Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7832548, 9 pages
https://doi.org/10.1155/2022/7832548

Research Article

Music Emotion Recognition Based on Bilayer Feature Extraction

Chen Wang[1] **and Yu Zhao**[2]

[1]*Xi'an University, Xi'an, 710000 Shaanxi, China*
[2]*Guangxi Arts University, College of Music Education, Nanning, 530022 Guangxi, China*

Correspondence should be addressed to Yu Zhao; 20100010@gxau.edu.cn

Received 7 April 2022; Revised 19 May 2022; Accepted 10 June 2022; Published 4 July 2022

Academic Editor: Zhiguo Qu

Music is a kind of art which is to express the thoughts and emotion and reflect the reality life by organized sounds; every piece of music expresses emotions through lyrics and melodies. Human emotions are rich and colorful, and there are also differences in music. It is unreasonable for a song to correspond to only one emotional feature. According to the results of existing algorithms, some music has obviously wrong category labels. To solve the above problems, we established a two-layer feature extraction model based on a spectrogram from shallow to deep (from primary feature extraction to deep feature extraction), which can not only extract the most basic musical features but also dig out the deep emotional features. And further classify the features with the improved CRNN neural network, and get the final music emotion category. Through a large number of comparative experiments, it is proven that our model is suitable for a music classification task.

1. Introduction

In terms of technology, the 21st century is an era of rapid development of information technology. As the core of information technology, artificial intelligence has promoted the vigorous development of various emerging technologies [1, 2]. From the perspective of life, with the development of economy and society, people put forward higher requirements for spiritual life, and people's entertainment life is more diversified. Based on the above two aspects, people have higher requirements for music appreciation. Music is a special social ideology. With a strong appeal, it can not only edify sentiment and regulate emotions but also develop thinking and coruscate emotions. Driven by technology, the digitization of audio has become an intermediate force in the digital trend. Audio is the product of digital processing and preservation of sound, and music is an important component of audio. In recent years, with the rapid development of digital devices, digital music has attracted a large number of scholars to carry out research on it. Compared with traditional music, the advantages of digital music are reflected in all aspects of music creation [3]. First, digital signal processing has lowered the cost of storing music, allowing more people to create it. Secondly, the application and popularity of the Internet has promoted the spread of digital music. The most important is that traditional music is based on text retrieval. With the development of audio retrieval technology and the rapid growth of music data, the traditional way is gradually difficult to meet the needs of users [4]. The retrieval based on audio content is gradually rising, and the research on digital audio content has gradually become the hot spot of audio digital technology.

However, from the perspective of music appreciation, its essential feature is musical emotion, and the biggest advantage of music is to arouse people's emotional resonance. Research on music information behavior indicates that emotion is an important criterion used by people in music retrieval [5]. The core of music retrieval based on emotion is music emotion recognition. As one of the important research directions of digital music, it is a multidisciplinary field. Research on musical emotion recognition covers not only musicology, musical acoustics, and audio signal processing but also psychology, natural language processing, machine learning, and other fields [6]. Scientists have conducted a number of relevant studies, and the results show that different audiences are generally consistent in judging the emotions expressed in music. Therefore, emotion recognition with high precision becomes possible.

The research on musical emotion recognition began in the middle of the 20th century and has a history of several decades. At the beginning of artificial intelligence, some people proposed the relationship between music content and emotion, and then, many universities and research institutes around the world have carried out relevant research. In recent years, with the development of artificial intelligence, music emotion recognition research progress is rapid and is successfully applied in various fields, including music emotion retrieval, music emotion recognition of the performing arts, and smart space design. Music emotion recognition technology has paved the way for the road in the field of music visualization research [7], and it has the extensive research prospects and important application value. At present, there are more and more researchers in the field of music emotion recognition, and the research results are updated quickly. However, because of the complexity of music emotion information, music emotion recognition is a long-term research goal [8]. Through the analysis of music emotion, the digitization process of music is promoted and the spiritual experience of music is improved.

2. Related Work

Musical emotion is a psychological process that conveys information in a unique way through music. It contains all kinds of human emotional factors generated in the interaction between people and music [9]. The purpose of music emotion recognition is to analyze music file data and the audio characteristics related to music emotion by computer means and to construct an emotion classification model according to a certain emotional psychological model, so as to establish corresponding relationship between music and various emotions and to provide services for the users [10]. The recognition of music emotion is a pattern recognition, and the recognition process needs to correctly map data from high-dimensional music feature space to low-dimensional emotion space. With the arrival of the era of big data, how to make a good job in a product, to optimize product structure, to meet user needs, and to improve user experience has become the primary task of Internet companies.

Digital music refers to a new type of music art created with computer digital technology, stored in digital format, and disseminated through the Internet and other digital media technologies [11]. Digital music is very different from traditional music in the way it processes, produces, and organizes sound. Digital music has a variety of sound effects, which break through the limitation of the number of timbre by means of the point oscillator. With a variety of creative ways, through human-oriented computer for the assistance or computer simulation of people's thinking to create, it has broken through the conventional way of creation. With timeliness and influence, digital music is spread through the Internet. With the rapid development of information technology, digital music has broken through the limitation of time and space [12].

In the research of musical emotion recognition, many research institutions have carried on this research. Xia et al. [13] used the continuous emotional psychological model and regression prediction model to drive the generation of robot dance movements according to the constructed emotional changes and music beat sequence. Schmidt and Kim [14] selected the continuous emotion model, connected the emotional content of music with the acoustic feature model, and established a regression model to study the change of musical emotion over time. Sordo et al. [15] studied a variety of acoustic features, including low-level features, melody, tone, and high-level genre, style, and other features, and then reduced these features to the D-dimensional space and associated them with semantic features, using k-nearest neighbor algorithm for automatic recognition of musical emotion. Bresin and Friberg [16] invited 20 music experts to express different emotions such as happiness, sadness, fear, and calm by controlling the numerical combination of 7 characteristic quantities such as rhythm and timbre in the device and obtained the relationship between characteristic quantities and musical emotions. Kuo et al. [17] established a recommendation model based on musical emotion and mainly studied the emotion conveyed by film music. Yang et al. [18] used the continuous emotional psychological model and regression modeling to predict the emotional value of music and used two fuzzy classifiers to measure the emotional intensity, so as to identify the emotional content of music. Sarkar and Saha [19] proposed the use of a convolutional neural network to identify the music model and compared it with the BP neural network and other commonly used classifiers. Yang et al. [20] established a multimodal music emotion classifier, mainly studying the underlying features of audio and song text features. On this basis, a music sorting algorithm based on a two-dimensional emotion model is established [21].

With the deepening of research, scholars studied the music label relabeling technology based on underlying features by using principal component analysis for CAL500 [22] and CAI 10K [23] data sets [21]. Considering the tagged words of Internet songs for emotional classification, Lin et al. [24] first made genre classification and then emotional classification. Zhang and Sun [25] carried out emotional recognition of web music and classified the music into a corresponding emotional label. Yan [26] studied multilabel music emotion recognition technology based on evidence theory and semantic cell model based on timbre and rhythm characteristics and adopted principal component analysis for feature dimension reduction. Wang et al. [27] studied the song emotion recognition algorithm based on a spiral model, extracted emotional features such as tone strength and speed, and adopted hierarchical K-nearest neighbor classifier to construct a spiral emotion model to realize the classification of song emotion. Sun [28] carried out research on key technologies of automatic analysis of music elements, mainly studying musical theory features such as mode, chord, and style. Liu [3] studied the melody extraction and segmentation technology of MIDI files and used the method of improved gene expression programming to build a music emotion cognitive model and compared it with the BP

neural network and other classification algorithms. Hu [7] took MIDI files as the research object, constructed a music emotion classification model based on the BP neural network, and analyzed the influence of music duration on classification accuracy. Zhuang and Ye [29] designed a neural network model combining an attention mechanism with a long- and short-term memory network for music emotion recognition and completed music recommendation with a convolutional cyclic neural network.

Compared with traditional music, digital music, with the help of high-speed development of digital technology, has formed new characteristics of the times. Digital music is virtual [30]. It is a virtual sound generated by digital technology, rather than a real performance or natural sound [31]. Digital music has the characteristics of interactivity and real-time transmission so that people can enjoy or create music at any time; digital music is popular and easy to transmit, store, and edit features, which greatly reduce the difficulty of creation; digital music has zero loss, is independent of material medium, and has no physical form, so it will not be lost over time. Although scholars have done a lot of researches in the field of music recognition, due to the rapid development of the Internet and a wide variety of music, traditional models need to be constantly improved to adapt to new music types, and relevant technologies also need to be further improved. At present, it is still difficult to construct training sets in this field, and there are few open musical emotion data sets. Most of the musical data studied are recorded or collected by themselves. The existing music feature analysis technology is limited in the application of basic music theory knowledge, so it is one of the difficulties to establish a reasonable music feature analysis model. The premise of music emotion recognition is to establish an appropriate emotional psychological model as a classification standard [32]. In the current emotional recognition system, there is no unified standard, and the emotional psychological model still needs to be further studied. What is the most important is that the current common method of music emotion classification is to extract the underlying physical features of music and to analyze and process these features. But the correlation between the underlying features and the high-level emotions is limited. In view of the above problems, this paper first analyzes the structure of audio files, so as to make the computer extract the relevant features through programming and then complete the things like extraction of the main melody and music emotion characteristic analysis through a series of algorithms.

3. Musical Emotion Recognition Model

Before classifying music emotion, it is necessary to establish a suitable emotional psychological model. That is, how many kinds of emotions can music be divided into, what are the dividing methods, and what are the connections or differences between different emotions. In the end, scientifically analyze the only major emotion of each piece of music and choose the number of emotion types reasonably in the process. Each category should be logically divided, and the different categories should be as comprehensive and nonoverlapping as possible. After the establishment of the emotional psychological model, we should construct the emotional classification model. That is, to judge after determining the input of emotional characteristics, the music will be classified into what type and what classification method. This paper analyzes several features such as average pitch, average pitch strength, and speed, constructs the input classification model of an emotion feature vector, and establishes a two-layer emotion extraction algorithm. The improved neural network algorithm can make the model accurately identify the emotion of music.

3.1. Feature Extraction of Music

3.1.1. Data Processing. In order to make the model parameters of training more realistic, select the music data that ranks top on the Internet. Songs were collected according to the emotional module of the song list. The original data classification had a high repetition rate, mainly including nostalgia, freshness, romance, sexy, sadness, cure, relaxation, loneliness, touching, excitement, happiness, quiet, and missing. However, some of the emotional labels overlap and contain, so the emotional labels are readjusted in the follow-up analysis. Change the options to excited, happy, relax, sad, and fear. The common audio data is 3-5 minutes, and its sampling rate is 16 kHz. For the two-channel data, the average is converted to single-channel processing and normalized. Usually, a song has a high chorus, a high frequency, and a gentle beginning and end. The shape of the sound channel is shown in the envelope of the short-time power spectrum of speech, while MFCC (Mel Frequency Cepstral Coefficients) is a feature that accurately describes this envelope [33]. In order to get more information of the audio, MFCC operation is now needed for the audio.

The Internet annotation data is subjective. In order to verify the accuracy of the annotation and the rationality of the emotional label of the data set, the extracted audio feature embedding dimension is visualized, and then, the data set distribution is adjusted to achieve small differences within the group and large differences between the groups. Common manifold learning methods can be divided into linear and nonlinear aspects. Principal component analysis (PCA) is a common linear manifold learning method. Isomap, LE (Laplacian Eigenmaps), and LLE (local-linear embedding) are all nonlinear manifold learning methods. Choose T-SNE when the equipment is good because it requires computationally efficient hardware. The uncertain high-dimensional data is projected to 2 or 3 dimensions by dimensionality reduction to achieve data visualization, and the specific image is observed to see whether the interval between classes is small but the interval between different classes is large. Compared with the original data, some information is lost after dimensionality reduction. If the data is divisible in low-dimensional space, the data set is divisible. If it is not divisible in a low-dimensional space, the data itself may not be divisible, or the data may not be projected into a low-dimensional space. T-SNE evolved from the SNE algorithm, which maps data points to probability distributions through affinity transformation. SNE constructs a

probability distribution among high-dimensional objects, so that similar objects have a higher probability of being selected and dissimilar objects have a lower probability of being selected. SNE constructs the probability distribution of these points in low-dimensional space, so the two probability distributions are as similar as possible. In order to increase the accuracy of the model, a series of data processing methods such as random shearing, translation, gain, equalization, compression, and noise increase were used to obtain the data set after data enhancement. And then, it was compared with the original data without data enhancement in the final model [34]. Through the use of data enhancement, weaken the emotional characteristics of music and increase the difficulty of learning. However, if the information can be learned even in this case, it can be said that it is quite reliable. But doing so may result in different emotional music deviation, because the music emotion has a lot to do with rhythm and tone. Data enhancement may lead to more confusion in classification results.

3.1.2. Primary Feature Extraction from a Spectrogram. Meier frequency is proposed based on the auditory characteristics of the human ear, and it has a nonlinear corresponding relationship with Hz frequency. Meier frequency cepstrum coefficient is the Hz spectrum characteristic calculated by using this relation between them. There are mainly the following steps: preweighting, frame segmentation, windowing, fast Fourier transform (FFT), Meyer filter banks, and discrete cosine transform (DCT). Preweighting is used to enhance high-frequency information, because high-frequency energy is generally lower than low frequency, avoiding numerical problems in FFT operations and possibly increasing the signal to noise ratio. We take the Fourier transform of the speech data and transform the information from the time domain to the frequency domain. But if you do FFT for the whole speech, you lose timing information. But if FFT is done for the whole speech, timing information is lost. Therefore, to retain information better, it is assumed that the frequency information in a short time t is unchanged, and the Fourier transform of the frame of length T is performed, so that the appropriate expression of the frequency domain and time domain information of the speech data can be obtained. After the signal is divided into frames, each frame is substituted into the window function, and the value outside the window is set to 0. This is done to eliminate spectral leakage that may occur at the both ends of each frame. Commonly used window functions include square window, Hamming window, and Hanning window. According to the frequency domain characteristics of window functions, the Hamming window is often used. Fast Fourier transform transforms the signal from the time domain to the frequency domain, removes the influence of points higher than the highest frequency of the sampled signal, and also reduces the dimension. Since the human ear has different sensitivities to different frequencies and it is nonlinear, the spectrum is divided into multiple Mel filter banks according to the sensitivity of the human ear. Within the Mel scale range, the center frequencies of each filter are linearly distributed with equal intervals but not equally spaced within the frequency range. The specific process is as follows.

(1) Assume that the i speech emotion signal is represented by s_i, then $s_i(m, n)$ is represented after windowing and framing, m is the number of frames, n is the frame length, and the Hamming window is used for windowing

(2) Window and perform the Fourier transform. Compute the Fourier coefficient $x_i(m, n)$, as shown in

$$x_i(m, n) = \sum_{m=0}^{N-1} s_i(m, n) e^{2\pi j/N} km, \tag{1}$$

where $k \in [0, N]$, N stands for the sequence length, and K stands for the ordinal number.

(3) The logarithmic energy method is used to generate the gray spectrum, and the gray values at points (a, b) are as follows:

$$g_i(a, b) = \log_{10} |x_i(m, n)|. \tag{2}$$

(4) The maximum and minimum normalization method is used to normalize the spectrum, and the normalized grayscale spectrum is obtained:

$$G_i(a, b) = \frac{g_i(a, b) - g_{\min}(a, b)}{g_{\max}(a, b) - g_{\min}(a, b)}, \tag{3}$$

where $g_{\max}(a, b)$ and $g_{\min}(a, b)$ are the maximum value and the minimum value in the degree level of spectrogram $G_i(a, b)$, respectively.

(5) The spectrogram was quantized into a grayscale image $G_i'(a, b)$ of 0-255

4. Feature Extraction of Deep Musical Emotion

Assume that the $M * N$ matrix represents the number of pixels in the spectrogram, and $G_i'(a, b)$ serves as the network input in the later stage. Fine-grained audio features are extracted as follows.

(1) In order to deeply extract musical emotional features, a block of size $k_1 * k_2$ is selected and features are extracted by sliding on the gray spectrum image, and each block becomes a $k_1 k_2$ dimensional vector. For the ith gray spectrum image $G_i'(a, b)$, all blocks are $X_i = [x_{i1}, x_{i2}, \cdots, x_{imn}] \in R^{K_1 K_2}$, where x_{ij} is the j th block of the ith spectrogram. Average $\bar{X}_i = [\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{imn}] \in R^{K_1 K_2}$ for each piece, and use the same processing method for all spectrograms to get the combination matrix $X = [X_1, X_2, \cdots, X_N] \in R^{K_1 K_2 N m n}$

(2) The filtering parameter of PCA processing in the first stage is L_1, which is used to represent the first L_1 eigenvalues after sorting the eigenvalues in PCA. Then, the corresponding eigenvectors of L_1 maximum eigenvalues are taken to form L_1 PCA convolution kernels. PCA was used to calculate sample set X:

$$\min_{V \in R^{k_1 k_2 l_1}} \left\| X - VV^T X \right\|_F^2, \tag{4}$$

$$V^T V = I_{L_1}, \tag{5}$$

$$W_l^1 = mat_{k_1 k_2}\left(q_l\left(XX^T\right)\right) \in R^{k_1 k_2}, \tag{6}$$

where I_{L_1} is the identity matrix of $L_1 L_2$, $l \in [1, 2, \cdots, L_1]$ and $\mathrm{mat}_{k_1 k_2}$ are functions that map vector cluster $R^{k_1 k_2}$ space, and $q_l(XX^T)$ represents the eigenvectors of the lth principal component of XX^T. The sample is transformed into a new space through training as follows.

$$F_i^l = \bar{X}_i \times W_l^1, \tag{7}$$

$$F = \left[F^1, F^2, \cdots, F^{L_1}\right] \in R^{k_1 k_2 \times L_1 N m n}, \tag{8}$$

where $i \in [1, N]$ and F_i^l represents the lth feature extracted after PCA calculation. F is the result of L_1 features.

(3) Repeat step 2 to perform similar operations on the convolutional image. The eigenvectors corresponding to L_2 maximum eigenvalues are taken to obtain L_2 PCA convolution kernels. The obtained feature transformation is as follows:

$$I_{\mathrm{i}}^1 = \left\{ \bar{F}_i^l \times W_w^2 \right\}_{w=1}^{L_2}. \tag{9}$$

(4) In order to enhance the expression ability of musical features, the Heaviside step function $O(x)$ is used to double the eigenvalues; the final results obtained after binarization and weighting are as follows:

$$S_i^l = \sum_{w=1}^{L_2} 2^{w-1} o\left(I_i^l\right), \tag{10}$$

where $i \in [1, N]$, the output S is divided into several blocks, and each block is analyzed by histogram statistics to obtain the final feature representation.

4.1. Music Emotion Recognition Based on CRNN. Traditional speech emotion recognition algorithms use LLDs or HSFs to extract features and then use statistical classification models such as HMM to classify emotions, but the performance of these algorithms is not particularly ideal. With the continuous development of deep learning, people use the deep neural network for speech emotion recognition, and many speech emotion recognition algorithms based on the deep neural network have been proposed. Based on the above research, we established a music emotion recognition algorithm based on the deep convolutional neural network [35].

(1) Deep convolutional neural network

There are generally several convolution layers in CNN, among which each convolution layer has many convolution kernels. The BP algorithm is usually used to optimize the parameters of the convolution kernels. By using a multilayer convolutional layer, the network can extract features of different levels from the input. The first few layers are used to extract low-level features, and the later layers can be used to extract features based on this, so as to extract higher-level features. In this paper, the improved deep convolutional neural network model used to analyze musical emotion is shown in Figure 1.

The main difference between CNN and DNN is that CNN uses convolution operation to extract features. For DNN, a fully connected design is used between the input layer and the hidden layer, with each neuron connected to all the neurons in the next layer. This is fine if the input image resolution is low. However, when the input is large, the computation becomes very heavy when using the fully connected operation. With the increase in network depth, the number of parameters will be higher and higher, and the training speed of the model will also be greatly affected. Different from the fully connected network, the connection between the hidden layer and the input layer is limited in CNN. So each neuron can only connect with a part of the neurons in the adjacent layer, and then, the number of parameters required for training in the network can be reduced. The area that each neuron connects to the next layer is usually called the neuron's receptive field.

When the size of convolution kernel is $3 * 3$, the size of receptive field of each neuron is also $3 * 3$. After the convolution of two layers with convolution kernel of $3 * 3$, the size of each neuron's receptive field is $5 * 5$. Input images of CNN usually contain several feature maps composed of neurons arranged in different rectangles. For different feature graphs, the model will use multiple different convolution kernels for feature extraction. For the same feature graph, the parameters of convolution kernel are shared between neurons. In the training process, the network constantly updates the parameters of the convolution kernel through the BP algorithm and finally obtains the optimal parameters. Parameter sharing operation directly reduces the number of network parameters and avoids the phenomenon of overfitting model.

(2) CRNN music emotion analysis model

RNN is a network used for processing time series data, which can extract time context information in features and has been widely used in the fields of speech recognition, text analysis, and speech emotion recognition. Long-short-term memory (LSTM) is a special form of RNN, which evolved from

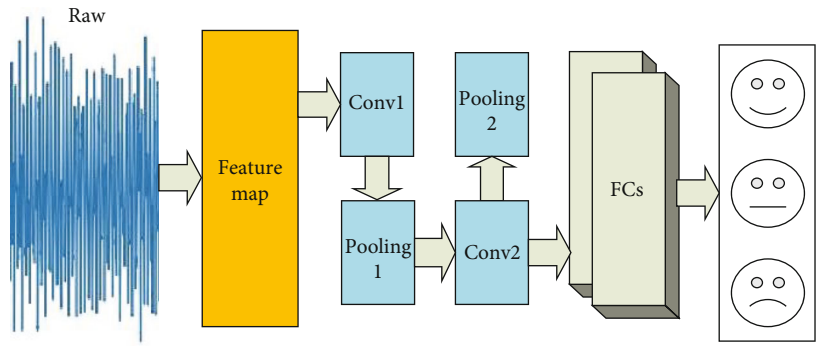

Figure 1: Deep convolution music emotion analysis model.

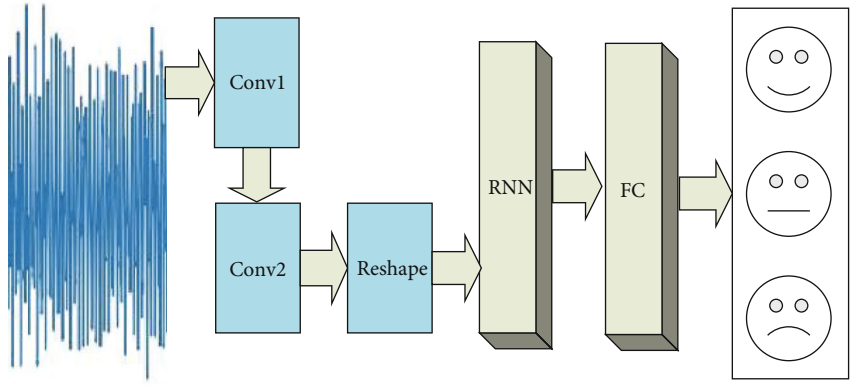

Figure 2: CRNN hybrid improved music emotion analysis model.

RNN and performs better than the standard RNN in completing many tasks. Almost all breakthrough results of RNN are achieved through LSTM [17–19]. Linking the previous information to the task at the current point in time is one of the key points of RNN, such as inferring subsequent speech from previous speech information at the time node. In some scenarios, such as using a language model to predict an upcoming word based on previous words, the prediction results can be obtained directly without context. The current task only needs to be performed according to the current information. In this case, the distance between the relevant information and the words to be predicted is very close, and the previous information contained is not large. In such application scenarios, RNN can learn to use the previous information, and the prediction results are often accurate. Because LSTM is designed to deliberately avoid long-term dependencies, remember that long-term information is the default behavior of LSTM in practice and does not need to pay a large price to obtain it. Therefore, in many practical issues, it has achieved better results. The hybrid improved model is shown in Figure 2.

The traditional speech emotion recognition algorithms are based on the deep neural network, and the input of the model is fixed length features or spectrograms. The algorithm divides the speech over the fixed value into equal length speech segments and makes up the insufficient speech segments with 0 to the fixed value and then extracts the speech spectrum from the speech segments. The label of each speech segment is the emotional label of the original speech. Although this ensures that the length of the input is fixed, thus reducing the difficulty of constructing a neural network, it causes other problems. It does not make sense to assign emotional labels to the entire speech for each subparagraph. For example, not the whole speech of a happy voice contains happy information, only a part of the speech segment contains happy emotional information, while other speech segments may only contain neutral emotional information. So models trained in this way may be less effective at identifying neutral and nonneutral emotions. The improved speech emotion recognition algorithm based on the deep neural network in this paper can input speech of different lengths and retain more complete emotional information, which is more suitable for music emotion analysis.

5. Experimental Analyses

5.1. Baseline Improvement Experiment Comparison. In order to test the experimental model more accurately, first classify the features of embedding by using the classical SVM method and then train with the method of OVOSVMs. Train 10 classifiers for multiple categories and then classify. The baseline is selected as a Gaussian kernel. At first, the penalty coefficient was set to 1, the kernel function was RBF, and the kernel coefficient was 0.02. Too large penalty coefficient may lead to overfitting and reduce the

Table 1: SVM multiclassification confusion matrix.

Category	0	1	2	3	4	Accuracy (%)
0	117	18	1	8	8	77.0
1	0	119	9	11	6	82.1
2	0	4	116	9	6	85.09
3	0	0	8	99	1	91.7
4	3	11	7	1	188	89.5
Accuracy (%)				85.2		

Category: 0—excited, 1—happy, 2—relax, 3—sad, and 4—fear and nervous.

Table 2: CNN multiclassification confusion matrix.

Category	0	1	2	3	4	Accuracy (%)
0	119	17	1	7	8	78.3
1	5	130	3	3	4	89.7
2	0	5	126	2	2	93.3
3	3	2	4	98	1	90.7
4	6	4	9	1	190	90.5
Accuracy (%)				88.3		

Table 3: Bi-LSTM multiclassification confusion matrix.

Category	0	1	2	3	4	Accuracy (%)
0	123	17	0	6	6	80.9
1	2	130	5	4	4	89.7
2	0	3	127	2	3	94.1
3	3	0	1	104	0	96.3
4	6	7	8	0	189	90.0
Accuracy (%)				89.7		

Table 4: Deep feature extraction confusion matrix based on CNN.

Category	0	1	2	3	4	Accuracy (%)
0	121	16	0	7	8	79.6
1	3	134	2	3	3	92.4
2	0	3	130	0	2	96.3
3	0	7	3	96	2	88.9
4	5	8	10	0	187	89
Accuracy (%)				88.3		

generalization ability of the model. According to the results, it was found that the data set was linear and indivisible, so the RBF Gaussian kernel was chosen for kernel function selection. The larger the gamma function is, the larger the support vector is. The smaller the gamma function is, the smaller support vector is. The default is the reciprocal of the number of categories. Because of the multiclassification and nonseparable linearity in this paper, through several experiments, the model accuracy was the highest when the gamma function was set as 0.02. One-versus-one was used to implement the multiclassification.

Table 5: Deep feature extraction confusion matrix based on CRNN.

Category	0	1	2	3	4	Accuracy (%)
0	122	15	0	7	8	80.2
1	3	134	2	3	3	92.4
2	0	2	131	0	2	97.0
3	0	6	3	97	2	89.8
4	5	7	9	0	189	90.0
Accuracy (%)				89.7		

According to the baseline results in Table 1, the total accuracy of SVM reached 85.2%. In terms of individual categories, the classification effect of the fourth category of sadness was the best, with an accuracy of 91.7%. The accuracy of the fifth category of fear and nervousness reached 89.5%, second only to that in the fourth category. Relatively speaking, the classification effect of the first category was not encouraging—only 77.0%, lower than the overall accuracy rate of 8.5 percentage points, but it was still appreciable on the whole.

Table 2 is a CNN multiclassification confusion matrix. The accuracy of the improved model based on CNN is 88.3%, which is higher than that of traditional machine learning. However, the model's judgment for the first category of excited is not accurate, and the judgment of the first two categories is prone to confusion. The model's judgment for the third category of relax is better than that of the first category, reaching 93.3%. And the accuracy of the fourth category and fifth category both reached 90%. The main reason is that CNN discovers the potential correlation among musical features through convolution operation.

Bidirectional LSTM can find the context of music features, so it has the best classification effect among these models. It is obvious from Table 3 that the overall accuracy is improved to 89.7%. Compared with the unidirectional LSTM, bidirectional LSTM improved the recognition accuracy of each category. Bi-LSTM improved the accuracy of the first category and the fourth category by 80.9% and 96.3%, respectively.

5.2. Experimental Comparison of Deep Feature Extraction. The focus of this paper is deep music feature extraction. Therefore then, we discuss the combination of deep feature extraction and the improved model. After deep feature extraction, the accuracy of the CNN model is improved compared with the previous accuracy. The specific experimental results are shown in Table 4. The accuracy of the second and third categories is improved after data enhancement while that of the first category is not, and the judgment of the latter two categories is not as good as that without data enhancement before. After CRNN used deep feature extraction, all kinds of emotion categories are improved; detailed data are shown in Table 5. It is a pity that the excitement category has made only marginal progress, while the most effective one is still the category three.

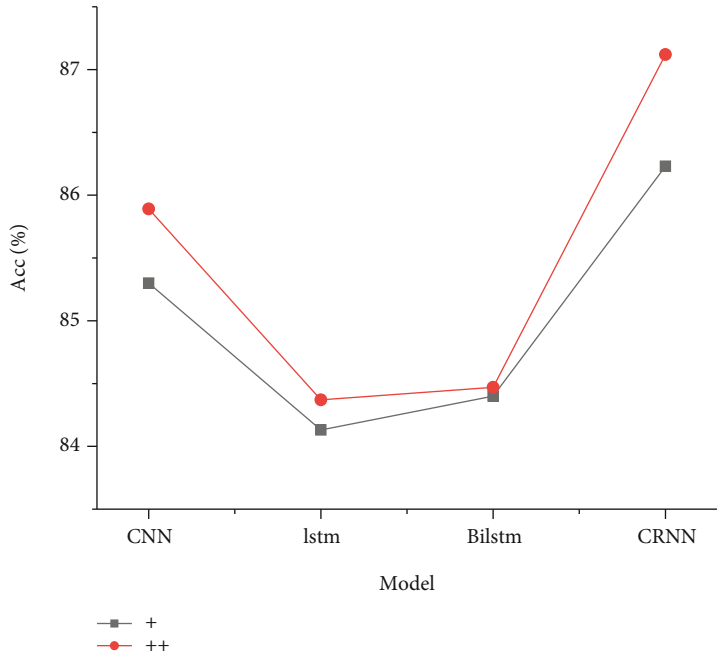

Figure 3: Complete music emotion analysis based on deep feature extraction. +: primary feature extraction, ++: deep feature extraction.

5.3. Decision Fusion Emotion Analysis based on Complete Music. Finally, we randomly tested the model effect of the whole music, as shown in Figure 3. According to probability output by softmax, multiple segments were weighted for voting prediction, and the final normalized result was the predicted probability of various emotional categories. As the experimental results show, after deep feature extraction, the experimental results of all models have been improved. Due to bi-LSTM's bidirectional time transfer characteristics, its improvement is not so obvious, but it still proves the effectiveness of our model. In terms of musical attributes, the model in this paper deeply extracts emotional features, so it can be well matched with the model based on CNN, and the experimental results are also in accord with the actual situation.

6. Conclusions

This paper discusses the speech emotion recognition algorithm, focusing on the speech emotion recognition algorithm based on the deep neural network. Also, the model structure and key techniques used in the algorithm are introduced. Aiming at the problem existing in the music emotion recognition algorithm such as low utilization of artificially designed features and oversimple feature extraction method, this paper puts forward the primary feature extraction algorithm based on language spectra and deep feature extraction algorithm from two aspects of feature extraction algorithm and model structure improvement. And the experiment has proven the superiority of the proposed algorithm. In the future, we hope to establish a more fine-grained musical emotion recognition model from the perspective of musical melody.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. I. Li-Ping, "On some issues in the development of information technology," *Journal of Hefei University of Technolog(Social Sciences)*, vol. 1, 2004.

[2] L. Yan and W. Xuemei, "International progress of information technology ethics research," *World Sci-tech Research and Development*, vol. 43, no. 4, p. 11, 2021.

[3] L. Yuxiang, *Research on Content-Based Music Analysis*, vol. 17, Tsinghua University, 2012.

[4] H. Xin, *Research on Content-Based Audio Information Classification Retrieval Technology*, Nanjing University of Science and Technology, 2007.

[5] Y. H. Yang and H. H. Chen, *Music Emotion Recognition*, CRC Press, Inc, 2011.

[6] A. Pras, M. Ghamsari, and M. Wanderley, "Combining musical tasks and improvisation in evaluating novel digital musical instruments," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, Marseille, France, October 2013.

[7] B. Hu, "Research on music emotion analysis based on feature vector," *Xidian University*, 2014.

[8] Z. Yongkai, *Research and implementation of multi-modal music emotion recognition technology*, China Engineering Science And Technology Knowledge Center, 2011.

[9] D. Vstfjll, "Emotion induction through music: a review of the musical mood induction procedure," *Musicae Scientiae*, vol. 5, no. 1, pp. 173–211, 2001.

[10] W. Shi and F. Shuang, "Research on music emotion classification based on lyrics and audio," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, October 2018.

[11] F. Ling, "The influence of digital media technology on contemporary music creation," *Performing Arts Technology*, vol. 2, p. 4, 2011.

[12] Z. Wenfei, "Virtual timbre and digital audio technology of computer music. Prose hundred schools traditional," *Chinese Education*, vol. 1, 2019.

[13] G. Xia, J. Tay, R. Dannenberg, and M. Veloso, "Autonomous robot dancing driven by beats and emotions of music," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, China, 2012.

[14] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," in *2010 Ninth International Conference on Machine Learning and Applications*, Washington, DC, USA, December 2010.

[15] M. Sordo, O. Celma, and D. Bogdanov, "MIREX 2011: audio tag classification using weighted-vote nearest neighbor classification," *Music Information Retrieval Evaluation Exchange*, vol. 2, 2011.

[16] R. Bresin and A. Friberg, "Emotion rendering in music: range and characteristic values of seven musical variables," *Cortex*, vol. 47, no. 9, pp. 1068–1081, 2011.

[17] F. F. Kuo, M. F. Chiang, M. K. Shan, and S. Y. Lee, "Emotion-based music recommendation by association discovery from film music," in *Proceedings of the 13th annual ACM international conference on Multimedia*, USA, November 2005.

[18] Y. H. Yang, C. C. Liu, and H. H. Chen, "Music emotion classification: a fuzzy approach," in *Proceedings of the 14th ACM international conference on Multimedia*, pp. 81–84, October 2006.

[19] R. Sarkar and S. K. Saha, "Music genre classification using EMD and pitch based feature," in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, Kolkata, India, January 2015.

[20] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. H. Chen, "Toward multi-modal music emotion classification," in *Pacific-Rim Conference on Multimedia*, Advances in Multimedia Information Processing - PCM 2008, pp. 70–79, Springer, Berlin, Heidelberg, 2008.

[21] Y. H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transaction Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.

[22] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query by semantic description using the CAL500 data set," in *The ACM International Conference on Special Interest Group on Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 439–446, USA, July 2007.

[23] D. Tingle, Y. E. Kim, and D. Turnbull, "Exploring automatic music annotation with acoustically objective tags," in *Proceedings of the international conference on Multimedia information retrieval*, pp. 55–62, USA, March 2010.

[24] Y. C. Lin, Y. H. Yang, and H. Homer, "Protecting the content integrity of digital imagery with fidelity preservation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 7, no. 3, pp. 1–20, 2011.

[25] K. J. Zhang and S. Q. Sun, "Web music emotion recognition based on higher effective gene expression programming," *Neuro Computing*, vol. 105, pp. 100–106, 2013.

[26] Y. Chengyi, *Research on multi-label music emotion recognition based on evidence theory and semantic cell model*, Zhejiang University, 2012.

[27] W. Zhiqiang, X. Zidong, and F. Xianghua, "Song emotion recognition algorithm based on spiral model," in *Harmonious Human-Machine Environment Joint Academic Conference*, Tsinghua University Press, 2011.

[28] S. Jiayin, *Research on key technology of automatic analysis of music elements*, Harbin Institute of Technology, 2011.

[29] L.-r. Zhuang and D.-y. Ye, "Text sentiment classification based on CSLSTM network," *Application of Computer Systems*, vol. 27, no. 2, p. 6, 2018.

[30] Y. H. Yang, D. Bogdanov, P. Herrera, and M. Sordo, "Music retagging using label propagation and robust principal component analysis," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 869–876, China, April 2012.

[31] W. Tianjiang, F. Liu, and C. Gang, "A Digital Music Emotion Recognition Method: CN, CN101599271 A," 2009.

[32] C. Zhixian, *Research on music emotion recognition method based on machine learning*, Hunan University, 2018.

[33] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques," *Tactics, Techniques, and Procedures*, vol. 2, 2010.

[34] L. Li, B. Lei, and C. Mao, "Digital twin in smart manufacturing," *Journal of Industrial Information Integration.*, vol. 26, no. 9, p. 100289, 2022.

[35] C. Yang and Q. Li, "Music emotion feature recognition based on Internet of things and computer-aided technology," *Computer-Aided Design & Applications*, vol. 19, no. S6, pp. 80–90, 2021.