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

Acoustic Feature Construction for Emotional Expression in Zhuhu Performance within Traditional Chinese Opera Heritage

Chengyao Duan[1,*]

[1] Hubei Normal University, Huangshi, Hubei, 435002, China

Abstract: *Opera is a treasure of Chinese art, and opera Zhuhu Performance is an indirect expression of human language and emotion, and the emotion analysis in opera plays a more important role in the understanding of opera and the transmission of music language. On the basis of introducing the basic theory and relevant features of musical signal, the article preprocesses the multimodal data of the Zhuhu Performance of opera by means of aggravation, frame splitting and windowing, and then completes the classification and recognition of the emotion of the Zhuhu Performance of opera based on the emotion recognition model of improved multimodal RCNN. The article uses the confusion matrix method to show how accurately the model classifies emotions in experiments, and it builds an audio signal feature where the model can identify the emotions of Zhuhu's opera performance with a relatively high accuracy rate and the classification accuracy rate of the emotion "happy" reaches the highest level.*

KEYWORDS: *multimodal RCNN; emotion recognition; audio features; Zhuhu performance*

1 Introduction

As an important carrier of information transfer, audio feature plays a crucial role in people's daily life [1]. It can carry rich environmental information, including various types of sound types, such as traffic noise, crowd conversation and natural environmental sounds. These sounds not only provide us with clues to perceive the world and help with spatial localization and contextual understanding, but also convey information about emotions and social interactions. For example, the hustle and bustle of the street, the ticking of the rain, or the sound of birds chirping reflect the state of the surrounding environment, but also contain deep emotional coloring and social context. It is these multidimensional audio signals that enable people to respond appropriately in complex environments and thus better understand the events that are taking place [2-5]. Therefore, audio is not only an important tool for information acquisition, but also constitutes an important bridge for human interaction with the world.

Traditional Chinese musical instruments serve as the tangible vessels of Chinese traditional music and culture, representing a vital component of the nation’s splendid cultural heritage. They enjoy high prestige both domestically and internationally, with broad market demand. Within the vast sonic landscape of traditional Chinese opera, one instrument holds a position of soul—the "Zhuhu"[6]. It acts as the helmsman of the entire opera, the most faithful follower and steadfast pillar of the performer’s vocal delivery. An outstanding "Zhuhu" player must share a profound understanding with the lead performer, possess a deep mastery of the musical essence of the opera genre, and be capable of guiding the rhythm and enhancing the emotional atmosphere with the fiddle’s voice amid the ever-changing dynamics of the stage [7]. It does not refer to a fixed instrument; its form varies with different operas. In the vigorous and resonant

*Chucierhu@126.com
https://doi.org/10.65102/is2026415
Published: 30 April, 2026

world of Peking Opera, it is the high-pitched and passionate Jinghu; in the gentle and flowing melodies of Yue Opera, it is the mellow and soulful Yuehu[8]. Thus, "Zhuhu" is more of an honorific title for an artistic role—the most sensitive nerve in the lifeblood of opera music, the indispensable "backbone" that upholds the charm and style of an entire performance.

The identification and categorization of musical genres as well as traditional musical instruments depend heavily on the acquisition of musical elements as they may be used to describe the characteristics of duplicated music [9, 10]. Over the past few years, the process of extracting musical features has been an area of interest within the scope of musical instrument identification and classification, which directly affects the effectiveness of identification and classification. Classification of musical features from various angles may give rise to diverse types of results. For instance, in order to address the issue regarding the difficulty of source separation leading to the difficulty in extracting audio features that occurs in the identification of polyphonic instruments using linear predictive coding (LPC) and linear predictive coding cepstrum coefficients (LPCC), Duan, Z et al. proposed using the discrete cepstrum (UDC) derived from a single spectral point of the source in the mixed spectrum for overcoming this issue, by means of which overfitting isolated spectral points can be avoided via employing more natural and locally self-adaptive regularizers [11]. Similar to this, Hung, Y N et al. created a large-scale database of synthesized polyphonic music with pitch and instrument labels at the "frame" level. In addition, they used the pitch and timbre information to perform instrument recognition using a multitask learning algorithm because of the unique qualities of pitch and timbre in audio features [12]. In order to conduct audio scene categorization, Mesaros, A. et al. converted the onset audio into characteristics such as Mel Frequency Cepstrum Coefficients (MFCC), Log Mel Spectral Coefficients, and Short Time Fourier Transform (STFT) [13].

Because audio is a time-series signal, research on neural networks for audio feature extraction mostly focuses on convolutional neural networks (CNN) as a backbone network architecture for the audio encoder. [14]. With a micro F1 score of 0.619 and a macro F1 score of 0.513, Han, Y et al.'s Convolutional Neural Network (CNN) structure for main musical instruments in chordal music is one of the most recent models for identifying musical instruments. [15]. However, Slizovskaia, O. et al. found that both the audio network alone and the multimodal approaches perform comparably to human performance on a subset of musical instruments in the YouTube-8M dataset after using multimodal data from the audio and visual domain to learn a representation of the two modalities using late fusion to recognize musical instruments in user-generated videos [16]. Furthermore, Liu, J and Xie, L used SVM classification method and compared four different audio feature sets to classify each of 13 traditional Chinese musical instruments and 13 Western musical instruments into four groups such as wind instruments, percussion instruments, string instruments and plucked instruments, while utilizing the spectral peak coefficients, the short-time Fourier transform coefficients, the spectral flatness coefficients, and the Mel cepstrum coefficients as four different audio feature sets, respectively. The experiment proved that Mel cepstrum coefficients provided the highest classification accuracy [17]. In order to improve the recognition of musical instruments, Abe´, J. et al. evaluated three CNN network architectures. All three network structures achieved an impressive result, identifying 11 musical instruments with a micro F1 value of 0.81 and a macro F1 value of 0.52, demonstrating the effectiveness of the attention mechanism-based CNNs in instrument classification [18].

At present, there is a gap in the extraction and recognition of audio features specifically for the main beard, and more research has been done on traditional Chinese musical instruments. For example, while creating an audio database of traditional Chinese musical instruments, Li, R., and Zhang, Q. created an eight-layer CNN and ResNet model for audio feature extraction and used the SVM method to classify all musical instruments by entering Mel spectral

characteristics [19]. Yang, J et al. constructed a comprehensive audio dataset containing four families of traditional Chinese musical instruments (blown, percussive, plucked and bowed) and extracted three time-frequency features, i.e., MFCC, Constant Q Transform (CQT), and timbre, in order to capture a wide range of audio features [R2020]. Jin, R et al. constructed a diverse dataset of 37 traditional Chinese musical instruments (e.g., erhu, pipa, guzheng, and flute, etc.), which contains audio features such as different pitches, timbres, and playing speeds to provide a comprehensive picture of Chinese musical instruments, and also used MFCC and spectrograms to capture important audio features in the sound of the instruments [21]. Despite the good results of some of the above traditional musical instrument recognition research, this research is still at a prospective stage. The ideal situation for musical instrument audio feature construction is to enable computers to extract different musical instruments as easily as humans and to adapt to various types of audio signals. However, due to the complexity of the sound of an opera Zhuhu performance, the demand for audio feature construction for realistic instrumental performances cannot be met at present.

The study first introduces the basic theory of musical signals, including the basic concepts of musical signals and related acoustic parameters. The multimodal opera main beard performance is preprocessed by means of weighting, frame-splitting and windowing, and the feature vectors of the opera frequency are extracted using MFCC, while the global and local features of the feature vectors are extracted by using RCNN network, and the temporal features of the data are obtained using the bidirectional LSTM algorithm, and then the trained data are weighted by the self-attention mechanism to output the classification results. Finally, the effectiveness of the model proposed in this paper in classifying the emotion of the main beard performance of opera is demonstrated through feature extraction and music emotion recognition experiments.

2 Method

2.1 Fundamentals of music theory

2.1.1 Overview of musical signals

Musical signals are generally composed of periodic waveforms and are created by adding various sound waves of different frequencies. The lowest frequency sound wave among these sound waves is called the fundamental frequency of the musical signal, or fundamental tone. This is an essential element of a musical signal that determines the pitch of the musical notes produced by it. A musical signal is made up of many high-frequency sound waves known as overtones in addition to the fundamental frequency. The harmonics of the musical signal are composed of these high-frequency sound waves and the fundamental frequency.

2.1.2 Acoustic parameters of musical signals

(1) Pitch

The frequency of the sound we perceive is called pitch 49, and it is expressed in Hz. The sound is loud when the frequency is high and quiet when the frequency is low. In general, pitch is determined by the fundamental frequency emitted by the instrument, usually expressed as f. We can perceive pitch even if some instruments lack a fundamental frequency in their sound. The pitch of an instrument is usually indicated by a pitch mark symbol. In addition, markers consisting of numbers and symbols can also be used to indicate pitch. In this case, the number indicates the octave, and the symbol indicates the elevation.

(2) Tone Length

Tone duration refers to the duration of the musical signal, which can be expressed in seconds, minutes, hours and other units. Tone length is an important part of musical rhythm, and has an important influence on the rhythm and rhyme of the piece. One may think of the tone duration as the musical sound wave's time-domain characteristic. In terms of time, the musical wave passes through the four stages of the ADSR cycle, as shown in Figure 1: onset, decay, hold and release. Different musical instruments have their own unique articulation modes, which make the duration of each stage different, thus presenting different time domain envelope shapes.

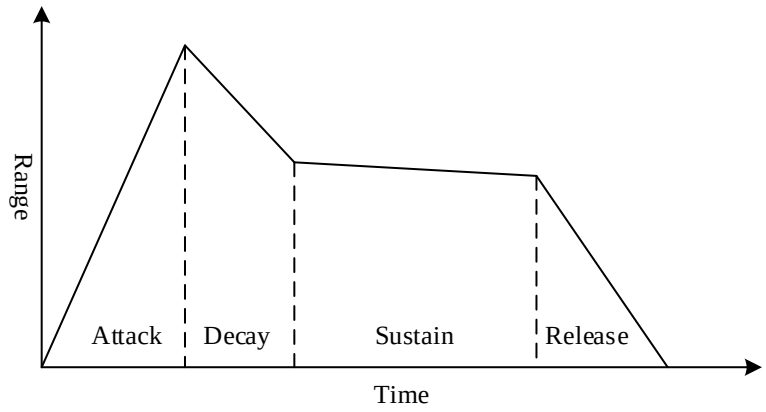


Figure 1: ADSR envelope

(3) Tone intensity

The loudness of the tone, which is determined by the organ's vibration amplitude, is referred to as tone intensity. The loudness of the tone generated is often reflected in the vibration amplitude. In real life, the loudness of sound is commonly referred to as volume, and decibels (dB) are used to quantify this loudness. The sound generated increases with the number of decibels. The amount of musical signal energy generated is indicated by sound intensity when recognizing musical instruments.

(4) Tone

Tone is the key feature to distinguish between different instruments, it allows the listener to distinguish between instrument types with the same pitch, duration and loudness. Different instruments have different combinations of fundamental frequencies and overtones, so they each have a unique timbre. The mutual strength and distribution of the harmonics of the musical signal are the main characteristics that make up the timbre of an instrument.

2.2 Pre-Processing for Opera Main Beard Playing

2.2.1 Pre-exacerbation

The signal's power decreases as its frequency increases. At this stage, appropriate methods should be used to repair the damaged signal in order to guarantee that the signal quality is improved. The high-frequency pre-emphasis should process the emotional signal to ensure that its high-frequency signal-to-noise ratio is enhanced. A first-order digital filter with the following equation (1) is used to achieve the high-frequency pre-emphasis:

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

where μ is the pre-emphasis factor, usually a decimal close to 1, and z is the input signal. The output representation of the signal at the nth moment after passing through the filter is shown in equation (2):

$$\overline{Y(n)} = Y(n) - \mu Y(n-1) \tag{2}$$

where $\overline{Y(n)}$ denotes the weighted output signal, and $Y(n)$ denotes the original input signal at the nth moment.

2.2.2 Splitting frames and adding windows

In terms of non-stationarity in long-duration signals, music signals are comparable to voice signals. Nonetheless, it is feasible to estimate the spectral characteristics and the eigenparameters as almost smooth over brief durations, such as frame lengths of about 10–30 ms. Consequently, in these situations, frame splitting procedures must be performed. To improve the frame continuity, the signal must also be windowed. Windowing refers to convolving the original signal using the selected windows, and the common windows include rectangular, Hanning, and Hamming windows, and the three kinds of window functions at the moment of n are shown in Eqs. (3), (4) and (5):

(1) Rectangular window

$$\omega(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & else \end{cases} \tag{3}$$

(2) Hanning Window

$$\omega(n) = \begin{cases} 0.5(1-\cos(2\pi n/(N-1))) & 0 \le n \le N-1 \\ 0 & else \end{cases} \tag{4}$$

(3) Hanming Window

$$\omega(n) = \begin{cases} 0.54-0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & else \end{cases} \tag{5}$$

Windowing is necessary to ensure that the dynamic musical data is accessible and undamaged. In most situations, the window shift would be around half of the window length. Both ends of the signal frame would be lighter than the central portion due to windowing. Nevertheless, the frames would appear to overlap rather than to intercept one another. The time interval between the initiation of two consecutive frames in this scenario is referred to as the frame shift and is usually considered as half of the duration of the frame, that is, 10 ms. The equation for determining the total number of frames of a musical signal is represented by (6):

$$N = \left[\frac{N_1 - N_0}{N_2 - N_0} \right] \tag{6}$$

where N represents the number of frames, N_0 represents the frame shift, N_1 represents the length of the music signal, and N_2 represents the frame length.

2.3 Musical Signal Correlation Characteristics

2.3.1 Time domain characteristics

(1) Zero Crossing Rate (ZCR)

The rate at which the time-domain signal crosses the zero axis in a second is known as the zero-crossing rate. It may be used to depict the harmonic content and rhythmic quality of the musical signal. The following formula is used to compute it:

$$ZC_i = \sum_{n=0}^{N-2} \left| \text{sign}\left[x_i(n) - x_i(n+1) \right] \right| \tag{7}$$

$$\text{sign}(x) = \begin{cases} 1, x \geq 0 \\ -1, x < 0 \end{cases} \tag{8}$$

where $x_i(n)$ denotes the ith frame of the musical signal and N is the frame length.

(2) Amplitude Envelope

The amplitude envelope describes the "time-volume" graph of the signal in the time domain, which represents the volume magnitude, volume decay, or growth of the audio signal over various time periods. The amplitude envelope is derived from the Root Mean Square Energy (RMSE), and its equation is expressed below:

$$A_{RMS} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} \left| x_i(n) \right|^2} \tag{9}$$

After the time-domain signal is subframed, the short-time root-mean-square energy of the time-domain signal is calculated for each frame to obtain the amplitude envelope.

(3) Autocorrelation coefficient

The autocorrelation coefficient can measure the correlation between the values of the same signal at different moments, and it is often used to describe the periodicity, repetitiveness, and smoothness of audio signals in the time domain, so it is applied to the timbre differentiation. The formula is as follows:

$$R_i(k) = \frac{1}{R_i(0)} \sum_{n=0}^{N-k-1} x_i(n) x_i(n+k) \tag{10}$$

(4) Short Time Energy

Short-time energy is often used for endpoint detection of audio signals, and refers to the time-domain energy sum of the audio signal within a very small window. The formula is as follows:

$$DSNL_i = \sum_{n=1}^{N} x_i^2(n) \tag{11}$$

2.3.2 Frequency domain characteristics

Frequency-domain feature calculations usually require a Fourier transform of the musical signal to normalize the amplitude spectrum. The magnitude spectrum can be used to represent the

distribution of energy in the signal as a function of frequency. Frequency domain features include power spectral density, spectral energy, spectral center of mass and spectral width.

(1) Power spectral density

The signal's energy distribution in the frequency domain is represented by the power spectral density, which is computed as follows:

$$P_i(k) = \frac{\left|X_i(k)\right|^2}{N} \tag{12}$$

where $X_i(k)$ is the signal amplitude spectrum of the i th frame and N is the frame length.

(2) Spectral Energy

The energy magnitude in each frequency segment is measured by the spectral energy. Specifically, the spectral energy may be used to describe the signal's strength and loudness; the mathematical formula for calculation is shown below:

$$E_i = \sum_{k=0}^{N-1}\left|X_i(k)\right|^2 \tag{13}$$

(3) Spectral Center of Mass

The spectral center of mass is an indicator that describes the frequency and energy distribution of a signal and can be used to characterize the brightness and darkness of a signal. The formula is as follows:

$$f_{scr_i} = \frac{\sum_{k=0}^{N-1} f(k)P_i(k)}{\sum_{k=0}^{N-1} P_i(k)} \tag{14}$$

where $f(k)$ denotes the frequency of the k th sampling point.

(4) Spectral Width

In signal processing, spectral width reflects the degree of signal spreading in the spectrum. Generally speaking, the wider the spectral width is, the more dispersed the signal is, and the spectral width can be used to measure the tonal characteristics of the signal. The formula is as follows:

$$f_{sw_i} = \sqrt{\frac{\sum_{k=0}^{N-1}\left(f(k) - f_{scr_i}\right)^2 P_i(k)}{\sum_{k=0}^{N-1} P_i(k)}} \tag{15}$$

2.3.3 Characterization of the inverted spectral domain

(1) Mel inverse spectral coefficient

The Mel frequency is a psychoacoustic frequency unit, related to the human auditory scale, which uses a nonlinear frequency resolution based on the auditory scale, and can be converted between linear and Mel frequencies using Equation (16).

$$Mel(f) = 2596 \times \lg\left(1 + \frac{f}{700}\right) \tag{16}$$

The extraction of MFCC features is mainly divided into several steps:

(a) Signal pre-processing: Pre-emphasis, framing, and windowing of the signal are examples of signal pre-processing. Pre-emphasis is the process of adding a high pass filter to the input signal in order to highlight its higher frequencies. The following is the High Pass Filter Transfer Function:

$$H(z) = 1 - uz^{-1} \tag{17}$$

where u is the pre-emphasis coefficient. After pre-emphasis, the signal is divided into frames, the frame length is generally taken as 10-30ms, the frame shift is taken as 1/2 or 1/3 of the frame length, and the window function is taken as Heming window.

(b) The short-time Fourier transform is performed on each frame of signal $x_i(n)$ obtained after preprocessing to obtain the spectrum $X_i(k)$. The Fourier transform formula is as follows:

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) e^{-j\left(\frac{2\pi}{N}\right)kn} \tag{18}$$

(c) Calculate the number of groups of Mel filters and process each frame of the spectrum through the Mel filter to obtain the Mel spectrum and calculate the logarithmic energy according to Eq. Equation (19) is expressed as:

$$S_i(m) = \log_{10}\left(\sum_{k=0}^{N-1} \left|X_i(k)\right|^2 H_m(k)\right), 0 \le m \le M \tag{19}$$

where k is the spectrum corresponding to the frequency domain of the signal, M is the number of Mel filters, and $H_m(k)$ is the transfer function of the m th Mel filter.

(d) Calculate the discrete cosine function according to Eq. (20) to obtain the cepstrum. Equation (20) is expressed as:

$$MFCC_i(n) = \sqrt{\frac{2}{N}} \sum_{m=0}^{M-1} S_i(m) \cos\left[\frac{\pi n(2m-1)}{2M}\right] \tag{20}$$

where n denotes the order of the MFCC, which is usually taken to be 12-16, and $MFCC_i(n)$ denotes the MFCC coefficients of the i th frame.

(2) Linear prediction cepstrum coefficient

LPCC is based on the linear prediction coefficients (LPC), and then the cepstrum coefficients of the signal characteristics. The full-polar model of the linear prediction coefficient channel model is expressed as:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-1}} \tag{21}$$

where p is the order of linear prediction and a_k is the prediction coefficient. The inverse of LPC is further taken as follows:

$$\begin{cases} c(1) = a_1 \\ c(n) = a_n + \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)a_k c(n-k), 1 < n \le p \\ c(n) = \sum_{k=1}^{p}\left(1-\frac{k}{n}\right)a_k c(n-k), n > p \end{cases} \tag{22}$$

Eq. (22) is further simplified to Eq. (23):

$$LPCC_i(n) = LPC_i(n) + \sum_{k=1}^{p}\left(\frac{n-k}{n}\right)LPCC_i(n-k)LPC_i(k) \tag{23}$$

where $LPCC_i(n)$ is the linearly predicted cepstrum coefficient of the ith frame of the signal.

2.4 Emotion Recognition Model Design for Opera Main Beard Performance

2.4.1 Multimodal attention fusion mechanisms

In the algorithm of this paper, a variety of information is used to analyze the sentiment of the main beard performance of the opera, so it is also necessary to use the attention mechanism (AM)7-8 to merge and quantify the information of the multimodal data in order to improve the accuracy of the sentiment prediction results. The attention mechanism used is a ternary function, whose inputs are query, key, and value, and the results are output according to the mapping relationship.

In the attention mechanism, summed attention and dot product attention are usually used for computation, and the query, key and value are denoted by Q, K, V respectively. The formulae for the summation and dot product attention mechanisms are:

$$\alpha_i = f\left(Q, V_i\right) = \frac{\exp\left(h\left(Q, V_i\right)\right)}{\sum_{j=1}^{N}\exp\left(h\left(Q, V_j\right)\right)} \tag{24}$$

$$o = f(Q, K)V = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{25}$$

where: $h(\cdot)$ is the implicit layer function. d is the Euclidean distance between Q and K. Meanwhile, in this paper, we use modal enhancement algorithm to enhance the information interaction between modes by using self-attention mechanism, i.e., interconnection by residual module to get the results. The modal features connected by residuals can be expressed as:

$$o' = \sigma_o\left(\left(o^{update} + o\right); \theta_o\right) \tag{26}$$

where: σ_o denotes the activation function of the fully connected layer of the residual module. θ_o is the weight value.

2.4.2 RCNN feature extraction network

Recurrent Convolutional Neural Networks (RCNN) 9-11 is a hybrid neural network that combines RNN and CNN.The model improves on CNN connectivity by using multiple layers of CNNs to accomplish the recursive operation, which in turn forms the form of an RNN network, which is then connected using a recursive convolutional layer (RCL).The RCL iterates over the length of the paths of all of the convolutional networks and the shortest paths go through only the forward feedback convolutional network, and the longest traverses all of the $T+1$ layer neural networks. Therefore, the traversal process of the RCL network inherits some of the characteristics of the recurrent neural network, and the RCL traversal process is shown in Figure 2.

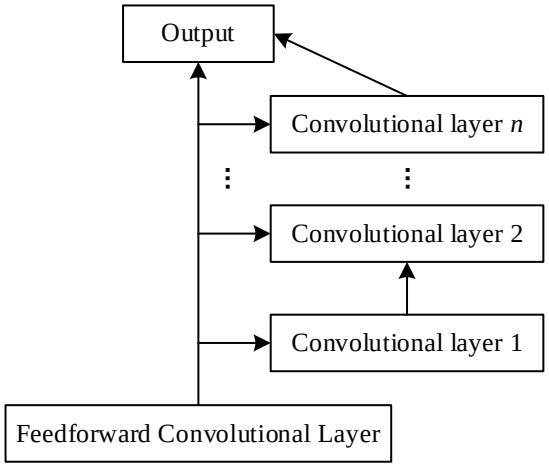


Figure 2: The RCL traversal process

The flow of RCL network information can be expressed as follows:

$$z_{ijk}(t)=\left(\mathbf{w}_k^f\right)^{\mathrm{T}}u^{(i,j)}(t)+\left(\mathbf{w}_k^r\right)^{\mathrm{T}}x^{(i,j)}(t-1)+b_k \tag{27}$$

where $\mathbf{w}_k^f$ and $\mathbf{w}_k^{\mathrm{r}}$ are the weight values of the front convolutional and recurrent networks, respectively. b_k is the bias value. $u^{(i,j)}$ denotes the feedforward network input information. $x^{(i,j)}$ denotes the iterative input information.

2.4.3 Bi-LSTM-based pre- and post-timing feature extraction

The RCNN model can obtain the initial features of the opera Zhuhu performance, i.e., the opera Zhuhu performance is divided into frames in the preprocessing stage. In order to get the correlation between different frames, the temporal features of the performance of the opera master are trained using a bi-directional LSTM model.

LSTM is also known as long and short term memory network. The network consists of three special structures: input gates, output gates and forgetting gates, in addition to internal states and candidate states.

The formulas and states involved in LSTM network are shown below:

$$i_t=\sigma\left(W_i\left[h_{t-1},x_t\right]+b_i\right) \tag{28}$$

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{29}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{30}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c_t \tag{31}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{32}$$

where i_t, o_t and f_t denote the data corresponding to the input gate, the output gate, and the forgetting gate, respectively. c_t is the neuron state. h_t is the input value of the unit at moment t. W_i, W_o, W_f, and b_i, b_o, b_f denote the weight values and bias values of the input, output, and forgetting gates, respectively. To further explore the correlation between the data, the past and future audio features are extracted using a bidirectional LSTM (Bi-LSTM) network.

2.4.4 Modeling

Fig. 3 shows the algorithm's complete design. Preprocessing, RCNN feature training, temporal feature training, and multimodal feature fusion modules make up the four components of the algorithm. The modality is divided into audio and text data, and the former is the information about the main beard performance of the opera in the dataset, while the latter is the background information of the main beard performance of the opera. Firstly, the preprocessing module preprocesses the data of the opera main beard performance according to the steps in Fig. to get the related spectral information. At the same time, the RCNN feature training module also preprocesses the feature information to obtain the features of global and local data. And the timing feature training module is responsible for extracting the timing information of different opera Zhuhu performance frames. Finally, the multimodal data are fused using the self-attention mechanism to obtain the emotion labels of the opera Zhuhu performance.

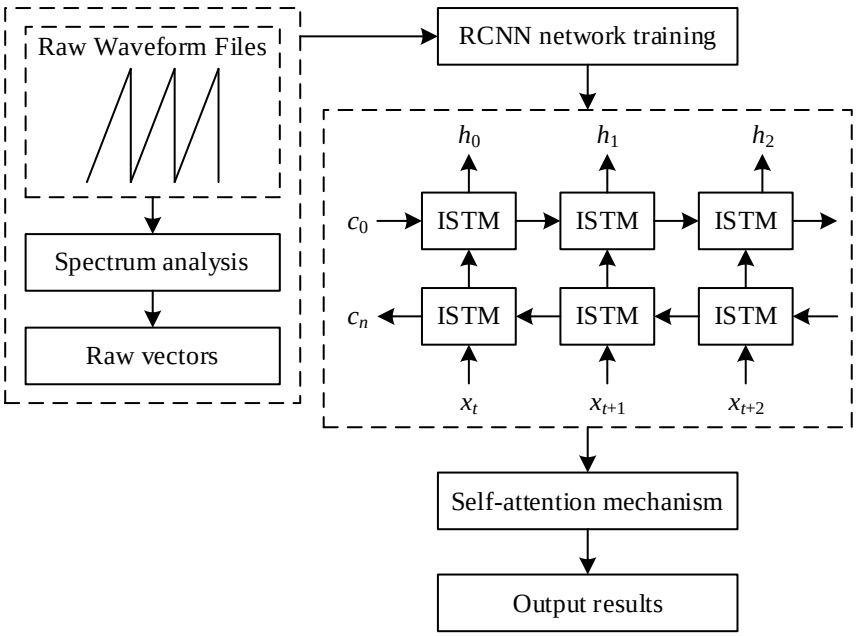


Figure 3: Improve the structure of the model

3 Results and Discussion

3.1 Experimental setup

3.1.1 Data sets

In this paper, we will use two datasets for experimental comparison, CASIA data and 1,000 opera Zhuhu performance data, which are taken from the opera “Jiangnan Rain”. Ten thousand phrases with different pronunciations of the six emotions—neutrality, happiness, anger, sorrow, fear, and surprise—make up the CASIA data, a Chinese emotion dataset gathered by an expert researcher from a research organization. The emotional intensity of feelings is shown by the CASIA data. This database has important experimental value for emotional analysis and acoustic characteristics of different emotions. 500 distinct emotion datasets with identical content were chosen for slicing and other preprocessing tasks in this paper. The audio file is sliced every 30 seconds to produce 1,400 speech samples, each with 1,620 frames. The sliced data is then divided into the training set and the testing set at an 8:2 ratio. The sampling rate of the audio is set at 16 kHz. The data for the performance of opera main beard has rich emotive elements, and the data of the performance of opera main beard used in this experiment has been collected from online audios of the web and has been preprocessed using slicing, etc., with each segment having a time of 30s, totaling 1,000 segments, with the training set and the test set split at the proportion of 8:2. For the sake of valid comparison experiments with the emotions of the CASIA data, six emotions: happy, neutral, angry, sad, scared, and surprised are still used to classify the emotions of the opera main beard performance.

3.1.2 Evaluation criteria

Having a metric that can be used to measure the outcome of the classification process becomes crucial while working with the classification challenge. In this experiment, a set of pertinent metrics for the task, a percentage of checking correctness (P), checking completeness (R), and F1 value, will be used to assess the results.The term "accuracy" can be used since precision is a measure of the percentage of anticipated accurate values among all properly identified objects. Recall is also known as checking accuracy, which indicates the percentage of anticipated accurate values among all correct values. In contrast, the F1 score is a weighted average of the first two metrics, giving the lower one more weight.

3.2 Experiments on Feature Extraction of Opera Main Beard Performance

3.2.1 Experimental environment

The hardware environment used in the experiment as mentioned in this chapter is as follows: intel core(TM) i7-3770 CPU @3.4 GHz CPU with 16 G RAM. The software environment considered here includes working under a Windows 10 64 bit operating system in an environment such as Anaconda with the use of programming languages like python with various packages.

3.2.2 Feature Extraction Experiments

In this chapter, the experimental music feature extraction part, all of which is based on the opera “Jiangnan Rain” as a sample, mainly focuses on two aspects of the extracted feature spectrograms and feature parameters. A total of 36-dimensional timbre features of MFCC and its difference are extracted, a total of 6-dimensional pitch attribute features related to the fundamental frequency are extracted, as well as a total of 12-dimensional sound quality features

of the first, second, and third resonance peaks, and a total of 10-dimensional feature parameters of the time-frequency domain features that represent the essential attributes of music. In this paper, the length of each frame of the opera music samples used in the experiment is 1035, and the frame shift length is set to be 1/2 of the frame length, and the time of each frame of the music samples in this paper can be known as 48ms through the calculation, then theoretically, there are 1,324 frames of the 30s music clips in the music library of this paper, but in fact, it only contains more than 1,295 frames.

1. MFCC

In this case, the value of L is fixed at 12 for the acquisition of Mel frequency cepstral coefficients of music signals and the differential of first order and second order MFCCs. Generally, the value of L of MFCCs is in the range of 12 to 16. Figure 4 shows the spectrogram of MFCC characteristics. The X-axis represents the time span of music clips, whereas the Y-axis shows the frequency information of the signal.

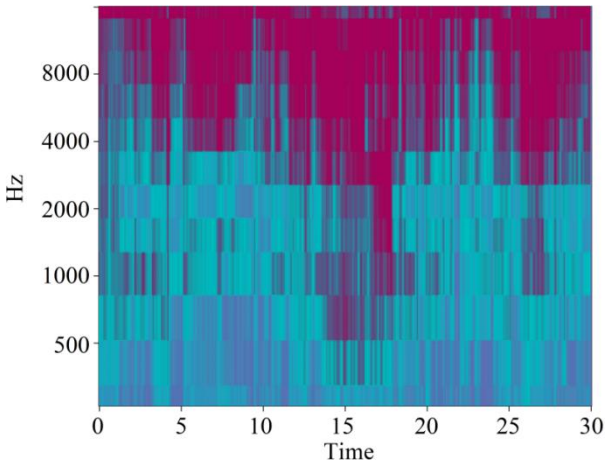
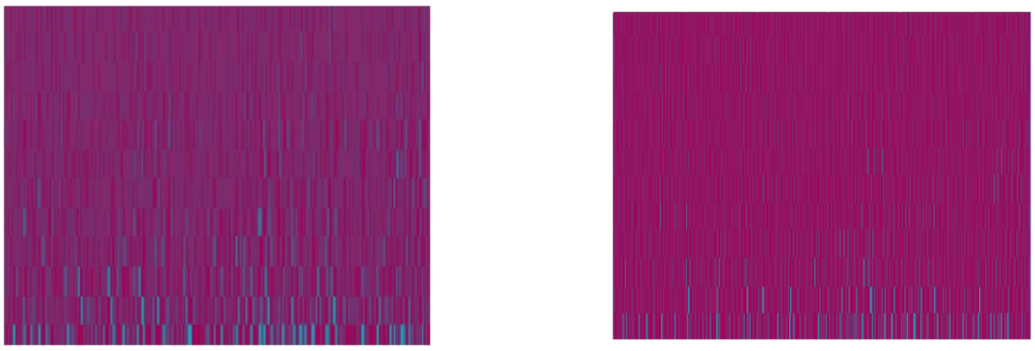


Figure 4: MFCC characteristic spectrum

The MFCC first-order difference and second-order difference feature spectra are shown in Fig. 5, with MFCC-delta in Fig. a and MFCC-delta in Fig. b. The values of the relevant parameters of the MFCCs are obtained by comparing the colors at the specified time points and frequencies on the feature spectra with the color plates, and thus obtaining the corresponding parameter values.



(a) MFCC-delta (b) MFCC-delta2

Figure 5: MFCC first-order and second-order difference characteristic spectra

Information on the relevant characterization parameters of MFCCs is shown in Table 1.

Table 1: Characteristic parameter information related to MFCCs

12-dimensional MFCC	MFCC	The first-order difference of MFCC	The second-order difference of MFCC
MFCC_0	-38.14	-0.8037	0.4675
MFCC_1	28.93	0.5952	-0.0898
MFCC_2	0.08	-0.1734	0.2996
MFCC_3	7.8	-0.12	0.1864
MFCC_4	1.29	-0.3041	0.0967
MFCC_5	4.63	0.052	-0.015
MFCC_6	-0.87	-0.2297	-0.0676
MFCC_7	1.22	0.4177	-0.2321
MFCC_8	-3.89	-0.0722	-0.0346
MFCC_9	1.33	0.0626	-0.2131
MFCC_10	-2.96	0.414	-0.3484
MFCC_11	-0.53	0.1079	-0.0102

2. Keynote Frequency

To a certain extent, the fundamental frequency can reflect the pitch of music, and the change of pitch can also reflect the change of music emotion. When the music emotion is relatively low and sad, the fundamental frequency is relatively low and does not change much. Through the detection of fundamental frequency, the value and change of music fundamental frequency can be observed through the fundamental frequency spectrogram, and the spectrogram of fundamental frequency characteristics is shown in Fig. 6, from which it can be seen that the fundamental frequency of the clip basically stays around 130Hz, and the range of change is not big.

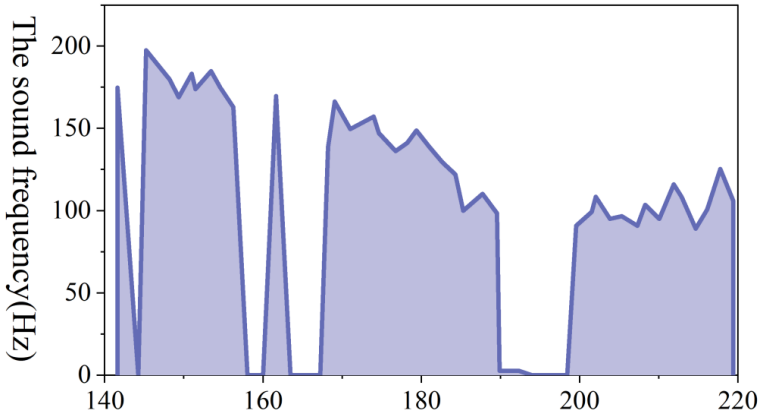


Figure 6: Phonetic gram

3. Resonance peaks

Because the resonance peaks can reflect the sound quality of opera music, the relevant characteristic parameters of the first, second and third resonance peaks are mainly extracted in this section of the experiment. Overall, the distribution of the three resonance peaks in the frequency domain is increasing, i.e., the first resonance peak is generally lower than the second and third resonance peaks. The characteristic parameters of the resonance peaks are shown in Table 2.

Table 2: Characteristic parameters of formant peaks

Resonance peak	Average value	Median	The median bandwidth occupied	Standard deviation
First formant	1.36E+03	1.32E+03	2.69E+03	281.4924
The second formant	2.34E+03	1.95E+03	2.69E+03	354.7018
The third formant	3.26E+03	3.37E+03	2.20E+03	326.9084

4. Time and Frequency Domain

The time domain and frequency domain information of the characteristic music signal can greatly respond to the essential information of the music signal. In this paper, when considering the attributes of music timbre, sound quality and pitch, we also add the consideration of the characteristics of the original music signal, and mainly extract the 10-dimensional characteristic parameters of the average, maximum, variance, center of gravity and intensity of the time-frequency domain. Fig. 7 displays the music signal's time domain waveform. Figure 8 displays the music signal's frequency spectrum.

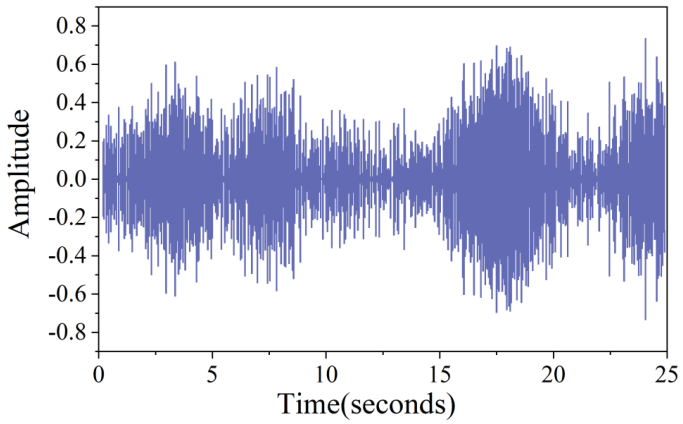


Figure 7: The time-domain waveform diagram of the music signal

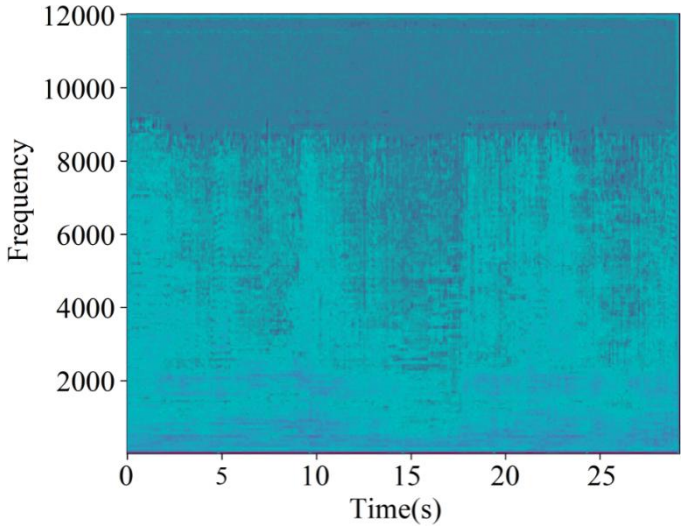


Figure 8: The spectrogram of the music signal

3.3 Experiments on Emotion Recognition in Opera Main Beard Performance

The four models are assessed using CASIA datasets in the studies that follow. First, two sets of experiments, the CNN_LSTM model and the AC_BiLSTM model, were used to validate the emotion recognition and classification capabilities of the suggested audio emotion recognition and classification model with improved multimodal RCNN. A single set experiment was used to confirm the superiority of the suggested model.

For the CASIA dataset, the classification parameters for emotion recognition are compared for CNN-LSTM, AC-BiLSTM, MHAtt-ResNet, and the proposed method. Classification parameters (%) for emotion recognition using different methods are summarized in Table 3. It can be seen from Table 3 that the proposed method has the highest accuracy for classifying emotions.

Table 3: Comparison of emotion classification parameters of different methods

Classification method	P	R	F1
CNN_LSTM	64.87	58.52	63.4
AC_BiLSTM	82.9	83.12	81.28
MHAtt_ResNet	70.01	73	69.85
Ours	83.75	82.62	85.29

Table 4 illustrates the classification accuracy of each emotion with various approaches. From the experimental result, it is clear that the classification accuracy of the proposed model is higher compared to the other three models, with a classification accuracy rate of 84.74%.

Table 4: The accuracy rates of various emotion classifications by different methods

Classification method	Happy	Neutral	angry	Sadness	Fear	Surprised	Accuracy rate(%)
CNN_LSTM	66.34	57.64	61.7	64.72	70.13	68.7	62.75
AC_BiLSTM	79.93	68.36	85.13	85.36	84.76	84.21	81.5
MHAtt_ResNet	78.66	63.48	70.12	73.23	65.52	71.48	70.38
Ours	93.39	70.78	83.28	89.47	86.03	85.56	84.74

After verifying the effectiveness of the model in this chapter for audio sentiment analysis of CASIA data, this group of experiments will test the model in this chapter with the data of opera main beard performance, and the comparison of the model accuracy is shown in Table 5. The results can be seen, the model in this paper compared to the MHAtt_ResNet model overall classification accuracy increased by 13.4%, the performance has been greatly improved, all types of emotions also have high recognition results, compared to several other emotions.

Table 5: Comparison of model accuracy

Method	Happy	Neutral	angry	Sadness	Fear	Surprised	Accuracy rate
MHAtt_ResNet	81.74	64.08	71.92	72.9	72.86	70.08	71.8
Ours	93.6	70.35	85.79	90.71	88.13	88.31	85.2

The results of comparing the accuracies of the two algorithms are provided in Fig. 9. In Fig. 9, we have depicted the trend of the accuracy of the two models for sentiment analysis in the case of opera main beard performance data with an increase in the number of iterations. It can be seen that the accuracies of the two models for sentiment analysis improve with the increase

in the number of iterations, and after a certain number of iterations, they tend to converge. From Fig. 9, it can be observed that the MHAtt_ResNet model begins to converge slowly after 38 iterations, whereas the model proposed in this paper converges slowly after 40 iterations, which indicates the fact that the proposed model is more capable of recognizing long time series, and therefore, the proposed model outperforms the other model in terms of sentiment analysis of opera main beard performance.

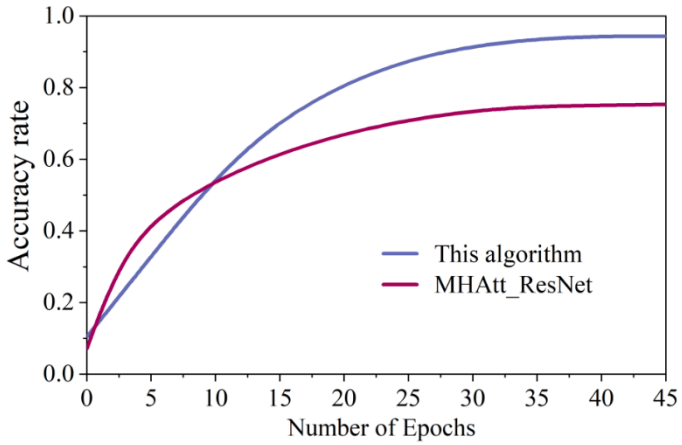


Figure 9: Comparison of accuracy rates of the two methods

This section concludes by giving the visualization results of the model's sentiment classification accuracy using the confusion matrix.Confusion Matrix separates the accuracies through the color's brightness level. As seen in Fig. 10, the emotion labels on the vertical axis represent the predictions derived from the test music clips, while those on the horizontal axis represent the actual emotions of the music clips. The anticipated probabilities of the music clips are represented in the matrix using various brightness levels, where a greater probability value is represented by a brighter color and a lower probability by a darker color. The right-side of the graph shows the probability levels indicated by the color brightness scales, where brighter color represents high predicted probability while darker color represents smaller predicted probability. Higher brightness indicates higher probability of predicting the particular emotion while darker color indicates lower probability. Analysis of color brightness of the matrix indicates that the accuracy of model in recognizing various emotions in the main beard performance of opera is very high and it has very low probability of being mistaken to be other emotion. From analysis of accuracy of the emotions, it is evident that "happy" emotion classification is the most accurate because there are very high differentiations between "happy" emotions and other features.

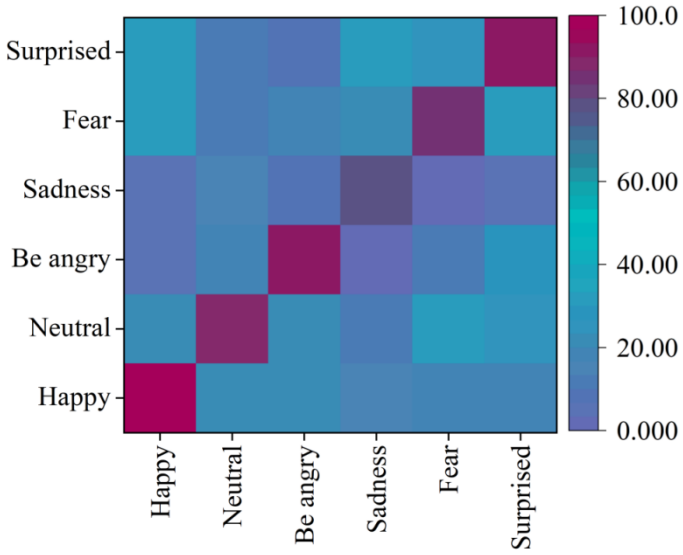


Figure 10: Confusion matrix

4 Conclusion

Accompanied by the development of national economy, the protection and inheritance of opera has been paid more and more attention. This paper proposes a multimodal model design scheme for the emotion recognition and classification algorithm of the zhuhu performance of opera based on the improved RCNN network, and verifies the effectiveness of this paper's model for the emotion recognition of the zhuhu of opera through experiments. The experimental conclusions drawn from the article are as follows:

(1) In the feature extraction experiments of opera Zhuhu performance, it is found that in the spectrogram of fundamental frequency features, the fragment fundamental frequency is basically kept around 130Hz, and the overall range of change is not large.

(2) The classification accuracy rate of the suggested approach is 13.4% higher than that of the MHAtt_ResNet model, according to an analysis of the models' accuracy. This is a great improvement in accuracy.

(3) The best accuracy rate for classifying emotions is achieved for happy emotions, and the overall accuracy rate for identifying emotions of other opera lead-beard playing is also quite high, as can be seen from the use of a confusion matrix to display the model's classification accuracy rate of emotions.

About the Author

Chengyao Duan, Hubei Normal University Music College associate professor, Supervisor of postgraduate. Teaching for more than 10 years, has presided over and participated in more than 10 projects in Hubei Province, city and level, wrote and published more than 20 academic papers, 8 papers above the core level, published 2 academic works and 1 academic textbook; 2 patents; My main research direction is Ethnomusicology and Intangible cultural heritage protection and research.

References

[1] Carron, M., Rotureau, T., Dubois, F., Misdariis, N., & Susini, P. (2017). Speaking about sounds: a tool for communication on sound features. Journal of Design Research, 15(2), 85-109.

[2] Lee, J., & Lee, J. S. (2018). Music popularity: Metrics, characteristics, and audio-based prediction. IEEE Transactions on Multimedia, 20(11), 3173-3182.

[3] Nadírova, A., & Ayapova, H. (2023). MUSICAL SOUND AND ITS CHARACTERISTICS. Modern Science and Research, 2(12), 1170-1173.

[4] Panda, R., Rocha, B., & Paiva, R. P. (2015). Music emotion recognition with standard and melodic audio features. Applied Artificial Intelligence, 29(4), 313-334.

[5] Kroher, N., & Díaz-Báñez, J. M. (2018). Audio-based melody categorization: Exploring signal representations and evaluation strategies. Computer Music Journal, 41(4), 64-82.

[6] Han, K. H. (2009). CHINESE MUSICAL INSTRUMENTS: A HISTORICAL ACCOUNT. Kaleidoscope of Cultures: A Celebration of Multicultural Research and Practice, 63.

[7] Ni, Y. (2021). The modern Erhu: Perspectives on gender, education, and society in the development of Erhu performance. Kent State University.

[8] Qiaoyi, Y. (2024). Revitalizing tradition: The role of the Erhu in modern music and global culture. International Journal of Education and Humanities, 4(4), 512-521.

[9] Cai, X., & Zhang, H. (2022). Music genre classification based on auditory image, spectral and acoustic features. Multimedia Systems, 28(3), 779-791.

[10] Juan, L., Jirajarupat, P., & Yinghua, Z. (2023). The Transmission of Guqin Musical Instrument Knowledge Literacy and Its Reflection Study in Guizhou Province, China. International Journal of Education and Literacy Studies, 11(2), 22-29.

[11] Duan, Z., Pardo, B., & Daudet, L. (2014, May). A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7495-7499). IEEE.

[12] Hung, Y. N., Chen, Y. A., & Yang, Y. H. (2019, May). Multitask learning for frame-level instrument recognition. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 381-385). IEEE.

[13] Mesaros, A., Heittola, T., & Virtanen, T. (2018, September). Acoustic scene classification: an overview of DCASE 2017 challenge entries. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 411-415). IEEE.

[14] Zhao, Z., Liu, H., & Fingscheidt, T. (2018). Convolutional neural networks to enhance coded speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(4), 663-678.

[15] Han, Y., Kim, J., & Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1), 208-221.

[16] Slizovskaia, O., Gómez, E., & Haro, G. (2017, June). Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (pp. 226-232).

[17] Liu, J., & Xie, L. (2010, May). SVM-based automatic classification of musical instruments. In 2010 international conference on intelligent computation technology and automation (Vol. 3, pp. 669-673). IEEE.

[18] Abeßer, J., Chauhan, J., Pillai, P. P., Taenzer, M., & Mimilakis, S. I. (2021, August). Predominant Jazz Instrument Recognition: Empirical Studies on Neural Network Architectures. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 361-365). IEEE.

[19] Li, R., & Zhang, Q. (2022). Audio recognition of Chinese traditional instruments based on machine learning. Cognitive Computation and Systems, 4(2), 108-115.

[20] Yang, J., Gao, F., Yun, T., Zhu, T., Zhu, H., Zhou, R., & Wang, Y. (2025). A Deep-Learning Framework with Multi-Feature Fusion and Attention Mechanism for Classification of Chinese Traditional Instruments. Electronics, 14(14), 2805.

[21] Jin, R., Zhou, Y., & Wu, H. (2024, December). The Comparison of Different Deep Learning Models for Chinese Musical Instrument Recognition. In 2024 IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISP) (pp. 1-5). IEEE.