructing a convolutional neural network The basic idea is to fit a residual mapping by adding shortcut connection branches to form a basic residual learning unit and utilizing stacked nonlinear convolutional layers.

Deep residual networks, as an extremely deep convolutional neural network framework, show good properties in terms of accuracy and convergence, etc. ResNet consists of many residual units, and each residual block, as shown in Fig. 3, can be represented as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{4}$$

$$x_{l+1} = f(y_l) \tag{5}$$

$h(x_l) = x_l$ represents a constant mapping, in the forward and back propagation phases of training, the signal can be passed directly by jumping, which neither introduces new parameters nor increases the computational complexity, but makes the training simpler, thus solving the problem of difficult training and performance degradation of deep networks.

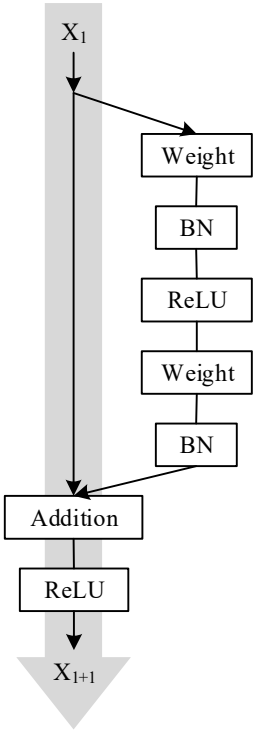


Figure 3: Residual block structure of the residual network

ResNet networks for shallow layers, such as ResNet18/34, have the same time complexity as ResNet networks for deep layers, such as ResNet50/101. The whole structure is generally referred to as a “building block”, and the structure of the deep ResNet network is specifically referred to as a “bottleneck design”.

Considering the experimental configurations and computational power, the ResNet50 network structure was finally chosen [16], and the bottleneck architecture uses a 3-layer stack, with the 1 × 1 convolutional layers in the first and third layers used to recover the dimensionality, and the 3 × 3 convolutional layers in the middle becoming the bottleneck with small dimensionality in order to reduce the number of parameters, which is nearly 16 times different compared to the shallow ResNet network structure.

II. C. 2) Optimizing the residual block structure

In this paper, we refer to the decomposition diagram of ResNet structure in related literature, and optimize and improve the ResNet50 network model to make it more suitable for the research in this paper. The schematic diagram of the original ResNet50 network architecture is given first, as shown in Fig. 4.

The ResNet50 model commonly uses a 7 × 7 convolutional kernel for network input, which is computationally large. Meanwhile, when the input dimension of the residual block in ResNet50 is not the same as the output dimension, the common method is to increase the dimension with a 1 × 1 convolutional kernel with a step size of 2 so that the input and output dimensions are the same, as shown by the content in the dotted box in Fig. 4. However, for the classification problem of fine granularity images, most of the redundant information will be lost when a convolutional layer with a step size of 2 is chosen, and 3/4 of the feature points are not involved in the computation, which will adversely affect the final computation results, and reduce the credibility of the feature information to a certain extent.

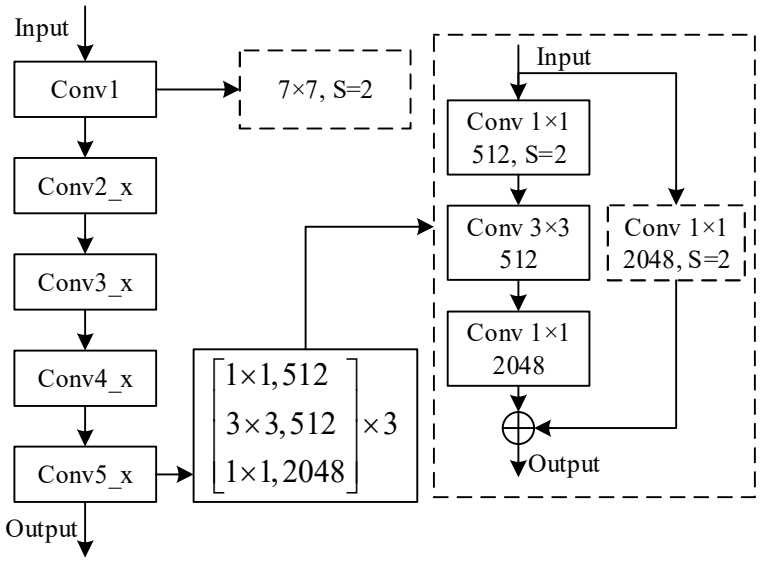


Figure 4: The original ResNet50 network structure

In this paper, the model is optimized and improved from the above 2 problems, and the improved ResNet50 network architecture is shown in Fig. 5.

First, the Inceptionv2 idea is borrowed to replace part of the 7×7 input convolution kernel with three 3×3 small-sized convolution kernels, which reduces the number of parameters involved in the computation while maintaining the same sensory field of the convolution kernels, thus reducing the amount of computation. For a 7×7 convolution, the total number of parameters used is 49channels, while the total number of parameters used with three 3×3 convolutions is 27channels, which significantly reduces the number of parameters and shortens the computation time. In order to retain as many parameter points as possible to participate in the computation, the convolution of 1×1 with step size 2 is replaced by the convolution of 1×1 with step size 1 on the shortcut path of the residual block, and a mean-pooling layer of 2×2 with step size 2 is added in front of this convolutional layer in order to retain the gradient and make the input and output dimensions consistent. Such an improvement will also lose some information, but compared to the convolution layer with a step length of 2 1×1, this method first undergoes selection and then loses the redundant information, and each feature point is involved in the calculation, so that it can retain most of the information of the feature points, which can make up for the loss of information of the original structure to a certain extent.

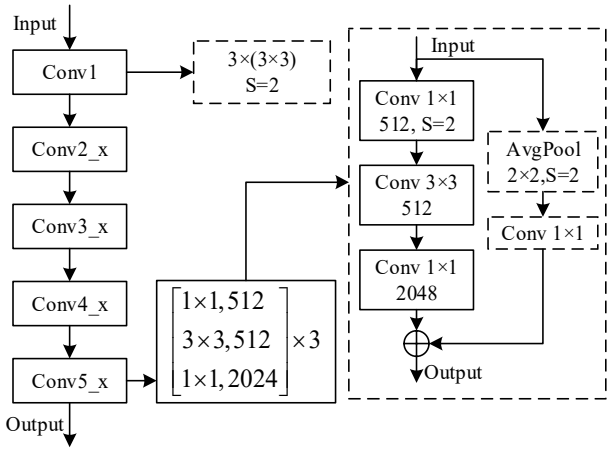


Figure 5: The optimized ResNet50 network structure

II. D.Improved Softmax Classifier

II. D. 1) Softmax Classifier

The Softmax function is often used by many researchers in usage scenarios where multiple classifications are implemented using deep learning. The Softmax function maps the extracted feature inputs to [0,1] and guarantees that the sum is 1 by a normalization operation. Softmax is of the form:

$$p_{y_i}(x_i) = \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{-i\omega t}} \tag{6}$$

where W_{y_i} and b_{y_i} are the weights and biases of the last fully connected layer corresponding to class y_i, and n is the number of classes. The cross-entropy function is often chosen as the objective function for multicategorization problems, i.e:

$$L_{cross} = -\sum_{k} t_k \log P(y=k) \tag{7}$$

In classification problems with convolutional neural networks, one-hot encoders are often used to process the predicted classes, and the current generalized Softmax function is to nonlinearly scale each input x to $\exp(x)$ of the form:

$$p = \frac{e^{x_i}}{\sum_{j} e^{x_j}} \tag{8}$$

As can be seen through the formula (6)~(8), the Softmax function will separate different categories of features, and there will be a certain distance between different categories, but the distance will remain unchanged after approaching a certain degree, so it is often the case that the distance between the same category may be greater than the distance between different classes. For the music emotion classification studied in this paper, the above problems also exist, so the Softmax classification function needs to be improved.

II. D. 2) Introducing the Centerloss function variant

In order to ensure that the classification model has the characteristics of intra-class convergence and inter-class separation, in recent years there are some scholars who have made corresponding improvements to Softmax, and the more commonly used ones are Angular-Softmax and Center-Softmax.

The idea of Angular-Softmax is to convert the separation characteristics between sample features into angular boundary learning, the specific formula is:

$$L_M = -\log\left(\delta g(x) e^{L_\theta} \sum_{j} e^{L_\theta}\right) \tag{9}$$

Among them:

$$L_\theta = \|W_{y_i}\| \|x_i\| \cos(\theta_{y_i}) + b_{y_i} \tag{10}$$

Angular-Softmax is mentioned to normalize its weights to $\|W\| = 1$ and make the bias 0. Finally, the effectiveness of this function is verified by face recognition experimental analysis. However, these features are still not well recognized, and the enhancement effect is limited with the increase of data volume.

The idea of Center-Softmax is to minimize the intra-class spacing, and control the feature center by introducing Center loss, the specific formula is:

$$L_{cs} = L_s + \lambda L_c = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_{y_i}^T x_i + b_{y_i}}} + \frac{\lambda}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \tag{11}$$

where c_{y_i} represents the feature center of category y_i, which will change with the change of features, and m represents the size of mini-batch to update the feature center. For the emotion image studied in this paper belongs

to the image classification of fine granularity, so we hope that the final classification model is intra-class aggregation and inter-class separation. While Center-Softmax only considers intra-class centering, there is still room for improvement.

This paper considers:

(1) The shortest distance from the training samples to the class centroids.

(2) The sum of the distances between the training samples and their non-corresponding category centers is maximum.

The Center loss function is improved by introducing the distance of non-corresponding classes, and on the basis of controlling the centroids of the same class, the distance between the centroids of different classes is guaranteed to be the largest as far as possible. The improved Center loss function formula is:

$$L_{c/s} = \frac{1}{2}\sum_{i=1}^{m}\frac{\left\|x_i - c_{y_i}\right\|_2^2}{\left(\sum_{j=1, j\neq y_i}^{k}\left\|x_i - c_{y_j}\right\|_2^2\right)+1} \tag{12}$$

The addition of 1 to the denominator is to prevent the denominator from appearing to be 0. The final representation of the improved classification function L_Center_Softmax in this paper is:

$$L = L_s + \lambda L_{c/s} = -\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_{y_i}^T x_i + b_{y_i}}} + \frac{\lambda}{2}\sum_{i=1}^{m}\frac{\left\|x_i - c_{y_i}\right\|_2^2}{\left(\sum_{j=1, j\neq y_i}^{k}\left\|x_i - c_{y_j}\right\|_2^2\right)+1} \tag{13}$$

The improved loss function combines the center loss and interclass distance, which can improve the discriminability of the features and ensure that the intra-class spacing of the features is reduced while increasing the differentiation between different classes to enhance the adaptability of fine-grained image classification.

III. Experiments and analysis of results

In this chapter, the proposed multimodal music emotion classification model is experimentally validated to ensure its effectiveness in the analysis of emotional expression in vocal performance.

III. A. Algorithm comparison experiment

III. A. 1) Experimental environment and dataset

The experimental environment for executing the proposed algorithm is shown in Table 1. In order to validate the algorithm of this paper, the world's largest classical piano dataset, GiantMIDI-Piano, which has a total of 10,854 classical piano songs, is used. Meanwhile, in order to get the real emotion labels, Python is used to write scripts and crawl 4000 piano songs with emotion label classification from the above dataset as the training and validation sets of the algorithm, and the ratio of the number of the two is 7:3. Since a piece of music may contain multiple emotions, the music is also segmented according to the data preprocessing step, and the time of each segment is 20 s. The emotion labels were categorized into 4 categories: calm, agitated, pleasant and pathos.

Table 1: Experimental environment parameters

Parameters	Configuration
CPU	Intel Core i7 10900K
GPU	NVDIA GTX 1660 Super
Memory /GB	256
Software version	Python 3.13.1

III. A. 2) Algorithm testing

The essence of this paper's algorithm for recognizing the emotion of music is a four-classification problem, so the experimental test indexes chosen for this experiment are accuracy, recall and F1 value. At the same time, in order to better verify the performance of the algorithm, this experiment also uses RF, XGBoost, LSSVM [17] and BP neural network as comparison algorithms. The preprocessed data is input into the model to be trained, and 30 piano music frames are randomly selected for 15 tests and averaged, and the classification results of the different

algorithms obtained are shown in Fig. 6.

It can be seen that the emotion recognition accuracy of this paper's algorithm is optimal among all the algorithms in the experimental tests, which indicates the good performance of the algorithm.

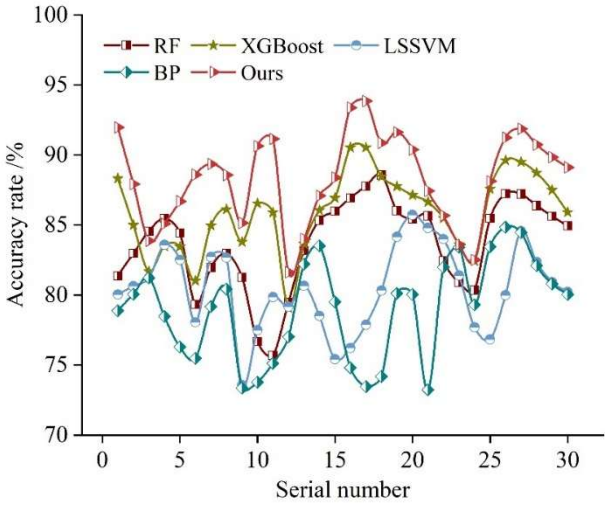


Figure 6: Recognition results of different algorithms

In addition to the accuracy metrics, the algorithm's recall and F1 value should also be tested, and the results of the metrics testing of different algorithms are shown in Table 2.

Analysis shows that the average accuracy of this paper's algorithm is 88.34%, which is 2.43% ahead of the XGBoost algorithm. Meanwhile, the recall and F1 value of this paper's algorithm are also the highest among the comparison algorithms. This proves that the algorithm with the addition of deep residual network and improved Softmax classifier can have more superior classification performance, ahead of other algorithms, which further illustrates the effectiveness and robustness of the model.

Table 2: Test results of indicators for different algorithms

Algorithm	Average accuracy rate /%	Recall rate /%	F1 value /%
RF	83.72	72.63	77.65
XGBoost	85.91	79.46	80.31
LSSVM	80.44	74.52	77.48
BP	79.03	71.47	75.29
Ours	88.34	83.34	86.52

In addition to the binary classification test, this experiment also conducted a four classification experiment on the model to verify the algorithm's ability to classify different results. The comparison algorithm is selected from the same type of algorithm XGBoost, and the results of the four classification experiment are shown in Table 3.

It can be seen that the two types of music recognition accuracy is higher in the two types of music, the reason is that the two types of music with a strong sense of rhythm, waveform file features are more obvious compared to the other two types. In addition, the algorithm in this paper is ahead of XGBoost in all indicators, which shows that its multi-classification performance is also superior.

Table 3: Test results of the four-classification experiment

Type	The algorithm of this article			XGBoost algorithm		
	Accuracy /%	Recall /%	F1 /%	Accuracy /%	Recall /%	F1 /%
Calm	77.94	78.41	77.94	74.62	76.85	75.41
Passionate	88.73	81.25	85.46	85.34	78.76	83.85
Pleasant	84.52	77.36	82.87	83.81	76.32	79.62
Sad and sorrowful	88.19	83.43	86.78	87.43	82.79	85.53

III. B. Experiments on Multi-Emotion Recognition in Chinese Folk Song Performance

In this section, the proposed multimodal music emotion classification model is used to identify the emotion of Chinese folk song performances and analyze the change rule of emotion in them.

III. B. 1) Sentiment label mapping results

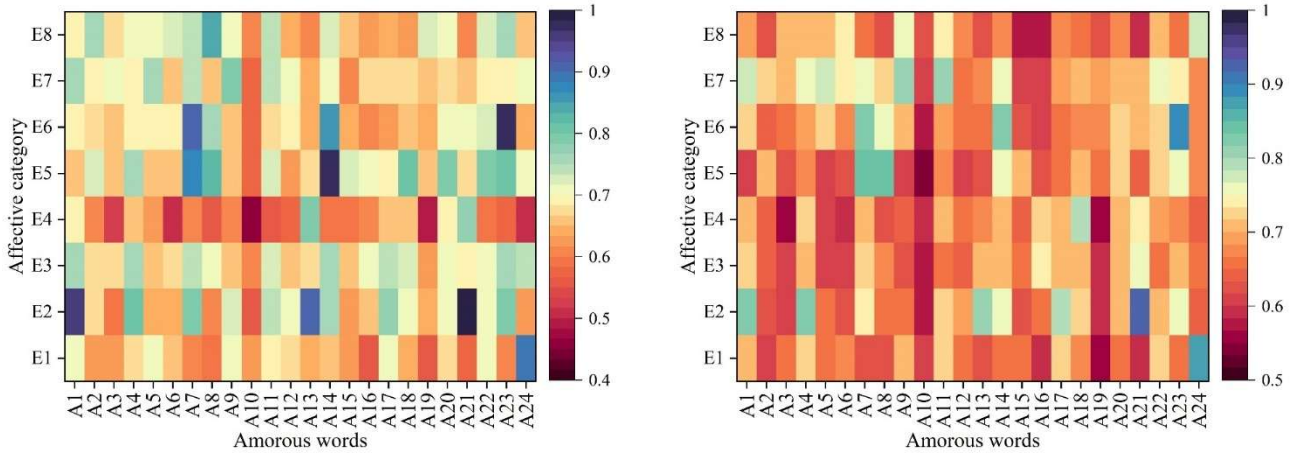
The folk song data in this paper comes from the “Songs and Music” section of the Chinese National Music Resource Database, which mainly includes lyrics, audio, emotion tags and other information. The 24 emotion words in the original data are used as the F-O dictionary, and the first and second level emotion words of the F-O dictionary are semantically enhanced to obtain the vector representation, and the cosine similarity is used as the semantic similarity to carry out the mapping experiments. Two mapping schemes, FO-L1 and F-O-L2, are used and the mapping effects of different schemes are compared. The mapping process of the scheme is as follows:

(1) F-O-L1 mapping scheme: for any sentiment word $word\_O$ in the F-O dictionary, respectively, and the first-level sentiment word $word\_L1$ under 8 sentiment categories to calculate the vector cosine similarity, if $word\_L1$ satisfies $sim(word\_O, word\_L1)$ is the maximum and not lower than the similarity threshold θ, then the mapping label of $word\_O$ is $word\_L1$, otherwise the mapping label is "Other".

(2) F-O-L2 mapping scheme: the vector cosine similarity is calculated for $word\_O$ and the second-level sentiment words $word\_L2$ under 8 sentiment categories, and the average similarity with all second-level sentiment words under each category is taken as the similarity between the label and the category, if there is an emotion category that satisfies the $Average(sim(word\_O, word\_L2))$ maximum and not less than θ, then the mapping label of $word\_O$ is the first-level sentiment word $word\_L1$ corresponding to the sentiment category, otherwise the mapping label is "other".

The preliminary calculation results of the F-O-L1 and F-O-L2 mapping schemes are shown in Fig. 7. Among them, (a) and (b) denote the calculation results of F-O-L1 and F-O-L2, respectively, A1~A24 denote the 24 emotion words of grief, exuberance, roughness, low, high-pitched, heroic, cheerful, lively, intense, firm, tense, vast, poignant, light, clear, soft, low, soothing, jumping, euphony, sadness, lilting, delightful, and solemnity, and E1~E8 respectively E1~E8 represent 8 emotional categories: sacred, sad, yearning, lyrical, light, happy, passionate, and lively.

It can be found that the similarity value of the F-O-L1 mapping scheme polarizes significantly and gives a clearer answer in label proximity judgment. Since the cosine similarity takes the value range of [0,1], and the mean value of similarity is calculated to be 0.67, the range of θ is initially determined to be among the four values of 0.5, 0.6, 0.7, 0.8, and the results of label categorization under the four values are compared and analyzed. When θ = 0.5 or 0.6, all original labels will have 2 and more mapped labels that exceed the threshold. When $\theta = 0.8$, more than half of the original labels will be mapped to the “other” class. The above is a less ideal situation, .when $\theta = 0.7$, it can better balance the above problems, so we take 0.7 as the label mapping similarity threshold to complete the mapping process under both mapping schemes.



(a) F-O-LI mapping scheme (b) F-O-L2 mapping scheme

Figure 7: Preliminary calculation results of the two mapping schemes

The final mapping results are shown in Table 4. The differences in the results of the two mapping schemes are mainly in three aspects: the degree of lexicon full mapping, the label mapping results, and the exception sentiment processing:

(1) In terms of lexicon full mapping, not every generic emotion word has a corresponding original label under the

F-O-L1 scheme, such as yearning and lyricism, while every generic emotion word has at least one original label corresponding to it under the F-O-L2 scheme. Chinese folk songs have distinct and strong emotions, and the emotion pointing of yearning and lyricism is too broad and does not specify what to yearn for and what kind of emotion to express, which is not applicable to Chinese folk songs, so the mapping result of F-O-L1 is more reasonable.

(2) In the label mapping results, there are 10 words with different attributions under the two schemes, and according to the Chinese context, the mapping results of words such as rough, lively, and jumping are more in line with the intuitive emotional cognition under the F-O-L1 scheme.

(3) In the exception sentiment processing, F-O-L1 thinks that firm and soft should be categorized as “other”, and F-O-L2 thinks that firm and vast should be categorized as “other”. Combined with Fig. 7, the similarity between the above words calculated by F-O-L2 and the generic words is very close to that of the generic words. Combined with Fig. 7, the similarity between the above words calculated by F-O-L2 and each generic word is very close and lacks differentiation, in contrast, the judgment result of F-O-L1 is more explicit. Based on the above reasons, this paper believes that the F-O-L1 mapping scheme is more reasonable, and therefore this paper chooses the F-O-L1 results as the final label mapping results.

Table 4: Final mapping results of the original sentiment labels

Serial number	Emotion	F-O-L1 mapping result	F-O-L2 mapping result
1	E1	A24	A24
2	E2	A1, A4, A13, A17, A21	A1, A4, A13, A17, A21
3	E3	/	A16
4	E4	/	A18
5	E5	A7, A14, A15, A17, A20, A22	A8, A15, A20
6	E6	A23	A7, A14, A23
7	E7	A3, A5, A9, A11, A12	A2, A5, A9, A11, A19, A22
8	E8	A2, A6, A8, A19	A3, A6
9	Other	A10, A16	A10, A12

III. B. 2) Folk Song Emotion Recognition Results

Based on the F-O-L1 mapping scheme, the 24 emotion labels in the original data have been mapped to sacred, sad, light, happy, passionate, vitality, and other 7 general emotion words, at this time each folk song in the dataset has 1-2 emotion labels, and the specific distribution of the number of labels after mapping is shown in Figure 8. Compared with the original distribution, the label semantics is more concentrated and the number distribution is more balanced, and there are 109 songs under the least number of sacred emotion, which enables the model to learn more useful information in the subsequent automatic emotion recognition task.

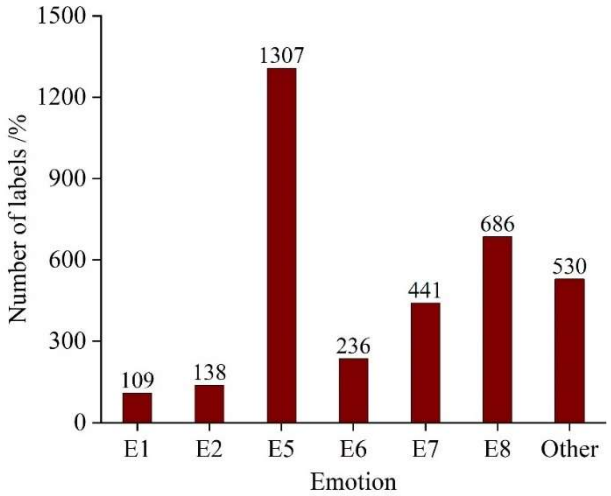


Figure 8: The distribution of label quantity after mapping

Using the multi-sentiment recognition model based on the improved ResNet50 network, experiments were

conducted on the unimodal model based only on lyrics and audio features and the multimodal fusion model respectively, and the parameters in the training phase were updated using the Adam optimizer. In the Sigmoid function normalization stage, the label prediction probability threshold was set to 0.6, and all labels with output probability > 0.6 were used as the emotion recognition results for the corresponding folk songs. The 3447 folk songs were divided into training and testing sets according to 8:2.

The results of the ablation experiments are shown in Table 5. The HL value indicates the percentage of incorrectly predicted labels among all labels, and a lower value indicates a better model performance. Comparing the effect of using only text features and only audio features, the model using only text features has better performance in Accuracy, Recall, F1, and HL metrics, especially in Recall metrics, which is 7.92 percentage points higher than the model using only audio features. This shows that for folk songs, the emotional information contained in the lyrics is more important than the audio. Comparing the unimodal model with the multimodal model, it can be found that the multimodal model has a significant improvement in most of the indicators, and the recognition accuracy reaches 85.62%, which is 6.15% and 4.19% higher than that of the textual unimodal and audio unimodal, respectively. Compared with single modality, multimodal fusion provides more effective information for the model, so the performance improvement is obvious, which further proves the effectiveness of this paper's model on the task of multi-emotion recognition of folk songs.

Table 5: Ablation experiment results

Evaluation indicators	Text	Audio	Text +Audio
Accuracy	0.7062	0.6749	0.7214
Precision	0.7947	0.8143	0.8562
Recall	0.7946	0.7154	0.7497
F1	0.7947	0.7617	0.7994
HL	0.1184	0.1395	0.1141

Based on the consistency between complete folk songs and fragmented folk songs in the text and audio feature space, the trained improved ResNet50 model is migrated to the fragment emotion recognition task to provide a solution to the problem of labeling difficulty and hard to start in fine-grained emotion recognition. The 3447 folk songs are segmented into 720 song fragments according to the lyric line cut sign, and the fragment emotion labels are identified using the improved ResNet50 model, from which 200 song fragments are randomly selected for manual label annotation, and analyzed in comparison with the results of automatic recognition. The evaluation indexes of the model on fragment data are shown in Table 6, which shows that the model trained on coarse-grained songs can effectively recognize the emotion of fine-grained fragment songs, which corroborates the reasonableness of the strategy of using model migration in this paper.

Table 6: Evaluation of model migration effect

Evaluation indicators	Take the value
Accuracy	0.8045
Precision	0.8234
Recall	0.8369
F1	0.8301
HL	0.0698

III. C. Visual Analysis of Vocal Performance

In this section, we analyze the Rubato processing of the ancient Chinese poetic art song “Guan Ju” by combining the score example, through the visualization charts such as the velocity-intensity curve and the IOI deviation curve, to provide a typical case for the visualization research of vocal performance.

III. C. 1) Visualization of speed-strength curves

Guan Ju is a song in D-feather seven-tone mode, a single three-part form with reproduction, an introduction, a connection and a coda. The analysis of Guan Ju is shown in Table 7. The track as a whole is soothing and gentle, narrative and emotional is very strong, different singers for the interpretation of this track in the speed and intensity have their own treatment.

Table 7: Analysis of the musical form of *Guan Ju*

Paragraph nature	Prelude	A				B			Connect	A'		Epilogue
Structural properties	Prelude	a	a'	b	b'	c	c'	d	Connect	a	a	Epilogue
Number of sections	9	4	4	6	5	4	5	4	5	4	4	4
Subsection positioning	1-9	10-13	14-17	18-23	24-28	29-32	33-37	38-41	42-46	47-50	51-54	55-58

In order to learn from it, this paper imports the singing audio into Vmus.net platform and analyzes it, and comes up with the speed and intensity analysis graphs of the versions of Shi Yijie, Song Zuying and Zhou Shulin. Taking Shi Yijie's version as an example, the speed-intensity curve of this version is shown in Figure 9. Shi Yijie sang the whole song almost according to the notation of strength and weakness in the score, but the black line, which should be smoother, fluctuates constantly, and he did not sing strictly according to the tempo in the score. The same is true of the other two singers' versions. The other two singers' versions are also similar. It is clear that the singers have added free play to lengthen and shorten the tempo according to their own understanding of the repertoire, which is the free tempo rubato in the singing treatment of vocal works.

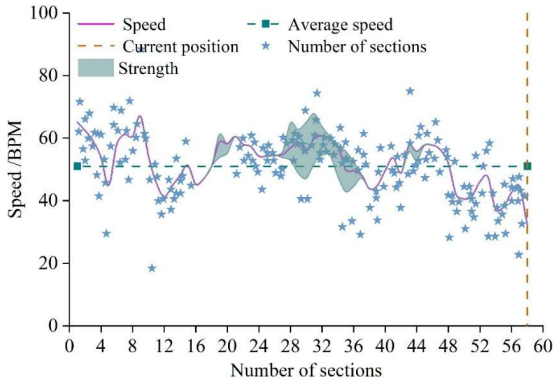


Figure 9: Shi Yijie's singing version of the speed-dynamics curve

III. C. 2) Visualization of free velocity Rubato

In order to understand more intuitively and in more detail on what principle the three singers used Rubato interpretation to complete the song “Guan Ju”, this paper imported some of the audio of their singing into Vmus.net for the computerized analysis of the IOI deviation of the starting interval. The IOI deviation curves of bars 10-17 of Shi Yijie's version are shown in Figure 10.

(1) Based on emotion, lexis and sentence comma

Measures 10-17 of the main song section, i.e., a and a' of the A section, are two crucial lines that set the emotional tone of the whole song. Observing the IOI deviation curves of Shi Yijie's version of bars 10-17, most of the notes have very obvious “temporal expansion”, reaching a maximum of +86% and a minimum of -34%. Song Zuying's and Zhou Shulin's singing processing rules are very similar to Shi Yijie's, which can be summarized in two main points.

One, when the IOI deviation is negative, it is the Rubato processing of shortening the time value. During the singing process of the three singers, the sentences with IOI deviation of more than -20% are positioned at 1-8 and 21-28, and almost half of the notes in these two sentences have negative IOI deviation. At this time, if the singer brings himself into the state of mind of the main character, it should be urgent, eager and longing.

Secondly, when the IOI deviation is positive, it is the Rubato treatment of prolonging the time value. First of all, the three singers coincidentally prolonged most of the notes of the two adjectives “fair” and “lady”, separated by rests. Adjectives are meant to be intensified in poetic recitations, and this is a double intensification. Secondly, the corresponding note IOI deviations of the three singers are positive at almost all sentence connections in the lyrics, up to a maximum of +86%. Here the musical repertoire and melody serve the breaks and emotional tone of the lyrics. Lastly, all three singers have made an average of +45% sustaining treatment for the important verb “to seek” in the A section “to seek”, which takes over the meandering melodic tone of the A section and the B section near the end of the A section.

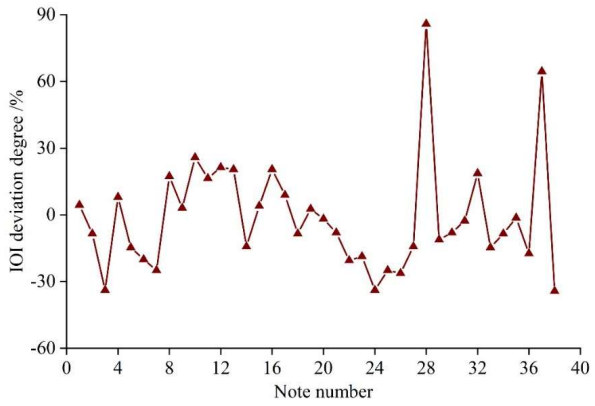


Figure 10: Shi Yijie sang the IOI deviation curve from bars 10 to 17 of the version

(2) “Compensatory Balance” based on pitch and general tendency

In the chorus section of this piece, the c and c' in the B section from measure 29-37 are the two most exciting lines, and also the difficult part of the piece.

The IOI deviation curves of the three singers are extremely similar, with a significant positive IOI deviation in the latter half of the eighth note "Cai" and the 29th note "Mi" at the highest pitch, as well as in the latter half of the 39th note and the connecting note "Le" at the end of the sentence. Taking Shi Yijie's singing version as an example, the highest deviation even exceeds+110%. In addition, almost all other sounds have negative IOI deviations, which also presents a classic Rubato expression: "compensatory balance".

IV. Conclusion

This study effectively solves the problem of objective quantitative analysis of emotional expression in vocal performance by constructing a multimodal music emotion classification model based on optimized residual network. Experimental validation shows that the proposed algorithm achieves an emotion recognition accuracy of 88.34% on the classical piano dataset, which is significantly better than traditional machine learning methods. The effectiveness of the multimodal fusion strategy is fully confirmed. In the emotion recognition task of Chinese folk songs, the recognition accuracy of the model fusing lyrics and audio features reaches 85.62%, which is 6.15% and 4.19% higher than that of a single textual and audio modality, respectively, indicating that synergistic effects of different modal information can significantly improve the recognition performance. The optimized residual network structure effectively mitigates the gradient vanishing problem of the deep network by introducing an improved convolutional kernel design and a jump connection mechanism, which enhances the feature expression capability. The improved Center-Softmax classifier enhances the discriminative ability of fine-grained sentiment classification by considering both intra-class aggregation and inter-class separation properties. The application of teacher-student model and data enhancement strategy further improves the generalization performance and robustness of the model. The visual analysis of vocal performance provides an intuitive quantitative tool for emotion expression, and reveals the emotion expression law of singers in artistic processing through the analysis of velocity-intensity curve and IOI deviation. The study provides scientific technical support for vocal music teaching, performance evaluation and artistic inheritance, and promotes the development of the deep integration of vocal music art and artificial intelligence technology.

References

[1] Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., & Eyben, F. (2017). The expression of emotion in the singing voice: Acoustic patterns in vocal performance. The Journal of the Acoustical Society of America, 142(4), 1805-1815.

[2] Dai, Y. (2024). Integration of Vocal Technique and Artistic Expression in Vocal Art Performances. Frontiers in Art Research, 6(2).

[3] Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?. Psychological bulletin, 129(5), 770.

[4] Zhou, L. (2023). Cultivation of artistic expression in college music and vocal music teaching. Art and Performance Letters, 4(12), 43-49.

[5] Dibben, N. (2016). Vocal performance and the projection of emotional authenticity. In The Ashgate research companion to popular musicology (pp. 317-333). Routledge.

[6] Olteţeanu, I. (2010). Vocal expression, music performance, and communication of emotions. Linguistic and Philosophical Investigations, (9), 311-316.

[7] Lv, Z. (2018, May). Exploration on the Importance of Singer's Emotion and Aesthetic Imagination in Vocal Performance. In 8th International Conference on Social Network, Communication and Education (SNCE 2018) (pp. 1024-1026). Atlantis Press.

[8] Fengqin, D. (2021). Influence of Articulation on Emotional Expression in National Vocal Music Singing. Frontiers in Art Research, 3(3).

[9] Condon, S. (2018). Preparing an emotionally expressive vocal performance. European Journal of Philosophy in Arts Education, 3(1).
[10] Neff, M. (2014). Lessons from the arts: what the performing arts literature can teach us about creating expressive character movement. Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters, 123-148.
[11] Heisel, E. (2015). Empathy as a tool for embodiment processes in vocal performance. Empirical Musicology Review, 10(1-2), 104-110.
[12] Tan, D., Diaz, F. M., & Miksza, P. (2020). Expressing emotion through vocal performance: Acoustic cues and the effects of a mindfulness induction. Psychology of Music, 48(4), 495-512.
[13] Wang, D. (2022). Emotional expression in vocal skill from the perspective of multiculturalism. Música Hodie, 22.
[14] Liu, J., & Zhou, M. (2021). The role of innovative approaches in aesthetic vocal performance. Musica Hodie, 21.
[15] Yuvaraj S. & Vijay Franklin J.. (2023). A dense layer model for cognitive emotion recognition with feature representation. Journal of Intelligent & Fuzzy Systems,45(5),8989-9005.
[16] Anthiyur Aravindan Abhinav, Kalyan Chappidi Sriram, Thumma Anirudh & Palanisamy Rohini. (2023). Prediction of Arousal and Valence State from Electrodermal Activity using Wavelet based ResNet50 Model. Current Directions in Biomedical Engineering,9(1),555-558.
[17] Shraddha A. Mithbavkar & Milind S. Shah. (2021). Recognition of Emotion in Indian Classical Dance Using EMG Signal. International Journal on Advanced Science, Engineering and Information Technology,11(4),1336-1345.