Hindawi
Journal of Mathematics
Volume 2023, Article ID 9830845, 1 page
https://doi.org/10.1155/2023/9830845

Retraction

Retracted: Music Emotion Research Based on Reinforcement Learning and Multimodal Information

Journal of Mathematics

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Hu, "Music Emotion Research Based on Reinforcement Learning and Multimodal Information," *Journal of Mathematics*, vol. 2022, Article ID 2446399, 9 pages, 2022.

Hindawi
Journal of Mathematics
Volume 2022, Article ID 2446399, 9 pages
https://doi.org/10.1155/2022/2446399

Research Article

Music Emotion Research Based on Reinforcement Learning and Multimodal Information

Yue Hu [1,2]

[1]*Shanxi Jinzhong Institute of Technology, Taiyuan 030600, China*
[2]*UCSI University, Faculty of Social Sciences and Liberal Arts, Kuala Lumpur, Malaysia*

Correspondence should be addressed to Yue Hu; 1002163449@ucsiuniversity.edu.my

Received 14 December 2021; Revised 14 January 2022; Accepted 18 January 2022; Published 9 February 2022

Academic Editor: Naeem Jan

Music is an important carrier of emotion and an indispensable factor in people's daily life. With the rapid growth of digital music, people's demand for music emotion analysis and retrieval is also increasing. With the rapid development of Internet technology, digital music has been derived continuously, and automatic recognition of music emotion has become the main research focus. For music, emotion is the most essential feature and the deepest inner feeling. Under the ubiquitous information environment, revealing the deep semantic information of multimodal information resources and providing users with integrated information services has important research and application value. In this paper, a multimodal fusion algorithm for music emotion analysis is proposed, and a dynamic model based on reinforcement learning is constructed to improve the analysis accuracy. The model dynamically adjusts the emotional analysis results by learning the user's behavior, so as to realize the personalized customization of the user's emotional preference.

1. Introduction

With the rapid growth of the number of digital music, the traditional music analysis and retrieval methods are more and more difficult to meet people's needs. In the ubiquitous information environment, anyone can connect with the network and obtain personalized information services through appropriate terminal equipment anytime and anywhere [1]. This new environmental change is bound to be accompanied by the changes of information generation mode, information dissemination channels, and information utilization mechanism and also objectively promote the vertical deepening, personalized, and diversified development of users' information needs [2]. For information service organizations, this situation is not only an opportunity but also a challenge. There are massive multimedia data on the vast Internet. How to effectively store, organize, and retrieve such a large amount of information has become an urgent problem to be solved [3]. The purpose of affective computing is to give computers the ability to recognize, understand, and express various emotions similar to humans so that computers can interact with humans more naturally and harmoniously [4]. Music is an emotional medium that conveys the true feelings of human beings. Therefore, music has specific emotional labels, and explicit emotional labels are conducive to the audience to quickly select the songs they want to listen to at the appropriate time and place [5]. In multimodal fusion, decision level fusion is the highest level fusion. The existing decision level fusion first models the multimodal data of decision, and then obtains the linear weighting of decision, so as to generate the decision results [6].

Music is a symbol used by performers to express their thoughts and convey their emotions. It contains rich emotional information. Therefore, emotion-based music retrieval is also one of the key research contents of music information retrieval systems [7]. The characteristics of cross-platform, multisource, heterogeneous, and high-dimensional information and the development trend of dynamic and active services make people pay more attention to the multimodal information fusion theory and technical methods initially applied in the military field and begin to

actively explore possible solutions to extend it to the field of information services [8]. With the development of digital storage technology and mobile Internet, digital music also has a serious problem of information overload [9]. Therefore, the automatic analysis of music emotion has also become one of the hotspots of today's research and has broad application prospects in music retrieval and recommendation [10]. With the gradual expansion of the music scale, the era of digital music has ushered in. At this time, the scientific management of music has attracted much attention [11]. Different from traditional manual retrieval methods, automatic retrieval will save a lot of labor costs. At the same time, compared with manual analysis, improving the accuracy of analysis will be a difficult problem for automatic analysis [12]. As an important means of automatic music retrieval, classifying music according to the expressed emotion is attracting the attention of researchers from different fields.

For music, emotion is the most essential feature and the deepest inner feeling. Music emotion recognition based on computer automation plays a key role in promoting the development of artificial intelligence [13]. For music emotion analysis, the most common method is to analyze the acoustic features extracted from music and get the emotional analysis results. However, the first mock exam is usually not satisfactory. The traditional single-mode research based on can only express some characteristics of the object, just as people observe the world through only one sense, which has considerable limitations [14]. In contrast, multimodal information has richer semantic information, and the information of each mode can complement each other. At the same time, the correlation between different modal data is also helpful to improve the accuracy of analysis results to a certain extent [15]. As an important branch of music labeling, emotion labels can reflect the artistic conception to be expressed by music to a certain extent. People can find music in line with Tun and emotion through emotion labels, which can relieve depression in their hearts and find happy resonance [16]. Compared with other music analysis problems such as genre analysis, emotion analysis is closer to people's perception, and the melody of music often contains the expression of emotion. In this paper, a music emotion analysis algorithm based on multimodal fusion based on reinforcement learning is proposed to improve the analysis accuracy.

In this article, the innovative concept of analysis of music characteristics is studied. Moreover, we discussed the analysis of music features based on lyrics and the music emotion feature analysis method based on multimodal fusion to improve the analysis accuracy. Music feature analysis is a very important step in the process of music emotion analysis and the multimodal music emotion analysis method is to analyze music emotion based on music content and lyrics, respectively, and then combine the two analysis results to get the final music emotion analysis. The relationship between music structured information and human emotion cannot be fully reflected by using existing common features. Therefore, we can further explore the feature extraction method with more musical emotion analysis ability.

The following is a summary of the research: Section 1 contains the introduction; Section 2 discusses the related work and background. Section 3 discusses the analysis of music characteristics. Section 4 discusses the music emotion feature analysis method is based on multimodal fusion; finally, the conclusion brings the paper to a finish in section 5.

2. Related Work

In the task of attribute-level emotion analysis, the literature [17] conducts joint learning through two tasks of attribute extraction and attribute-level emotion analysis, which greatly improves the performance of the attribute-level emotion analysis task. Literature [18] puts forward a joint model, which can simultaneously model the two tasks of emotion analysis and emotion cause recognition and effectively improve the recognition performance of the two tasks. Emotion analysis and emotion analysis are two different subtasks in emotion analysis. Because of the strong relationship between emotion tags and emotion tags, the two tasks are closely related. Literature [19] improves the performance of the two tasks by labeling an extra data set with emotional tags and emotional tags. However, it is difficult to obtain similar data sets in real scenes. Literature [20] adopts integer linear programming (ILP) to study emotion and emotion analysis tasks jointly and obtains the connection between the output of the emotion analyzer and the output of the emotion analyzer through constraints.

Literature [20] shows that music lyrics do contain some special semantic information, including emotion. Therefore, the comprehensive utilization of audio and lyrics modes can effectively improve the accuracy of music emotion analysis. We can analyze the relationship between lyrics, music modes, and human perception and explore the intrinsic relevance between the two modes and complement each other to improve the accuracy of the analysis. Literature [21] has proposed some simple multimodal fusion methods, which comprehensively use the information of lyrics and audio modes to analyze music. The experimental results prove that using multimodal information can improve the accuracy of emotion analysis to a certain extent compared with using only a single mode. Literature [22] preliminarily uses deep neural networks to extract advanced feature representation from original audio data and verifies the effectiveness of deep neural networks in speech emotion recognition. Literature [23] uses a convolution neural network to extract audio features to train audio data, and the accuracy of audio emotion analysis has been greatly improved. Based on reinforcement learning technology, this paper studies the emotional analysis of music from the perspective of audio visualization. According to the demand analysis of music emotion analysis, this paper explores a model framework of music emotion analysis based on multimodal information fusion function and level.

3. Analysis of Music Characteristics

The characteristics of music are sound (overtone, duration, amplitude, pitch, and timbre), melody, rhythm, structure or form, expression, and texture.

3.1. Music Feature Analysis Based on Audio. Music feature analysis is a very important step in the process of music emotion analysis. Different music features may show different emotions. Therefore, the main task of music feature analysis is to find an optimal feature space to represent music [24]. This feature space can not only reflect the emotion of music but also have a certain degree of discrimination, which can distinguish music with different emotions. The framework of multimodal music emotion analysis is shown in Figure 1.

Music is mainly composed of several basic elements, including sound nest, sound length, sound intensity, sound color, and so on [25]. Then, two or more basic elements are integrated to form the basic characteristics of music, mainly including ① rhythm: the rhythm of music reflects the speed and urgency of music tunes, in which the emotion expressed by gentle music is calm and gentle, while the emotion expressed by sudden rhythm music is strong. ② Melody: what people used to call melody actually refers to the melody. It is the most basic element of music. It is a series of organized and rhythmic sequences composed of several musical sounds by artists according to a certain pitch, time value, and volume. Melody can reflect the emotion expressed in music. For example, the emotion expressed by music with a light melody should also be light. ③ Strength: strength can also express the emotion of music. For the same music, the emotion expressed by different degrees is different. Usually, the greater the intensity, the louder and more exciting the music. The smaller the degree, the more soothing and soft the music is. ④ Timbre: timbre refers to people's sensory characteristics of different sounds so that people can distinguish different sounds. Different people or musical instruments produce different timbres.

The choice of music emotion model is the basis of music emotion analysis. Music carries a variety of emotions. The music analyzed by the early music emotion research is mostly classical music. Among them, the vocal content is small, and the emotional characteristics are mostly reflected by the rhythm, melody, pitch, and timbre expressed by musical instruments, and a piece of classical music may contain several completely different emotions. The study of this music needs to intercept a music fragment for analysis. According to the basic and complex characteristics of music, the overall characteristics of music are identified, including music form structure, style, and emotional connotation. The specific structure is shown in Figure 2.

For the music emotion analysis task, the feature extraction method is an important component module, and a good feature extraction method has a great influence on the result of the analysis task. Feature extraction solves the problem of how to better represent the analysis sample set. Usually, samples are converted into feature vector

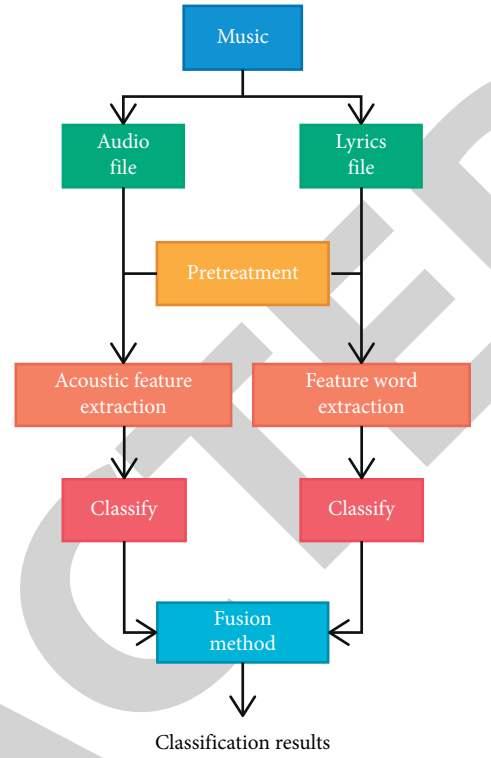

Figure 1: The framework of multimodal music sentiment analysis.

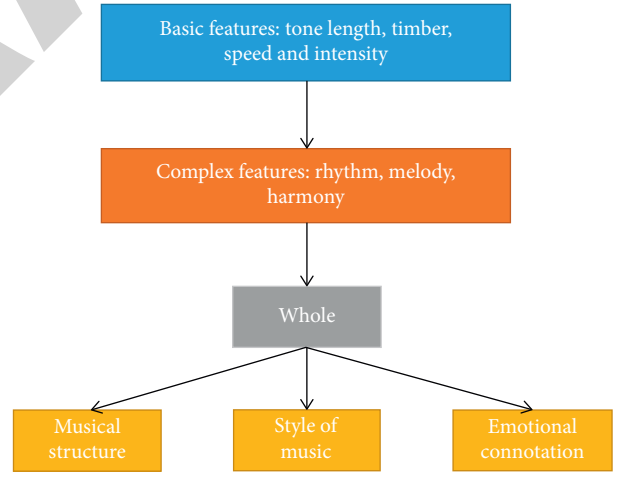

Figure 2: Composition of music form.

representation for the analysis model. Music feature acquisition is an important link in music emotion analysis. Early music acquisition mainly focused on audio attributes of sound killing. The basic audio features of a song, such as a rhythm, timbre, tone, volume, melody, and harmony, can reflect the emotional characteristics of music to varying degrees. Due to the structural heterogeneity between audio features and text features, there is an insurmountable gap between the emotions expressed by the two features. This makes it a serious problem to mine the correlation between the two expressions in emotional expression for multimodal analysis.

Time-domain characteristics of music refer to the time-domain parameters of each post calculated from music signals. Typical time-domain features include short-time energy, short-time average amplitude, short-time average zero-crossing rate, short-time autocorrelation function, and short-time average amplitude difference function.

The short-term energy of the nth frame music signal is defined as follows:

$$E_n = \sum_{m=n-(N-1)}^{n} [x(m)w(n-m)]^2. \tag{1}$$

In the formula, $w(n-m)$ is the moving wins function, N is the effective width of the window, and n is the time position of the window. It can be the starting point of the window or the midpoint or end of the window.

Short-term energy E_n is a time series, from which we can see how the signal energy changes with time. Generally speaking, the short-term energy of voiced sound is much larger than that of unvoiced sound, so it is easy to distinguish voiced sound from unvoiced sound by short-term energy sequence. In addition, the short-time energy sequence can also be used to determine the starting and ending points of music.

The process of extracting the characteristics of pitch and time value of music performance is shown in Figure 3.

Because the calculation of short-time energy needs a square operation, which enlarges the difference between magnitude and amplitude, it cannot accurately reflect the characteristics of signal short-time energy changing with time. Therefore, a short-term average amplitude describing the time-varying characteristics of signal energy is proposed, which is defined as follows:

$$M_n = \sum_{m=n-(N-1)}^{n} |x(m)w(n-m)| = \sum_{m=n-(N-1)}^{n} |x(m)|w(n-m). \tag{2}$$

The different signs of adjacent sampled values are called zero crossing, and the number of zero crossings per unit time is called the zero-crossing rate. The short-term average zero-crossing rate of a frame of music signal is defined as follows:

$$\begin{aligned} Z_n &= \frac{1}{2N} n \sum_{m=n-(N-1)}^{n} \left|\operatorname{sgn}[x(m)w_R(n-m)] - \operatorname{sgn}\{x(m-1)w_R[n-(m-1)]\}\right| \\ &= \frac{1}{2N} \sum_{m=n-(N-1)}^{n} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|. \end{aligned} \tag{3}$$

In the formula, $\operatorname{sgm}[x(m)]$ is the symbolic function of $x(m)$, defined as follows:

$$\operatorname{sgn}[x(m)] = \begin{cases} 1, & x(m) > 0, \\ 0, & x(m) = 0, \\ -1, & x(m) < 0. \end{cases} \tag{4}$$

In order to overcome the short-term average zero-crossing rate is very sensitive to noise, the formula (3) can be modified as follows:

$$Z_n = \frac{1}{1N} \sum_{m=n-(N-1)}^{n} \{|\operatorname{sgn}[x(m) - A] - \operatorname{sgn}[x(m-1) - A]| + |\operatorname{sgn}[x(m) + A] - \operatorname{sgn}[x(m-1) + A]|\}. \tag{5}$$

Whether or not it crosses zero is not judged by the different signs of the adjacent sampled value of the signal but judged by the different sign after the adjacent sampled value of the signal exceeds a set appropriate positive and negative limit. This eliminates false zero crossings caused by noise. Normally, the short-term average zero-crossing rate of unvoiced and noise is much larger than that of voiced sounds, so the short-term average zero-crossing rate can be used to distinguish them easily. The short-term autocorrelation function is defined as follows:

$$\begin{aligned} R_n(k) &= \sum_{m=n-(N-1)}^{n} [x(m)w(n-m)]\{x(m+k)w[n-(m+k)]\}, \\ n-m &= m \sum_{m=0}^{N-1-k} [x(m+n)w(m)][x(m+n+k)w(m+k)]. \end{aligned} \tag{6}$$

In the formula, k is the autocorrelation lag time. Equation (6) shows that $R_n(k)$ of each frame of the signal is a sequence with lag time k as an independent variable. The formula of the short-term average amplitude difference function is as follows:

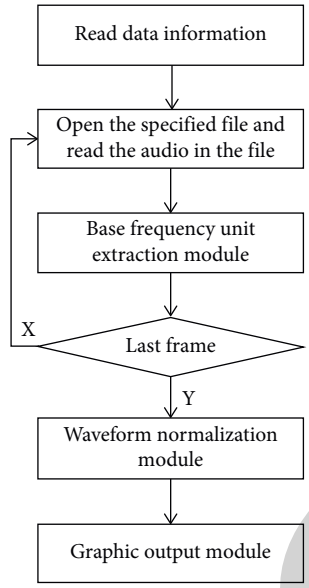

Figure 3: The process of extracting the pitch and time value feature of music performance.

$$\gamma_{k=}\sum_{m=0}^{N-1}\left|x(n+m)w_1(m)-x(n+m+k)w_2(m-k)\right|. \quad (7)$$

In the formula, $w_1(n)$ and $w_2(n)$ are rectangular windows with widths N and $N+K$, respectively, where K is the maximum possible hysteresis value. For any periodic signal, when the lag time is equal to the period or an integer multiple of the period, there will be a short-term average amplitude difference function $\gamma_k = 0$. The voiced signal is approximately a periodic signal, so γ_k will reach its minimum value at a lag time point equal to the pitch period or an integer multiple of the pitch period. Using this property, we can distinguish between voiced and unvoiced sounds based on the γ_k curve and estimate the pitch frequency of voiced sounds.

Pitch depends on the frequency and loudness of sound. Bass gives people a thick and deep feeling, while treble gives people a bright and sharp feeling. Audio features have strong objectivity and can be easily extracted from songs by digital signal processing. The key problem of audio feature extraction is which features are extracted. The results show that the energy, rhythm, melody, and timbre of music are the four characteristics that can best reflect music emotion. Therefore, in the digital signal processing of music, we should focus on these four characteristics. In the existing research, the music characteristics based on audio are usually borrowed from the parameters of speech signals, and the characteristics of speech signals change with time, but the changes are slow. Therefore, it is usually divided into short segments with phase dimensions, and each segment is processed separately by the processing method of a stationary random signal, which is the short-time processing technology of speech signal. The energy characteristics of music are closely related to the degree of motivation that music can bring to people. The higher the energy of music, the stronger the sensory stimulation to the listener. Songs such as metal and rock generally have higher energy value, while songs such as light music generally have lower energy value.

3.2. Analysis of Music Features Based on Lyrics. As an important part of music, lyrics also contain rich emotional information. Therefore, mining emotions from lyrics is a good supplement to music emotional analysis. The core problem of sentiment analysis based on lyrics is how to construct a feature space that can reflect lyrics sentiment, which mainly focuses on the selection of the expression model of lyrics text and the selection of feature selection methods. Lyrics data usually incorporate the expression of the music writer's own emotion, so it has rich semantic information related to emotion. How to extract this emotion from sparse and messy lyrics files will be a great challenge. A typical text emotion recognition system is shown in Figure 4.

Assuming that a document is composed of m feature words, the contribution of each feature word to the document is reflected by its weight. Expressed by a mathematical formula is as follows:

$$D = D(t_1, w_1; t_2, w_2, \ldots, t_m, w_m), \quad (8)$$

where w_i is the weight and $1 \le i \le m$. The similarity of the two documents is expressed by finding the cosine of the angle between the corresponding vectors. The formula is as follows:

$$\mathrm{sim}(D_1, D_2) = \cos\theta = \frac{\sum_{i=1}^{n} w_{1i} \times w_{2i}}{\sqrt{\left(\sum_{i=1}^{n} w_{1i}^2\right)\left(\sum_{i=1}^{n} w_{2i}^2\right)}}. \quad (9)$$

Among them, w_{1i} and w_{1i} represent the weight of the w_{2i} feature item of documents i^{th} and D_1, respectively.

Suppose there is $\mathrm{sim}(D_1, D_2)$, a document set with a total of n documents. After preprocessing, a total of m feature words are extracted, and a matrix of "feature words-documents" can be constructed:

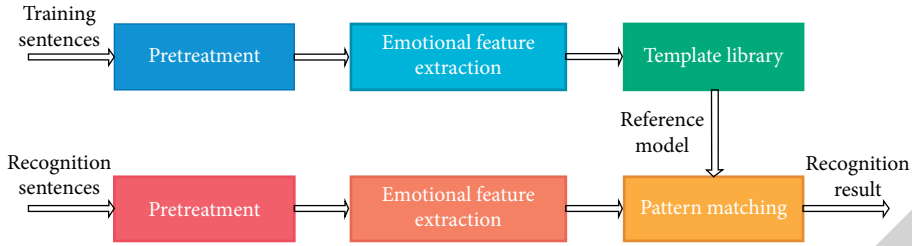

Figure 4: Principles of text emotion recognition.

$$X_{m\times n} = \left[x_{ij}\right]. \tag{10}$$

Among them, x_{ij} represents the weight of the i^{th} feature word in the j^{th} article search. The weight is used to measure the distinguishing ability of the feature word to the document, or the degree of contribution to the analysis. Since the length of each document is different, the weight tends to favor longer documents. In this case, it can be normalized when calculating the weight to avoid this situation. This leads to the following formula:

$$w\left(t_i, d_j\right) = \frac{\left(\log_2\left(1 + tf\left(t_i, d_j\right)\right)\right) \times \log_2\left(N/N_t\right)}{\sqrt{\sum_{t_i \in d_j}\left[\left(\log_2\left(1 + tf\left(t_i, d_j\right)\right)\right) \times \log_2\left(N/N_{t_i}\right)\right]}}. \tag{11}$$

Among them, $w(t_i, d_j)$ represents the weight of feature word t_i in document d_j. $tf(t_i, d_j)$ represents the number of times the feature word t_i appears in the document d_j, and $(1 + tf(t_i, d_j))$ is to prevent the occurrence of $tf(t_i, d_j)$. N represents the number of documents in the document set, and N_{t_i} represents the number of documents in the document set that contain characteristic words.

4. Music Emotion Feature Analysis Method Based on Multimodal Fusion

Multimodal information fusion is an information processing process that comprehensively utilizes natural language processing, semantic analysis, statistical analysis, and other technical methods to detect, correlate, estimate, combine, and analyze multimodal information in multiple levels and dimensions. The multimodal music emotion analysis method is to analyze music emotion based on music content and lyrics, respectively, and then combine the two analysis results to get the final music emotion analysis. If stress is defined as anxiety and happiness, and energy is defined as vitality and calmness, the final analysis result is determined by combining the two analysis results. To study the local form of the melody line, we should not only look at the connection of two notes but at least look at the ups and downs of four or five notes and five or six notes in a bar, so as to see the characteristics of linear form from the harmony interval and disharmony interval in music acoustics and law, for example, Table 1.

The energy of the audio signal changes significantly over time, and its short-term energy analysis gives an appropriate description method to reflect these amplitude changes. For the signal $\{x(n)\}$, the short-term energy is defined as follows:

$$E_n = \sum_{m=-\infty}^{\infty}\left[x(m)w(n-m)\right]^2 = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) = x^2(n) \times h(n). \tag{12}$$

Among them, $h(n) = w^2(n)$. Equation (12) represents the short-term energy when the window function is started at the nth point of the signal. The short-term energy can be regarded as the output of the square of the audio signal through a linear filter, and the unit impulse response of the linear filter is $h(n)$, as shown in Figure 5.

If $x_w(n)$ is used to represent the signal after $x(n)$ is windowed, the length of the window function is N, and the short-term energy is expressed as follows:

$$E_n = \sum_{m=n}^{n+N-1} x_w^2(m). \tag{13}$$

The requirement of music emotion analysis is based on the multimodal, complex, multisource, and heterogeneous characteristics of music emotion. The service has the quick adaptability of universal access, aggregation on demand, context processing, and seamless application and can realize the interoperability and autonomous cooperation of heterogeneous data across fields and platforms. Through the evaluation of the direction of notes, take the bar as the unit. No matter whether the notes go down or up, as long as a series of notes with the same direction appear continuously, an upward or downward melody line can be generated, which means that the evaluation value is higher.

The feature vector of lyrics is extracted based on the reinforcement learning model, and the feature value of each dimension is calculated. Then labeled lyrics are clustered to get a cluster set, and the similarity between lyrics and each cluster and the similarity of each cluster and the ratio of each category in the cluster are tested. The assignment of melody weights is shown in Table 2. The relationship between melody weight and melody trend is shown in Figure 6.

Table 1: Harmony degree of overtone.

Partial name	Chord vibration length	Degree of concord
1st part	1/10	Absolute concord
2nd part	1/20	Complete concord
3rd part	1/30	Absolute concord
4th part	1/40	Incomplete concord
5th part	1/50	Incomplete concord
6th part	1/60	Disharmony

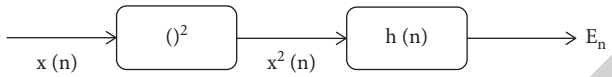

Figure 5: Short-term energy graph.

Table 2: Melody weight data.

Same trend degree of melody	Melody weight
2	0.9
4	0.75
6	0.60
8	0.45
10	0.30
Greater than 10	0.15

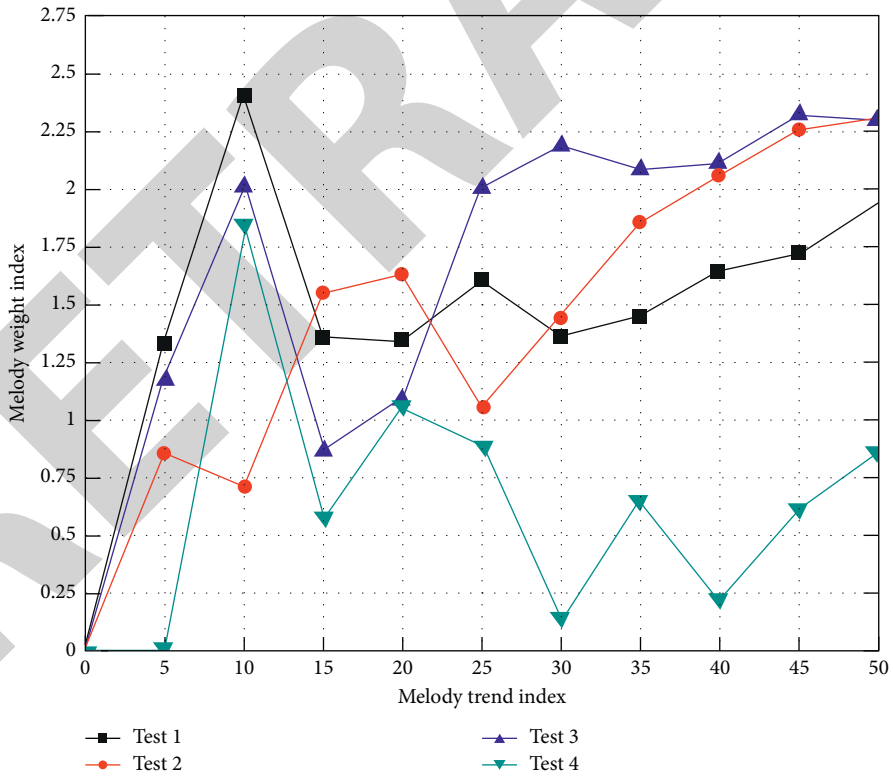

Figure 6: The relationship between the weight of the melody and the degree of the direction of the melody.

Compared with sentence-level emotion classification of automatic encoder, the accuracy of sentiment classification of article-level lyrics based on word vector sentence coding is improved. The dual-mode fusion method based on the neural network has a remarkable effect in audio emotion classification because it can set the weight of each mode. The linear regression curve is calculated according to the stepwise multiple linear regression equation, as shown in Figure 7. The ability of music emotion analysis based on the reinforcement learning

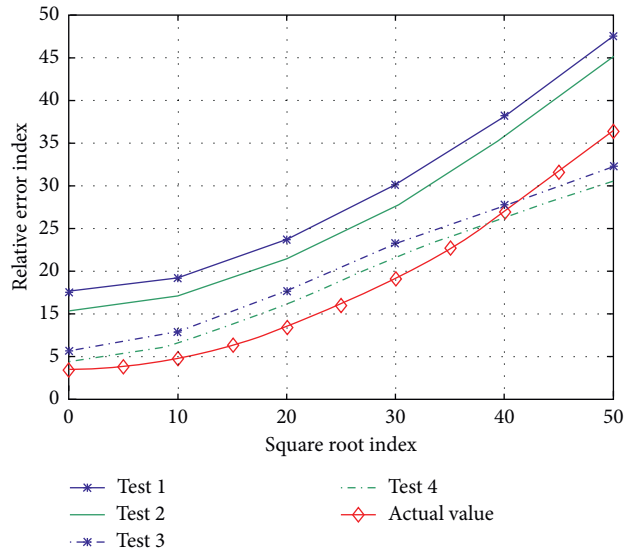

Figure 7: Analysis of the relationship between the error and the number of iterations.

feature extraction model is stronger than that of common multimodal information emotion analysis.

In the speech modality experiment, the method of capturing context information is better than other baseline methods, whether in the main task or the auxiliary task. The above baseline method only uses the phonetic features of the main task or auxiliary task. Due to the weak representation ability of speech modal features and the very small number of samples of "disgust" and "fear" categories, our model cannot predict the corresponding categories, and its performance in individual categories cannot reach the best. Different modes of music data often have a certain correlation in emotional expression; that is, different modes are not independent of each other. This correlation can often enhance the accuracy of sentiment analysis. Because the feature extraction methods used in different modal music data are different from each other, the dimensions and attributes of the obtained features are quite different. This difference makes it impossible for music features of different modes to operate and calculate each other directly, which makes it difficult to fully explore and apply the correlation between music data of different modes. In order to make full use of the temporal correlation of music data of different modes and improve the accuracy of emotion analysis, it is necessary to design an effective mechanism to aggregate music features of different modes and different time scales according to emotion categories.

5. Conclusions

In order to effectively manage music resources and help people efficiently obtain interesting content from massive music, music emotion analysis has always been a hot spot for scholars. Under multimodal fusion, based on the fusion of existing linear weighted decision-making layers, the reinforcement learning method is introduced, which can highly fuse different types of analysis effects of multiple modes, so as to guarantee the overall fusion effect. Based on reinforcement learning technology, this paper studies the emotional analysis of music from the perspective of audio visualization. According to the demand analysis of music emotion analysis, this paper explores a model framework of music emotion analysis based on multimodal information fusion function and level. The experimental results on multimodal emotion analysis data set show that this method can greatly improve the performance of emotion analysis tasks through emotion auxiliary information, and at the same time, the performance of the emotion analysis task is also improved to a certain extent. Music emotion analysis is an important means for automatic music retrieval. The heterogeneity and semantic gap between different modal music data make it a challenging problem to use multimodal information for music emotion analysis. The relationship between music structured information and human emotion cannot be fully reflected by using existing common features. Therefore, we can further explore the feature extraction method with more musical emotion analysis ability.

Data Availability

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

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] G. Wang, "Analysis of teaching strategies to strengthen students' emotional experience in middle school music teaching," *Teaching Management and Education Research*, vol. 2, no. 15, pp. 89-90, 2017.

[2] X. Li, L. Han, J. Li, and J. Zhou, "Multimodal music sentiment classification based on optimized residual network," *Computer and Modernization*, vol. 304, no. 12, pp. 87–93, 2020.

[3] K. Chen and L. Han, "Research on music emotion classification based on audio and lyrics," *Electronic Measurement Technology*, vol. 41, no. 22, pp. 15–20, 2018.

[4] Z. Tang, X. Liu, and H. Yang, "Image emotion annotation method based on multimodal information fusion," *Computer Integrated Manufacturing Systems*, vol. 26, no. 1, pp. 134–144, 2020.

[5] Z. Gong and X. Shao, "Multi-modal music recommendation system," *Journal of Nanjing University of Information Technology (Natural Science Edition)*, vol. 59, no. 1, pp. 72–80, 2019.

[6] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 38, pp. 1–19, 2021.

[7] H. Kaya and A. A. Salah, "Combining modality-specific extreme learning machines for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 139–149, 2016.

[8] X. Wang, "Exploration of music emotion teaching in junior middle schools," *China New Communications*, vol. 18, no. 15, p. 135, 2016.

[9] L. Bo, "A probe into the emotional experience education of students in the teaching of music appreciation," *Xue Weekly*, vol. 9, no. 9, pp. 176-177, 2017.
[10] B. Lv, Y. Zhang, and W. Liu, "Objective evaluation and regulation treatment of depression based on multimodal emotional brain-computer interface," *Chinese Journal of Psychiatry*, vol. 54, no. 4, pp. 243–251, 2021.
[11] D. Zheng, "Music emotion recognition and classification algorithm based on forward neural network," *Information & Technology*, vol. 337, no. 12, pp. 65–69, 2019.
[12] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 2, pp. 103–112, 2020.
[13] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, no. 10, pp. 217–230, 2017.
[14] S. Landry and M. Jeon, "Interactive sonification strategies for the motion and emotion of dance performances," *Journal on Multimodal User Interfaces*, vol. 14, no. 2, pp. 167–186, 2020.
[15] Y. Wang and W. Shi, "The application of multimodal English songs in children's English teaching," *Weishi: Modern Management*, vol. 131, no. 11, pp. 93–95, 2018.
[16] Y. Hao, Y. Zuo, and P. Wang, "Application of multimodal motion combined with music imaging in patients with head and neck tumor radiotherapy," *Journal of Changzhi Medical College*, vol. 154, no. 6, pp. 71–74, 2020.
[17] C. Jin, C. Hou, and Z. Cheng, "Research on intelligent composition based on multimodal neural network and rule algorithm," *Journal of Communication University of China (Natural Science Edition)*, vol. 26, no. 5, pp. 12–18, 2019.
[18] K. Sun, "Multimodal discourse analysis from the perspective of modern linguistics," *Journal of Jiangxi Vocational and Technical College of Electric Power*, vol. 109, no. 2, pp. 162-163, 2018.
[19] Y. Wang, "Research on the optimization of general education teaching based on multi-modal concepts," *Education Modernization*, vol. 5, no. 5, pp. 290–292, 2018.
[20] X. Li, G. Lu, and J. Yan, "A review of multi-modal dimension emotion prediction," *Acta Automatica Sinica*, vol. 44, no. 12, pp. 2142–2159, 2018.
[21] J. Jia, H. Jiang, and T. Zhang, "A review of multimodal emotion recognition," *Journal of Minzu University of China (Natural Science Edition)*, vol. 29, no. 1, pp. 54–58, 2020.
[22] T. Fan, P. Wu, and Q. Cao, "Research on multi-modal fusion netizen emotion recognition based on deep learning," *Journal of Information Recording Materials*, vol. 10, no. 1, pp. 39–48, 2020.
[23] J. He, Y. Liu, and Z. He, "Research progress in multimodal emotion recognition," *Application Research of Computers*, vol. 35, no. 11, pp. 3201–3205, 2018.
[24] S. Chen, S. Wang, and Q. Jin, "Multimodal emotion recognition in multicultural scenes," *Journal of Software*, vol. 29, no. 4, pp. 168–178, 2018.
[25] W. Tiger, "Discussion on the emotional expression of singing in vocal music teaching," *Northern Music*, vol. 37, no. 01, p. 137, 2017.