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

Combinatorial Press

Analysis and Enhancement Strategies of Emotional Expression Based on Pattern Recognition in Vocal Performance

Jinwei Zhang[1,]✉

[1] *Taizhou College, Taizhou, Jiangsu, 250200, China*

ABSTRACT

The essence of music is the carrier of human emotion expression, with the continuous deepening of music science and technology research, how to realize more accurate music emotion recognition has become the focus of public attention. This paper constructs a music emotion recognition model based on discrete emotion space (WLDNN_SAGAN). After pre-processing the collected audio data of vocal performances, the attention mechanism is introduced to weight and fuse the extracted low-level and middle-high-level music emotion features, and then the fused feature information is inputted into the WLDNN_SAGAN network to classify music emotions. The experimental results show that the model in this paper will improve the recognition accuracy of different emotions. Compared with the comparison model, the accuracy of this paper's model reaches 60% and above on three DIFFERENT datasets. The emotional vein of Chinese folk song performance identified by the model is lightness towards sadness and sacredness, which is consistent with the historical facts of Chinese folk song creation. In conclusion, the emotional expression of vocal performance can be enhanced by understanding the cultural connotation, applying singing techniques and body language.

Keywords: Music Emotion Recognition, Discrete Emotion Space, Self-Attention Module, MFCC Feature, WLDNN_SAGAN Network

1. Introduction

Emotional expression in vocal performance is that the singer conveys his inner emotions to the audience through his voice and performance skills, so that the audience can truly feel the emotions expressed in the song [1-2].

An excellent song often requires the artist to integrate his or her own emotions in order to make the song produce a stronger infectious force, emotional expression is not just simply follow the lyrics of

✉ Corresponding author.
E-mail address: temporubato11088@163.com (J. Zhang).
Received 18 February 2024; Revised 10 April 2024; Accepted 25 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-306

the content of the song, but also requires the artist to deeply understand the connotation of the song, through the voice and expression to convey the emotions and moods of the song to be expressed [3-6]. To achieve proper expression of emotions, the artist is required to analyze and understand the song in depth and master the connotation of the lyrics and the story behind it [7-8]. An excellent singer, in addition to technical training and solid musical skills, also needs to pass his or her emotions to the audience in a real way through a variety of ways of expression, so that the audience can resonate with the emotions expressed in the song, so as to realize the highest level of musical performance, so that the audience is truly moved by the music and produce a common emotional experience [9-12]. In addition, artists need to pay attention to the transformation of emotions when singing, and integrate their own emotions into the songs to make them more personalized [13-14]. Through the change of voice and the display of expression, the emotion is fully conveyed to the audience, causing their empathy and emotional resonance, and only by doing the emotional expression to the right place, can the vocal singing be more moving and touching [15-17].

In this paper, the music dataset is first pre-processed with file format conversion and data enhancement. Secondly, low and middle-high level music emotion features are extracted based on eGeMAPS feature set and music composition elements respectively. Since the middle and high level music emotion features are less standardized, feature selection is performed to obtain features with closer emotional relevance, and then the above data are processed according to the feature fusion method to obtain multi-layer features, which provide data for the construction of the model. Then the vocal emotion recognition model is constructed based on the discrete emotion space, and the most emotionally representative music clip is extracted by applying Schenkel analysis, which is inputted into the WLDNN_SAGAN network with the addition of self-attention mechanism and GAN network. Cross-sectional comparison experiments on the performance of the model are carried out in different datasets, and then the Chinese folk song performance data are used to verify the model's emotional expression analysis effect.

2. Recognition of emotional expression in vocal performance based on discrete emotional space

In the vocal emotion recognition model based on emotion space, feature extraction is a key step of pattern recognition, and effective features are extracted from vocal signals by pattern recognition techniques. These features can reflect the emotional information in the vocal signals, and then these features are input into the vocal emotion recognition model based on discrete emotion space to make a technical pavement for the analysis of emotional expression.

2.1. Data processing and model building process

The extraction and selection of music emotion features, and the design and construction of music emotion recognition model are all key factors affecting the recognition accuracy. From the above two aspects to carry out research, its data processing and model construction process is shown in Figure 1.

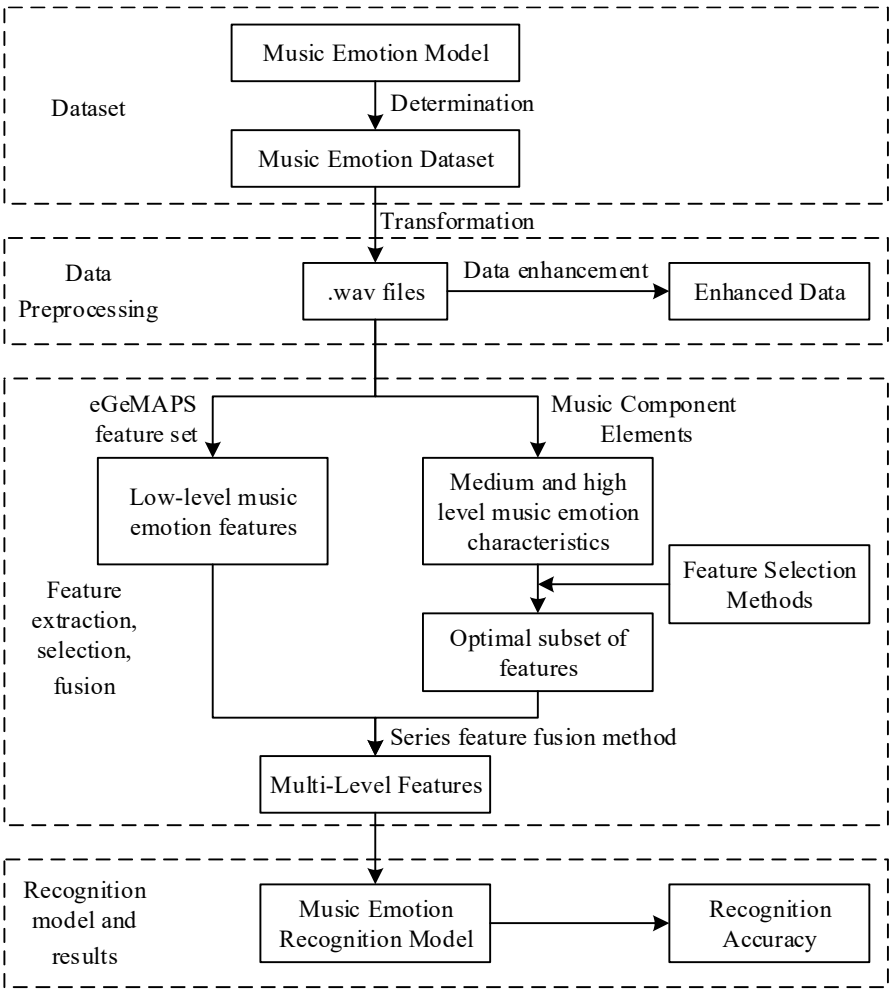

Fig. 1. Data processing and model build process

The data preprocessing operations for the experiments include format conversion and data enhancement; format conversion is the transformation of music files in MP3 format into WAV format. Data enhancement involves applying a series of morphing data to the experimental dataset.

The rationalization phase uses the AudioSegment tool to convert MP3 music files into WAV music files. In order to improve the model recognition accuracy, the WAV music files were enhanced using the Audiomentations tool using the data enhancement methods of Gaussian noise, changing the playback speed, adjusting the pitch, and time scrolling, respectively.

Assuming that the original music signal after format conversion is $p(n)$, the recognition accuracy of the above four data enhancement methods exceeds that of the original audio, and the expression of the data enhancement function is shown in Equation (1).

$$Augmentation() = compose([AddGaussianNoise(), Shift(), TimeStretch(), PitchShift()]) \tag{1}$$

The original audio data is passed through the data enhancement function to obtain a new music signal $s(n)$ as shown in Equation (2).

$$s(n) = Augmentation\big(p(n)\big) \tag{2}$$

2.2. Music Emotion Feature Extraction

2.2.1. Low-level music emotion feature extraction. To address the problem that the number of features is large and the recognition model is not flexible enough, resulting in low recognition accuracy, the eGeMAPS feature set is selected as the low-level music emotion feature for the experiments. eGeMAPS feature set features and correlation between features are theoretically and practically verified to be a standardized audio emotion feature set with standard normality. It is widely used in audio emotion related research such as speech emotion recognition and music emotion recognition.

The experiments are based on the eGeMAPS feature set, and the OpenSMILE tool is used to extract the continuous-time low-level music emotion features from the audio data, and the content of each song is represented as a time × feature form, which is saved to a .csv file. In this section, 88-dimensional features are extracted based on the eGeMAPS feature set, and the low-level music sentiment feature of a song is defined as $H^{X\times Y}=\{h_1,\ldots,h_X\}$, where X denotes the temporal dimension and Y denotes the feature dimension.

2.2.2. Middle and Upper Music Emotion Feature Extraction. The middle and high level music emotion features are closer to the emotion and richer in information. In terms of feature extraction, it is more complicated to obtain the middle and high level features from WAV files using manual calculation, so the MIRToolBox tool is chosen to extract the features. Under the guidance of professional musicians, according to the knowledge of middle and high level music emotional features introduced in related theories, combined with the middle and high level music emotional features that can be extracted by MIRToolBox tool, we extracted eight dimensional features from audio data, namely pitch, rhythm, beat transition, pitch dominant, tone brightness, roughness and regularity, and change of music state.

2.3. Selection of Musical Emotional Characteristics

2.3.1. Packaging method. The wrapping method trains the learner based on different subsets in the feature set, and the best feature subset is obtained when the loss value is minimized. The number of features to be selected for the experiment is small, so the greedy strategy is utilized to find the best subset of features and the multiple linear regression model is chosen as the learner. The features $g_{i,p}$ and label truth values y_i use multiple linear regression to obtain the model as shown in Eq. (3), where w and b are the learnable parameters, and $\hat{y}_i$ is the regression value of the ith first music sentiment data point.

$$\hat{y}_i = f\left(g_{i,p};w,b\right)=w^T g_{i,p}+b \tag{3}$$

To enhance the model generalization ability, the k-fold cross-validation method is used to ensure that the original data are all involved in the training and verified to reduce the model error. The number of feature subsets is used as a hyperparameter to observe the effect of different number of combinations on the fitting ability of the linear model.

2.3.2. Filtration. Filtering methods based on regression training methods often use both correlation coefficient and mutual information. The correlation coefficient method is a statistical measure of the linear correlation between each feature and the label truth value. The mutual information method is a measure of the amount of information shared between features and labels, and is used to capture the degree of linear and nonlinear correlation between each feature and the label truth value. The experiments use the above two methods for feature selection from both linear and nonlinear perspectives of features and labels.

1) Correlation coefficient method. In statistics, the three commonly used correlation coefficients are Pearson [18], Spearman [19] and Kendall Correlation coefficients. Kendall is applicable to test the categorical variable correlation, which is not consistent with the experimental training method. Two correlation coefficients, Pearson and Spearman, which are relevant to regression training, are chosen for experiments to study the linear correlation between features and label truth values for feature selection.

Pearson is used to measure the linear correlation between features and label truth values, which is greatly affected by nonlinear data, and is calculated as shown in Equation (4).

$$pearson = \frac{\sum_{i=1}^{N}\left(g_{i,p}-\bar{g}_p\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(g_{i,p}-\bar{g}_p\right)^2\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}} \tag{4}$$

where $g_{i,p}$ denotes the p-dimensional sentiment feature of the ith song, y_i denotes the label truth of the sentiment data point of the ith song, $\bar{g}$ is the mean value of the variable g_p, and $\bar{y}$ is the mean value of the variable y.

Spearman's measure of linear correlation between the rank order of features and labeled truth values is used to assess the monotonic relationship between features and labeled truth values, which is computed as shown below.

(1) Sort the features $g_{i,p}$ and label truth values y_i to obtain their rank order values $rg_{i,p}$ and ry_i.

(2) Calculate each feature and label order difference $d_{i,p}$ using equation (5).

$$d_{i,p} = rg_{i,p} - ry_i \tag{5}$$

(3) Calculation of Spearman's correlation coefficient according to Eq. (6) through the known conditions in (2), where N is the total amount of data.

$$Spearman = 1 - \frac{6\sum_{j=1}^{N} d_{i,p}^2}{N\left(N^2-1\right)} \tag{6}$$

2) Mutual Information Method. Mutual information of continuous dimension type data mainly uses $k-$ nearest neighbor estimation method, the feature $g_{i,p}$ and the label truth value y_i form the point $z_{i,p} = \left(g_{i,p}, y_i\right)$, and the maximum of the Euclidean distance is chosen as the k nearest neighbor criterion in the process of calculating the specific of the mutual information computation process is shown below.

(1) The distance from $z_{i,p}$ to the kth neighbor is defined as $\varepsilon(i)/2$, and the distances of this distance projected to the $g_{i,p}$ subplane and the y_i subplane are denoted as $\varepsilon_{g_{i,p}}(i)/2$ and $\varepsilon_{y_i}(i)/2$, which can be known to take the value of $\varepsilon(i)$ according to the k nearest neighbor criterion, which is computed as shown in the following equation.

$$\varepsilon(i) = \max\left\{\varepsilon_{g_{i,p}}(i), \varepsilon_{y_i}(i)\right\} \tag{7}$$

(2) The count of points whose distance to $g_{i,p}$ is less than $\varepsilon(i)$ is $n_{g_{i,p}}(i)$, and the count of points whose distance to y_i is less than $\varepsilon(i)$ is $n_{y_i}(i)$, which yields a mutual information of the feature $g_{i,p}$ and the mutual information $I\left(g_{i,p}, y_i\right)$ of the label y_i.

$$I\left(g_{i,p},y_i\right)=\Psi(k)-\left\langle\Psi\left(n_{g_{i,p}}(i)+1\right)+\Psi\left(n_{y_i}(i)+1\right)\right\rangle+\Psi(N) \tag{8}$$

where $\Psi(\cdot)$ is the Digamma function and the iteration rule is $\Psi(x+1)=\Psi(x)+\frac{1}{x},\Psi(1)\approx-0.5772156$.

2.4. WLDNN_SAGAN

The continuous emotion recognition model of WLDNN_GAN improves the accuracy to a certain extent compared with the current continuous emotion recognition model, but it still has several disadvantages as a training model for music features as follows:

1) When adding discrete music emotion features, the music information to be processed becomes more, and in order to keep the data from overfitting on the model, the model is improved by adding Self-attention [20] module to the original.

2) The role of the fully-connected layer in the model is limited to the regression prediction processing of music sentiment, which fails to provide help for the sentiment classification task and will reduce the interpretability of the model.

To address the problems mentioned above, this paper constructs the WLDNN_SAGAN model, firstly adding the Self-attention module to replace the traditional convolutional feature maps with self-attention feature maps to help solve the remote and multi-level dependency problems encountered in modeling. Secondly, we add the spectral paradigm normalization to the generator and discriminator modules in the GAN network, so that while the discriminator satisfies the 1-lipschitz limit, it can avoid the generator from generating gradient anomalies due to too many parameters, and can also make the whole model more efficient and smooth during training. The reason that the regularization of the discriminator slows down the learning of the GAN network is because of the training in real situations, where multiple update steps are required for each generator, and therefore this chapter will be applied to the generator and the discriminator with a separate learning rate (TTUR).

2.4.1. Model structure. In this section, the SAGAN module is constructed based on the traditional GAN network with the addition of the self-attention mechanism module, because according to recent studies, effective regulation of the generator can affect the performance of the GAN network, so the spectral regularization is added to the generator of the GAN network in order to get better results. The self-attentive module has a good balance of statistical efficiency, computational efficiency, and the ability to process long sequential inputs, so this chapter constructs the Self-Attentive Generative Adversarial Network (SAGAN).

GAN is one of the hottest generative algorithms today, widely used in image generation, speech generation and other scenarios.GAN consists of two models: discriminator and generator.

The objective function to be optimized by the network is:

$$\begin{aligned}\min_G\max_D V(D,G)=&E_{x\sim pdata(x)}\left[\log D(x)\right]\\&+E_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]\end{aligned} \tag{9}$$

Specifically the objective function for the optimization of the G network is:

$$L=\log\left(1-sigmoid\left(D\left(G(z)\right)\right)\right)\approx-\log\left(sigmoid\left(D\left(G(z)\right)\right)\right) \tag{10}$$

The objective function of optimization for the D network is:

$$\begin{aligned}L&=-\log\left(simoid\left(D(x)\right)\right)+\log\left(1-sigmoid\left(D\left(G(z)\right)\right)\right)\\&\approx-\log\left(sigmoid\left(D\left(G(z)\right)\right)\right)\end{aligned} \tag{11}$$

Due to the problem of limited capacity and vanishing gradient of traditional neural networks, only short-distance dependencies can be established in practice. Self-attention model can "dynamically" assign the weights of different connections, which is a good solution to this problem.

Self-attentive models usually use the QKV model, assuming that the input sequence is $\mathrm{X}=\left[x_1,\ldots,x_N\right]\in \mathrm{R}^{\mathrm{D_x}**\mathrm{N}}$ and the output sequence is $H=\left[h_1,\ldots,h_N\right]\in R^{D_v*N}$, mapping each input to three different spaces to obtain three vectors: the query vector $q_i\in \mathrm{R}^{\mathrm{D_k}}$, the key vector $k_i\in \mathrm{R}^{\mathrm{D_k}}$ and the value vector $v_i\in \mathrm{R}^{\mathrm{D_k}}$.

$$Q=W_qX\in R^{D_k\cdot N} \tag{12}$$

$$K=W_kX\in R^{D_k\cdot N} \tag{13}$$

$$V=W_vX\in R^{D_k\cdot N} \tag{14}$$

$$Attention\left(Q,K,V\right)=soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{15}$$

The attention mechanism has a very good performance in serialized data such as speech recognition, machine translation, and lexical annotation, and has been proved to be one of the excellent models for solving the problem of music emotion recognition.

In this paper, notation $W_g\in R^{\bar{c}*c}$, $W_f\in R^{\bar{c}*c}$, and $W_h\in R^{c*c}$ are the weight matrices of the learning, which are all realized by convolution with $1*1$, and experiments are conducted using $\bar{C}=C/8.\mathrm{f}$. f and g are two formulas for feature extraction, where $f(x)=W_fx,g(x)=W_gx^{\circ}$, and $g(x)=W_gx$. Denote by $\beta_{j,i}$ the degree of influence of the model on the ith position at the jth region with the following formula:

$$\beta_{j,i}=\frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N}\exp\left(s_{ij}\right)},s_{ij}=f\left(x_i\right)^T\,g\left(x_j\right) \tag{16}$$

The output of the attention layer $o=\left(o_1,o_2,\ldots o_j,\ldots o_N\right)\in R^{C*N}$, where:

$$o_j=\sum_{i=1}^{N}\beta_{j,h}h\left(x_i\right),h\left(x_i\right)=W_hx_i \tag{17}$$

In addition to this, during the output of o, the scale parameter γ needs to be multiplied and added to the input, with the overall output shown in the following equation:

$$y_i=\gamma o_i+x_i \tag{18}$$

First γ is initialized to 0, after which progressively more weights are assigned to the non-local features, and the loss expression used by SAGAN is as follows:

$$\begin{aligned}L_D=&-E_{(x,y)\sim pdata}\left[\min\left(0,-1+D(x,y)\right)\right]\\&-E_{Z\sim pz,y\sim pdata}\left[\min\left(0,-1-D(G(z),y)\right)\right]\end{aligned} \tag{19}$$

$$L_G=-E_{z\sim pz,y\sim pdata}D\left(G(z),y\right) \tag{20}$$

The WaveNet module was first proposed to generate for audio signals, mainly to solve the problem of the original RNN module's excessive computation, and consequently the number of convolutional layers used is low, the size of the convolutional kernel becomes large, and the choice of inflated causal convolution, such a treatment can indeed reduce the complexity of the network, but it also poses a certain problem that the features it extracts and learns from are confined to the low-dimensional

rough feature vectors. Using a complex neural network can be helpful for feature extraction, so this paper combines two modules, WLDNN and SAGAN.

After the emotion features have passed through the WaveNet network, they will be inputted into the SAGAN network for the final regression prediction, and this paper chooses the direct isotropic connection.The most important feature of the WLDNN_SAGAN network is that it divides the two parts of the model differently, and puts the emotion extraction in the WLDNN, and puts the regression prediction in the SAGAN.

As an artificial neural network for continuous emotion model, firstly, the inflated RNN layer in WLDNN is able to design the repetitive module according to the actual number of input features, the dimensionality reduction and weighted fusion of the two emotion features is responsible for by the linear layer, the mining of temporal relationship in the emotion features is responsible for by the SAGAN, and the normalization and prediction processing of the regression data is done by the final fully connected layer, the various processes of the detailed parameters of each process will be given below.

2.4.2. Multimodal feature fusion. The audio features in this chapter are still chosen as MFCC features and RP features.MFCC features in this section add a phase residual for complementary audio information.RP is the cosine of the phase function in the parsed signal derived from the residuals of the linear prediction of the music signal. A given music sample $s(n)$ can be predicted as a linear combination of the past p music samples predicted:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{21}$$

$$e(n) = s(n) - s(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{22}$$

where p refers to the order of prediction, the coefficients $a_k, k = 1,2,\ldots,k$, $k = 1,2,\ldots,k$ are the linear prediction coefficient set (LPCS), and the prediction error $e(n)$ is defined as the error between the actual value $s(n)$ and the predicted value $s(n)$ \$ between the actual value $s(n)$ and the predicted value $s(n)$, and this error is also known as the residual of the music signal.The LP residual contains a lot of information about the music emotion, and the resolved signal is $r_a(n) = r(n) + jr_h(n)$ by $r(n)$, where j is a fixed coefficient, and $r_h(n)$ is $r(n)$ of the Hilbert variation with the following formula:

$$r_h(n) = IFT\left[R_h(W)\right] \tag{23}$$

where IFT is the Fourier inverse variation and $R(w)$ is the Fourier transform of $r(n)$, calculated as shown below:

$$R_h(W) = \begin{cases} -jR(w), 0 \le w < \pi \\ jR(w), -\pi \le w < 0 \end{cases} \tag{24}$$

The final resolved signal can be obtained:

$$r_a(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{25}$$

Corresponding down to the remaining phase is calculated as:

$$\cos(\vartheta(n)) = \frac{R_e\left(r_a(n)\right)}{\left|r_a(n)\right|} = \frac{r(n)}{h_e(n)} \tag{26}$$

Finally, the MFCC features and RP features are then weighted and combined to obtain a brand new feature vector F_{MF_RP}, and the specific computational flow is shown in Fig. 2:

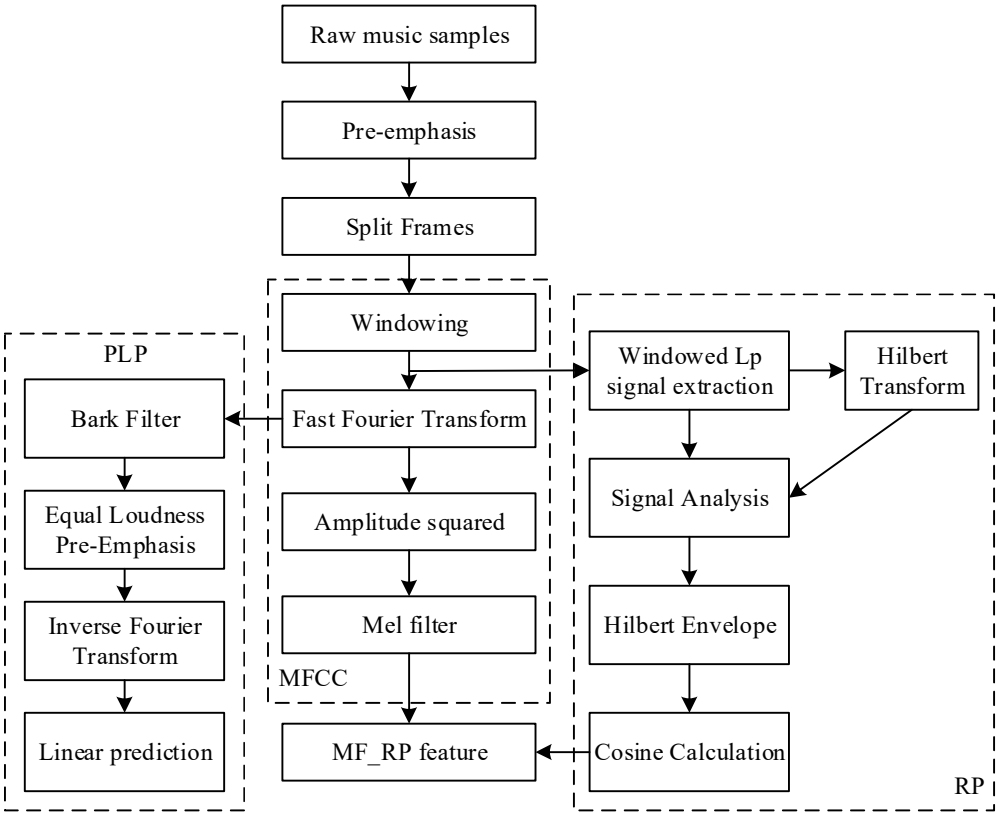

Fig. 2. Flowch art for multimodal music feafure extraction

3. Model performance evaluation experiments

3.1. Introduction to the data set

The EMA dataset was collected and produced in a borrowed manner, with all emotion category music coming from the Internet, and the original music collected in both MP3 and MIDI format files. For the pre-processing of the subsequent experiments, the function of AudioSegment class function in the Pydub library was used to uniformly convert the MP3 format of all emotion category music to WAV format.

The EMA dataset consists of pure music with a total of four emotion categories. In the research process, in order to improve the convenience of processing, the first 60 seconds of each song was simply selected, and those with a duration of less than one minute were subjected to the zero complementary operation.

The Emotion dataset consists of music in MP3 format, and music emotions were categorized into 4 categories. For the convenience of the experiment, only the first 30 seconds of each song was used, and those less than 30 seconds were subjected to a make-up-zero operation.

The 4Q-emotion dataset consists of music in MP3 format, and the music emotions are not classified according to emotion words, but according to the four tags Q1, Q2, Q3 and Q4. Only the first 30 seconds of each piece of music is used, and the complementary zero operation is performed for less than 30 seconds.

3.2. Loss Functions and Evaluation Indicators

The loss function used in the experiments in this chapter is the cross-entropy loss, which mainly portrays the distance between the actual output (probability) and the desired output (probability), the smaller the value the closer the two probability distributions are, and the better the model is used

for multi-label classification tasks. Where N denotes the number of samples i, M denotes the number of categories, when the category is the same as the category of sample i y_{ic} is 1, otherwise it is 0, and 0, p_{ic} denotes the predicted probability of the sample i belonging to the category C for the sample i.

$$L = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -\sum_{c=1}^{M} y_{ic}\log\left(p_{ic}\right) \tag{27}$$

For the simplest binary classification problems, common evaluation metrics are accuracy, precision, recall and $F1$. One class in the distinction to be made is set as a positive class, and the other class is set as a negative class, T indicates a correct prediction, and F indicates an incorrect prediction.

The accuracy rate represents the ratio of the number of correctly categorized samples to the overall number of samples, as shown in equation (28):

$$acc = \frac{TP + TN}{total} \tag{28}$$

The precision rate indicates the proportion of the number of samples in the true positive category to the proportion of all samples predicted to be in the positive category, as shown in equation (29):

$$pec = \frac{TP}{TP + FP} \tag{29}$$

The recall rate indicates the proportion of the number of samples predicted to be in the positive category to the number of samples in all true positive categories, as shown in equation (30):

$$rec = \frac{TP}{TP + FN} \tag{30}$$

Precision rate and recall rate indicators there is a certain contradictory relationship, one increase under the condition that the other will decrease, in a large sample set to show mutual constraints, in order to achieve a certain combination of trade-offs for the purpose of setting $F1$ indicators, $F1$ corresponds to the weighted sum of the average of these two indicators, as shown in equation (31):

$$F1 = \frac{2 \cdot pec \cdot rec}{pec + rec} \tag{31}$$

For a multi-classification problem, assuming that the number of sample labels is k, when using the confusion matrix for a binary classification problem, the $k - j$ class can be regarded as a positive class, then the remaining $k - j$ classes can be regarded as negative classes, and TP, FN, FP, and TN of the i classes can be calculated. Further calculations can obtain the accuracy, precision, recall and $F1$ of the category, so that the corresponding evaluation indicators of k categories can be calculated in turn.

3.3. Experimental design

The EMA dataset, the Emotion dataset and the 4Q-emotion dataset were randomly assigned the training set and the test set in the ratio of 9:1. Fig. 3-Fig. 6 represent the 3-frame MF_RP feature timing diagrams extracted from the music of the 4 emotion types in the EMA dataset.

It can be seen that the two music emotions, nervousness and excitement, have very different MF_RP feature profiles for different frames than the other music emotions. Cheerfulness and pleasure have similar features in the middle and high frequency bands, while the differences are larger in the middle and low frequency bands. Therefore, the MF_RP feature can improve the ability to extract emotional information from music signals, and the MF_RP feature is effective in capturing the differences even for emotions that are subjectively closer to a person.

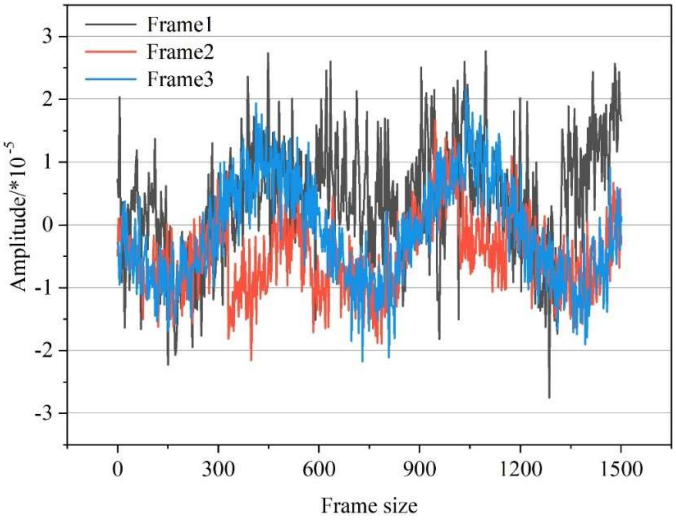

Fig. 3. The characteristics of the stress of music

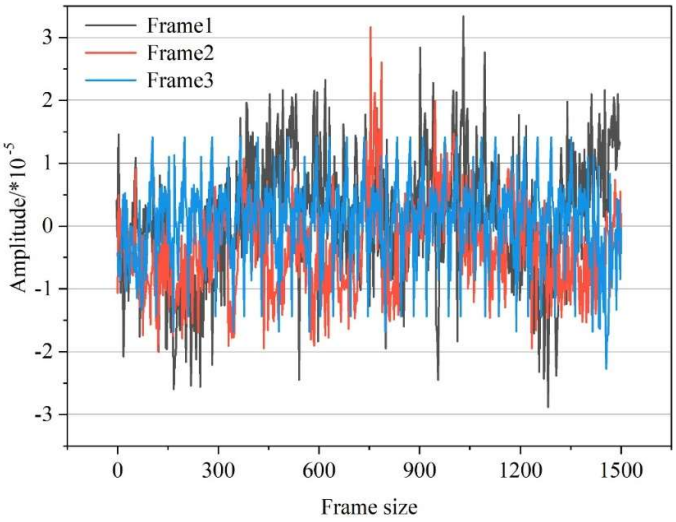

Fig. 4. The characteristics of the boil of music

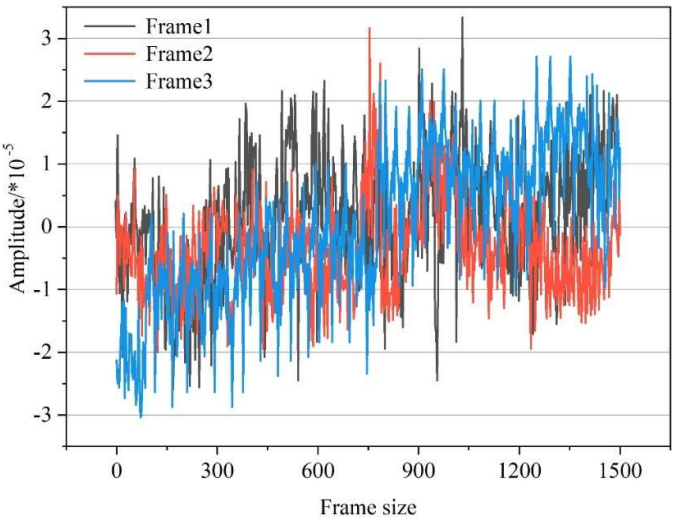

Fig. 5. The characteristics of the cheerfulness of music

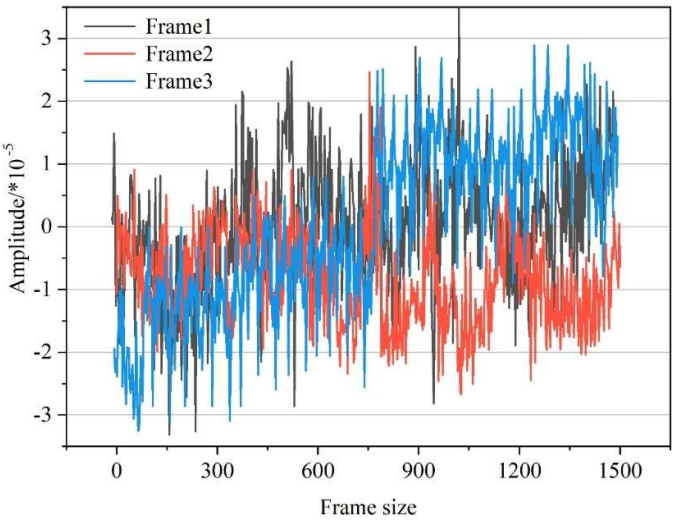

Fig. 6. The characteristics of the pleasure of music

3.4. Experimental results and analysis

The confusion matrix of the vocal emotion recognition model using only fused features is shown in Fig. 7, where darker colors indicate higher accuracy in this classification. It can be seen that the model without adding the Self-attention module has a higher classification accuracy in the emotion categories of Excited and Nervous. This is because the audio features of both are unique and highly recognizable, while the two categories of Cheerful and Pleasure are less effective in classification due to their more similar audio energies, even though they are supplemented with features using RP features, the classification accuracy is 19% and 15% for Cheerful and Pleasure, respectively, as they are still based on audio features and cannot be feature extracted from other aspects.

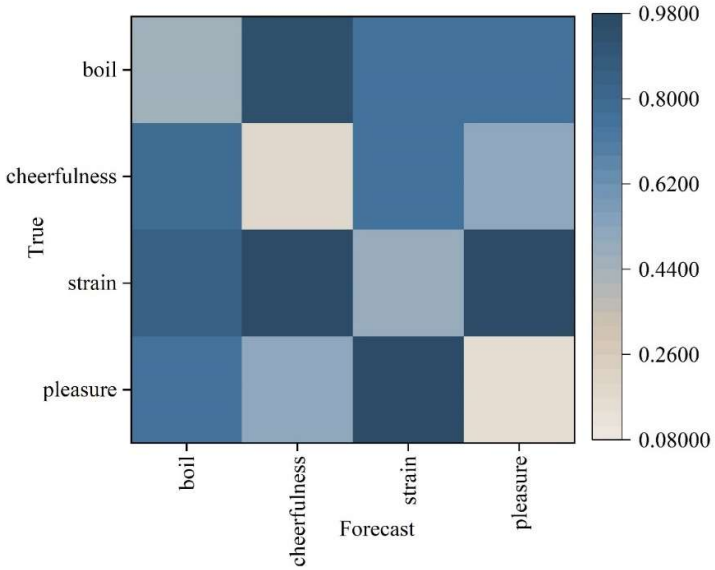

Figure 7 Only use the emotional recognition model of fusion characteristics

The model confusion matrix after adding the Self-attention module is shown in Fig. 8, and the recognition accuracies of the two categories of cheerfulness and pleasure are significantly improved to 55% and 45%, respectively. This is because the model in this paper combines the MFCC features extracted from the continuous emotion space with RP for weighting to represent a more complete and

comprehensive music emotion feature, giving the model a multimodal feature combination, which further proves the effectiveness of the model in this paper.

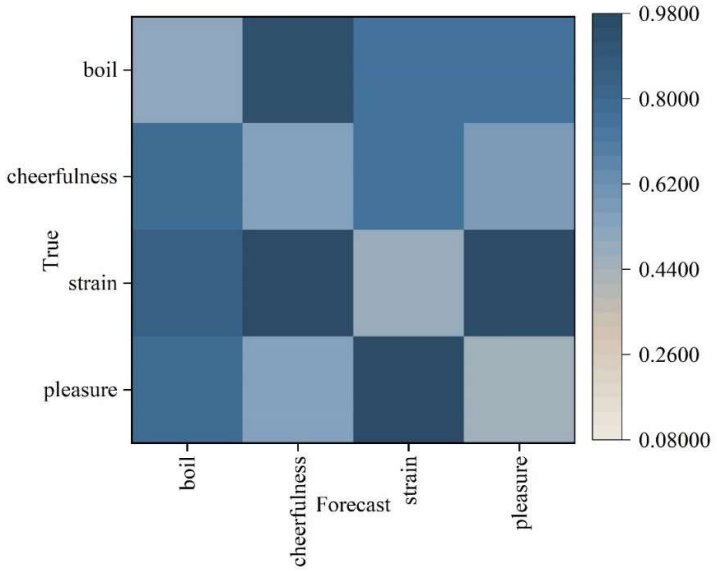

Fig. 8. The model confusion matrix after the attention module was added

Tables 1 to 3 show the results obtained when each model is recognized in the EMA dataset, Emotion dataset and 4Q dataset, respectively, and a specific analysis of the contents of this table shows that the accuracy of the model in this chapter for music emotion classification in the EMA dataset reaches 61.9%, which is 7.7% higher than that of the MCCLSTM, the RCNNBL, and the LSTM_BLS, respectively, 4.3% and 0.6%, respectively.

The classification accuracy of the model in the Emotion dataset reaches 64.9%, which is 0.8% better than the best performing LSTM_BLS. For the 4Q dataset, the best performing LSTM_BLS model is inferior to this paper's model in terms of both training efficiency and model classification accuracy, and this paper's model combined with multimodal feature fusion achieves the best results on the EMA dataset, the Emotion dataset, and the 4Q dataset.

Table 1. Model classification comparison (EMA)

Model	Accuracy	Precision	Recall	F1	Training time (s)
MCCLSTM	0.542	0.596	0.541	0.573	531.87
MCCBL	0.530	0.572	0.535	0.551	90.66
RCNNLSTM	0.547	0.583	0.540	0.568	1098.44
RCNNBL	0.576	0.588	0.573	0.587	115.28
LSTM_BLS	0.613	0.634	0.603	0.617	324.58
This model	0.619	0.632	0.620	0.625	462.75

Table 2. Model classification comparison (Emotion)

Model	Accuracy	Precision	Recall	F1	Training time (s)
MCCLSTM	0.587	0.633	0.543	0.585	275.13
MCCBL	0.581	0.594	0.520	0.553	40.82
RCNNLSTM	0.57	0.606	0.507	0.554	615.52
RCNNBL	0.582	0.600	0.619	0.604	55.93
LSTM_BLS	0.641	0.653	0.618	0.634	169.19

This model	0.649	0.667	0.628	0.640	216.84

Table 3. Model classification comparison (4Q)

Model	Accuracy	Precision	Recall	F1	Training time (s)
MCCLSTM	0.535	0.59	0.541	0.580	532.42
MCCBL	0.529	0.555	0.510	0.568	118.20
RCNNLSTM	0.532	0.602	0.510	0.559	1120.18
RCNNBL	0.564	0.575	0.561	0.577	101.85
LSTM_BLS	0.640	0.629	0.626	0.623	314.77
This model	0.653	0.658	0.638	0.644	462.49

4. Analysis of the pattern of change in vocal emotion

In this paper, we use Chinese folk song performance data to analyze the emotional expression of vocal performance. The folk song data comes from the "Songs and Music" section in the Chinese National Music Resource Database, which mainly includes information such as lyrics, audio, and emotional labels. The time span of folk songs is from ancient times to modern times, and the geographical scope covers 30 provincial administrative regions such as Anhui, Fujian, Gansu, etc. The themes of the songs include labor, current affairs, rituals and customs, love, children, and so on. The dataset contains 24 manually labeled emotion labels such as cheerfulness, euphony, sadness and anger, among which each folk song contains 1-2 labels. After the steps of weight removal and missing value removal, a total of 1000 Chinese folk song performances were obtained for subsequent experimental analysis. The poignant and sad all express negative sentimental emotions, and the label proximity is categorized, and finally all emotions are grouped into 7 categories.

4.1. Characterization of the Long-Term Pulse

In the calculation of the frequency of emotional fluctuations, the number of fluctuations of all the songs was firstly counted, and a total of 2548 fluctuations occurred in 1000 songs, among which the 20 songs with the highest F_{fluc} and their frequency values are shown in Table 4. It can be seen that most of the songs with frequent emotional fluctuations were created in the revolutionary period, reflecting the arduous journey of the Chinese people rising up against enemy aggression and moving from suffering to glory under the leadership of the CPC, and the F_{fluc} of "Graduation Song" and "Ode to the Huangpu River (II)" have reached 0.747 and 0.744, respectively.The revolutionary struggle was difficult and tortuous, and the emotions of the red songs correspondingly The revolutionary struggle was difficult and complicated, and the emotions of the red songs were correspondingly ups and downs. In addition, "Under the Silver Moonlight" and "Senjidema" also have a high frequency of emotional fluctuations, both songs depict the unswerving love of young people and the end of their love, and the emotions change frequently between anticipation, elation and sadness.

Table 4. Folk Songs with the Highest Fluctuation Frequency

Serial number	Folk song	F_{fluc}	Serial number	Folk song	F_{fluc}
1	Graduation song	0.747	11	Long live the motherland of youth	0.626
2	The huangpu river is in song (2)	0.744	12	Red army discipline song	0.623
3	jinggang	0.713	13	In the silver moonlight	0.621
4	Chairman MAO is the red sun of the people of all ethnic groups	0.689	14	Chairman MAO is great	0.615
5	Gutian conference resolution directs the direction	0.679	15	What does chairman MAO say? Let's do it	0.611
6	No tears, no sorrow	0.670	16	Chairman MAO, we wish you longevity	0.607
7	Be sure to put the flag of victory in Taiwan	0.641	17	We should believe in the masses and we should believe in the party	0.598
8	Big knife march	0.636	18	Great leader MAO zedong	0.585
9	Policy and strategy are party life	0.635	19	Sengedama	0.575
10	Soldiers love to sing revolutionary songs	0.627	20	I love the great motherland	0.572

In terms of the distribution of the emotional starting point - end point of the folk song, the proportion and flow of the first and last emotions are shown in Figure 9. Observing the proportion change, it can be found that the proportion of lightness at the end of the song has a more significant decline, and the proportion of sadness and sacredness increases significantly. Observing the specific flow of emotion, analyzing from the starting point of emotion, it can be found that the flow of emotion change of lightness targets mainly for vitality, other and happiness. From the analysis of emotional end point, it can be found that the source of emotional change of sadness is mostly lightness and passion, and the source of sacred emotion is mostly lightness and other. Based on the analysis of the emotional labels at the beginning and end of the songs, the emotional vein of Chinese folk song performances shows the flow direction of [lightness] → [sadness, sacredness]. Folk songs are mostly based on the production life of the working people, and start from the depiction of daily life and scenery, so the starting point of emotion is often lightness. The classic folk songs were written before the founding of New China, and they often expose social contradictions and criticize the reality, which is the source of the sad ending. The folk songs created in the post-revolutionary period reflect the destruction of people's lives by the atrocities of the invaders, and praise the Red Army's spirit of arduous struggle and glorious course, which is in line with the tone of sadness and sacredness, which is consistent with the characteristics of the long-time pulse and the pattern of change of emotions extracted from the emotion recognition model of this paper on Chinese folk songs, which illustrates the validity of this paper's model, and allows for a comprehensive analysis of emotions in vocal performance.

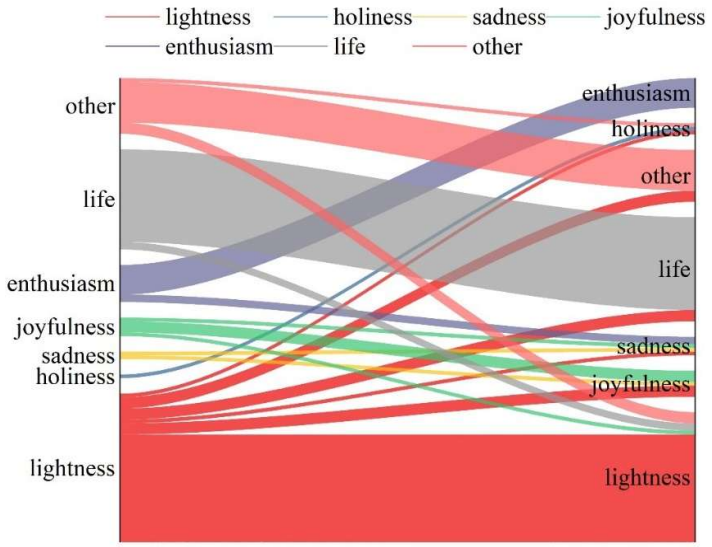

Fig. 9. The emotional flow of folk songs

4.2. Analysis of short-term fluctuation patterns

The second observation is the fluctuation of a single emotion. The addition and disappearance of emotions can be understood as the transformation between different emotions. Based on this the combination of emotion label fluctuation data refinement processing, to get seven kinds of emotion label two by two the frequency of correlation between the change, and presented in the form of network graph, the emotion network is shown in Figure 10. The width of the edges in the network graph is proportional to the frequency of bi-directional changes between two nodes. It can be intuitively seen that the emotional fluctuation pattern of Chinese folk song performances shows the pattern of lightness as the center, radiating to the rest of the surrounding emotions, or converting from the rest of the emotions to lightness, and the probability of bidirectional fluctuation with sadness, vitality, and other emotions is significantly higher than that of the other three emotions. Compared with the profound and serious western music, Chinese folk song performances are devoted to portraying life scenes, with a strong sense of life, and in terms of language and tunes, they all present the characteristics of lightness, flexibility and variability, which is consistent with the overall direction of the characteristics of the long-time vein of the vocal performance of Chinese folk songs, and once again verifies the accuracy of this paper's model in the analysis of short-term emotional fluctuation patterns.

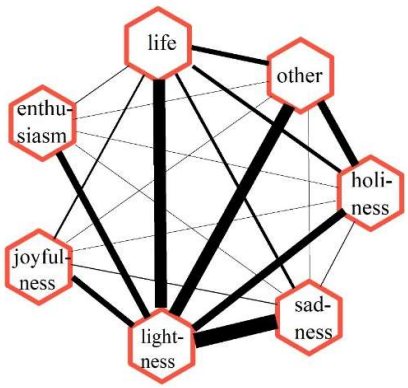

Fig. 10. Network of Emotional Fluctuation Pattern

5. Strategies for Enhancing Emotional Expression in Vocal Performance

In the performance of vocal works, to effectively improve the expression of emotion, we need to rely on the singer's own cultural literacy to deeply understand the cultural connotation of the work, master the content and emotional tone of the work, grasp the rhythm and melody of the work, realize the performance of the main body of the work to identify with the work, and through the appropriate singing treatment to show the artistic image and emotional connotation of the work, so as to give the audience a beautiful experience.

5.1. Deep understanding of the cultural connotation of vocal works

Each note and each phrase in the vocal works are cohesive with the thoughts and feelings of the authors of the lyrics and compositions. To accurately and profoundly express the emotional connotation of the works, it is necessary to comprehensively grasp the specific content of the works and the subtle changes in the emotions. First of all, we should pay attention to the background analysis of the singing works, and excavate the deep cultural connotation from the creative meaning of the works. Many works have a strong period characteristics and national flavor, full of the author of the contemporary society, the feelings of real life, and therefore to penetrate the understanding of the works of the times, national culture and regional customs, accurate positioning of the works of art style and the expression of the spiritual connotation of the times.

5.2. Reasonable use of singing techniques for emotional expression

Singing skills are the “hardware conditions” for the emotional expression of vocal works, which serve the emotional expression. In order to make the singing “sound and emotion”, it is necessary to reasonably design the singing skills of the work, such as biting and spitting, breathing control, timbre changes, etc., so as to make the student convey the accurate emotion.

First of all, the biting and spitting of words should be combined with the language characteristics of the work and its pronunciation rules, so as to achieve accurate pronunciation, standardized biting and clear spitting, so as to make the audience understand the content of the work more easily.

Secondly, the breathing control should be natural and smooth, and a variety of breathing techniques should be used flexibly to show the changes of emotions. Singing breath control is based on the work of the rhythm, melody, syntax and emotional changes and other factors to determine the general cheerful music to feel the fragrance of the flowers as to inhale; sad music to be like “crying” in general to adjust their own breath.

Finally, the tone should be reasonably used to express the emotional tone of the work. Tone is the basic positioning of emotional expression, different works need to be sung with different tones.

5.3. Appropriate use of body language to assist emotional expression

The art of vocal performance includes two important elements: “singing” and “acting”, i.e., the use of singing skills and body language to create a vocal work, and the perfect presentation of the work's artistic image from the auditory and visual points of view. Body language is a silent form of communication that expresses emotions through facial expressions, body movements and postures. Body language is driven by emotion and changes with the change of emotion, which can bring more direct visual impact and emotional experience to the audience and further enhance the infectious force of the performance. Facial expression is the most direct, distinctive and graphic part of the expression of inner emotion, which can convey joy, anger, sadness and other emotions.

6. Conclusion

In this paper, we propose a low-middle-high level music feature extraction based on discrete emotion space and combined with the optimized deep learning model WLDNN_SAGAN for emotion expression recognition in vocal performance.

The extracted vocal features provide multimodal music emotion feature input to the model, and the experimental data show that the addition of the self-attention module to the vocal emotion recognition model in this paper increases the accuracy of music emotion recognition, improving the recognition accuracy of the two categories of cheerfulness and pleasure to 55% and 45%, respectively. In addition, this paper's model has higher classification accuracies than the comparison model on three different datasets, with accuracies of 61.9%, 64.9% and 65.3% on the EMA dataset, the Emotion dataset and the 4Q dataset.

The model of this paper was migrated to the task of emotion recognition of Chinese folk song fragments, and the long-time pulse characterization of emotion and short-time fluctuation pattern analysis were carried out. It was found that the emotion fluctuation pattern of the Chinese folk song performance showed a lightness as the center and the rest of the emotions radiating around. The overall emotional pulse of Chinese folk song performances showed the flow direction of [lightness] → [sadness, sacredness]. The fact that the folk songs were created before the founding of New China, often expose social contradictions and criticize reality, and praise the Red Army's spirit of hard struggle and the sacred course of history is consistent with the model of this paper, which confirms that the model of this paper is effective in the task of recognizing the pattern of emotional expression.

Finally, this paper designs the enhancement strategy of emotional expression in vocal performance from the understanding of cultural connotation, the reasonable application of singing skills and the auxiliary application of body language, which contributes to the enhancement of the artistic charm of vocal singing and brings better emotional experience to the audience.

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] Yanhong, Z. (2021). Discussion on Aesthetic Thinking in Vocal Performance. Frontiers in Art Research, 3(3).

[3] Athanasopoulos, G., Eerola, T., Lahdelma, I., & Kaliakatsos-Papakostas, M. (2021). Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music. PLoS One, 16(1), e0244964.

[4] Juslin, P. N. (2013). Vocal affect expression: problems and promises. Evolution of emotional communication, 24, 252-73.

[5] Quinto, L., Thompson, W. F., & Taylor, A. (2014). The contributions of compositional structure and performance expression to the communication of emotion in music. Psychology of Music, 42(4), 503-524.

[6] Chen, L. (2022). THE IMPORTANCE OF VOCAL SKILLS IN VOCAL PERFORMANCE ART AND ITS IMPACT ON IMPROVING THE AUDIENCE'S POSITIVE PSYCHOLOGY. Psychiatria Danubina, 34(suppl 4), 780-780.

[7] Meissner, H. (2021). Theoretical framework for facilitating young musicians' learning of expressive performance. Frontiers in psychology, 11, 584171.

[8] Beveridge, S., & Knox, D. (2018). Popular music and the role of vocal melody in perceived emotion. Psychology of music, 46(3), 411-423.

[9] Wenqi, L. (2022). Imagination in vocal performance art and its cultivation measures. emotion, 4(7), 71-75.

[10] Bowling, D., Gingras, B., Shui'er Han, J. S., & Opitz, E. (2013). Tone of voice in emotional expression and its implications for the affective character of musical mode. Journal of Interdisciplinary Music Studies, 7(1-2), 29-44.

[11] Hakanpää, T., Waaramaa, T., & Laukkanen, A. M. (2021). Comparing contemporary commercial and classical styles: emotion expression in singing. Journal of Voice, 35(4), 570-580.

[12] Sundberg, J., Salomão, G. L., & Scherer, K. R. (2021). Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation. Journal of Voice, 35(1), 52-60.

[13] Coutinho, E., & Scherer, K. R. (2017). The effect of context and audio-visual modality on emotions elicited by a musical performance. Psychology of Music, 45(4), 550-569.

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

[15] Correia, A. I., Castro, S. L., MacGregor, C., Müllensiefen, D., Schellenberg, E. G., & Lima, C. F. (2022). Enhanced recognition of vocal emotions in individuals with naturally good musical abilities. Emotion, 22(5), 894.

[16] Chen, T. T. (2016, July). Theoretical Analysis of The Art Forms of Vocal Music. In 2016 5th International Conference on Social Science, Education and Humanities Research (SSEHR 2016) (pp. 1511-1515). Atlantis Press.

[17] Scherer, K. R., Trznadel, S., Fantini, B., & Sundberg, J. (2017). Recognizing emotions in the singing voice. Psychomusicology: Music, Mind, and Brain, 27(4), 244.

[18] Ji Ting Mo,Fang Liu & Xin Xing Deng. (2025). A group decision making model with non-reciprocal fuzzy preference relations based on Pearson correlation coefficient. Expert Systems With Applications,276,127095-127095.

[19] Ryo Yokoyama,Kai Wang,Shunichi Suzuki,Shuichiro Miwa & Koji Okamoto. (2025). Determination of Spearman's rank correlation for melt spreading-solidification dynamics through the combination of integrated experiments and Monte Carlo method. International Journal of Heat and Mass Transfer,242,126831-126831.

[20] Muhammad Shahroze Ali,Afshan Latif,Muhammad Waseem Anwar & Muhammad Hashir Ashraf. (2025). Multiscale self-attention for unmanned ariel vehicle-based infrared thermal images detection. Engineering Applications of Artificial Intelligence,149,110488-110488.