

Research on Emotion Recognition Algorithm of Piano Music Based on Multimodal Learning

Qiaolin Yu^{1, a}, Ying Lin^{2, b}

¹School of Liaoning Normal University, Dalian 116081, CN;

² Omnimedia Tech Center, Shenyang Radio and Television Station, Shenyang 110300, Liaoning, China

^a yuqiaolin@lnnu.edu.cn; ^b wolongfengchu2021@163.com

Abstract: The purpose of this study is to propose an emotion recognition algorithm for piano music based on multimodal learning. Multi-modal learning is a learning method that combines various types of data. By combining different modal information, learning tasks can be completed more comprehensively and accurately. In the field of music emotion recognition, multimodal learning can extract the emotional features of music from different angles, and improve the recognition accuracy and robustness. In this study, the multi-modal learning method is mainly used, and the representative features are extracted by combining the audio signal, score information and the performance of the piano music, and the machine learning algorithm is used to classify the emotions. The audio signal is preprocessed, including noise reduction, cutting and other operations, and then the features of the audio signal are obtained by feature extraction methods, such as short-time Fourier transform (STFT) and formant analysis (RPA). Digitally convert the score, extract the structure, melody, harmony and other characteristics of the music, and obtain the emotional information of the music by analyzing the relationship between symbols and notes in the score. By analyzing the performance skills, strength, speed and other parameters of the player, the characteristics related to the emotional expression of the player are extracted. Machine learning algorithm is used to train and classify the fusion features, and the emotion recognition of piano music is realized. In this study, the emotion recognition algorithm of piano music based on multimodal learning can mine the emotion information of music at different levels, improve the accuracy and robustness of emotion recognition, and provide useful reference for the field of music emotion calculation.

Keywords: multimodal learning; ; Research on Emotion Recognition Algorithm of Piano Music

1. 1 Introduction

Piano music has rich emotional expression ability, which can arouse people's strong emotional resonance. Therefore, it is of great practical value and theoretical significance to carry out emotion recognition research in the field of piano music. At present, the emotion recognition of piano music is mainly realized by analyzing audio signals. However, it is often difficult to fully capture the emotional information conveyed by piano music only by audio information. In contrast, visual information also plays an important role in identifying emotions. For example, the expression, posture and body movements of a pianist can convey emotional information. Therefore, we propose an emotion recognition algorithm for piano music based on multimodal learning, which aims to recognize the emotion of piano music more accurately by combining audio and visual information. Specifically, we use convolutional neural network (CNN) and recurrent neural network (RNN) to process audio signals, and use convolutional neural network and long-term and short-term memory network (LSTM) to process visual information, and fuse the two modes through attention mechanism. The main contributions of this paper are as follows: First, we propose an emotion recognition algorithm for piano music based on multimodal learning, which can understand the emotion conveyed by piano music more comprehensively and extract more abundant features from it. Secondly, we introduce attention mechanism to improve the model's attention to key features, so as to better understand the relationship between audio and visual information. Finally, through the experimental verification, we proved the effectiveness of the proposed method.

2. Music classification model design

2.1 Mel spectrum cepstrum coefficient

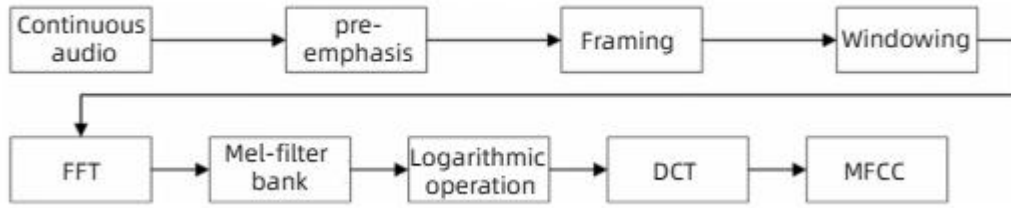


Fig. 1 Extraction process of MFCC

Mel-frequency cepstral coefficients (MFCC) is a commonly used feature extraction method in speech signal processing. By simulating the characteristics of human auditory system, it transforms the linear interval on the frequency axis into Mel Scale, and extracts a set of characteristic coefficients with good robustness and discrimination by cepstrum analysis. The extraction process of MFCC features includes the following steps: Pre-emphasis: preprocessing the original audio signal, and enhancing the high-frequency components of the signal makes the subsequent processing more stable. Framing: the pre-emphasized audio signal is segmented according to a fixed frame length, and the frame length of 20-40ms is usually selected. Windowing multiplies the audio signal of each frame by a window function (such as Hamming window) to reduce the influence of abrupt changes at the frame boundary on spectrum analysis and increase continuity. Fast Fourier Transform (FFT) performs fast Fourier transform on each frame signal, and converts the time domain signal into the frequency domain signal. Mel Filter Bank converts frequency spectrum into energy distribution on Mel scale through a set of filters. These filters are usually triangular, covering a range of frequencies on the Mel scale. Logarithm operation performs logarithmic operation on the output of each Mel filter to enhance the characteristics of the low frequency part. Discrete Cosine Transform (DCT) is used to transform the logarithmic energy spectrum and extract MFCC coefficients. Usually only the first few coefficients are kept because they contain the most important information. The MFCC feature extraction process is described above. Through these steps, audio signals can be transformed into a set of MFCC coefficients with good discrimination and robustness, which can be used in speech recognition and speaker recognition.

MFCC is a cepstrum parameter extracted in the frequency domain of Mel scale, which describes the nonlinear characteristics of human ear frequency, and its relationship with frequency is shown in Formula (1):

$$\text{Mel}(f) = 2595 \times \lg \left[1 + \frac{f}{700} \right] \quad (1)$$

The calculation method of MFCC is shown in Formula (2):

$$MFCC(t, i) = \sqrt{\frac{2}{N}} \sum_{j=1}^N \lg [E(t, j)]$$

2.2 emotion classification model combining LSTM and AM

This paper plans to classify the emotional categories of music in this model. The neural network model, which combines the two-layer LSTM structure and attention mechanism, consists of three layers of LSTM, attention layer and output layer. Its model framework is shown in Figure 2.

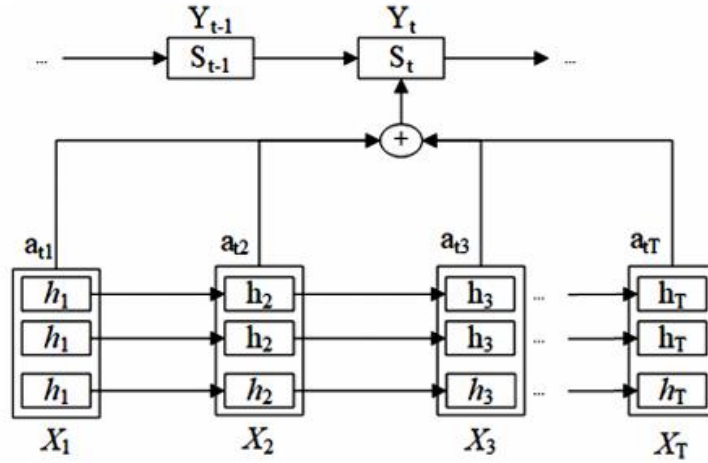


Fig. 2 LSTM attention neural network model

The audio signal is processed as MFCC feature as the input of neural network. Two-layer LSTM structure is used to learn the characteristics of time series, in which the output of the first layer LSTM is used as the input of the second layer LSTM to enhance the representation ability of the model. The attention layer is added behind the LSTM structure to adaptively adjust the importance of different time steps and improve the discrimination ability of the model. Through a fully connected layer, the attention-weighted LSTM output is mapped to the emotion category, and finally the emotion category prediction result of music is obtained. Generally speaking, this neural network model adopts multi-layer structure and attention mechanism, which can effectively extract the time series characteristics of music and adaptively adjust the information at different time steps, thus realizing the classification of music emotion categories.

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

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

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

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

The traditional RNN model is easy to encounter the problem of gradient disappearance or gradient explosion when dealing with long sequences, which will make it difficult to continue learning and optimizing network parameters. As a special type of RNN, LSTM (Long-term and Short-term Memory Network) can effectively solve this problem through the design of gate structure. LSTM introduces a memory unit called cell state, which can retain and transmit information. The gate structure in LSTM includes forget gate, input gate and output gate, whose purpose is to control the information flow in the cell state. The forgetting gate determines which information in the previous cell state should be discarded, the input gate determines which information in the current input should be added to the cell state, and the output gate determines which information in the cell state should be output. In this way, LSTM network can selectively retain, update or output important information, so as to better handle and predict important events in time series. In addition, LSTM also introduces a concept called candidate cell state, which is used to calculate a new cell state under the control of the input gate. This enables LSTM to adapt to different input situations by increasing or decreasing the information of cell state, so as to better handle and model long-term dependencies.

In the field of natural language processing, attention mechanism is widely used in machine translation, text summarization and other tasks. For example, in the task of machine translation, attention mechanism can help the model pay attention to the part of the input sequence that is

related to the current output position selectively, thus improving the quality and fluency of translation.

Usually, the attention mechanism is realized by calculating the attention weight of each input position. These weights represent the importance of each position to the current output. Usually, these weights are calculated by a feedforward neural network, which takes the features of the current output and all input positions as inputs and outputs the corresponding attention weights.

Specifically, given the input sequence and the output sequence, the attention mechanism can calculate the attention weight of each input position in the following way. $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_m)$

$$a_i = \frac{\exp(f(x_i))}{\sum_j \exp(f(x_j))}$$

Usually, it can be calculated by dot product, additivity or splicing. In this way, the model can learn different degrees of attention to different input positions according to the information of the current output and all input positions.

3. piano music recommendation model based on multimodal learning

3.1 User's Emotional Analysis

This model combines low-level descriptors LLD, such as MFCC, formant, short-term energy, pitch frequency and short-term zero-crossing rate, as well as LSTM and AM (attention mechanism). Specifically, the model diagram is as follows (assuming Figure 3):

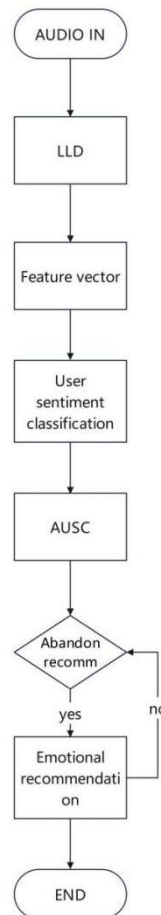


Figure 3 User Emotion Judgment Model

In this model, you can use MFCC(Mel frequency cepstrum coefficient) as a sound feature extraction method, which can transform the sound signal into a set of features similar to human ear perception. Formant, short-term energy, pitch frequency and short-term zero-crossing rate are also commonly used sound feature descriptors. Next, LSTM (Long-term and Short-term Memory Network) will be used to model the user's historical listening records. LSTM can capture the time dependence between audio sequences, and process long sequences through gate structure and memory unit. Finally, attention mechanism (AM) is introduced to enhance the performance of the model. Attention mechanism can help the model focus on the most important part of historical listening records, so as to judge users' emotions more accurately.

3.2 Feature Extraction of Historical Listening Music

The feature extraction process is basically consistent with the previous emotion classification process. In the feature extraction of historical listening music, it is necessary to obtain the LLD description of music and process the music in segments, each of which lasts for 3 seconds. Each piece of music needs to contain 50-dimensional original audio features such as 18-dimensional MFCC, 8-dimensional formant, 10-dimensional short-term average energy, 6-dimensional audio rate and 8-dimensional short-term zero-crossing rate. Then, the obtained feature sequence is weighted and averaged. In the selection of historical listening records and emotional classification, you can choose historical records according to the ranking of playing times in different time ranges. Generally speaking, the ranking of the number of plays in the past week is more valuable for reference. You can select the top k songs played in the past week, extract the feature sequence of each song, and then carry out weighted average processing on these sequences. Through the above process, we can get high real-time feature sequences, which will be used in the next emotion classification task.

3.3 Selection of Historical Listening Records and Emotional Classification

In the selection of historical listening records and emotional classification, you can select the top k songs played in the past week and extract their feature sequences for each song. Then, these feature sequences are weighted and averaged to get a new feature sequence with high real-time performance. The formula of weighted average has been given in your description. Then, this new feature sequence is transmitted into the three-layer LSTM structure as input data. Through the screening of attention layer and hidden layer, the corresponding relationship between output and emotion category is strengthened, and finally the obtained data is mapped to softmax layer for multi-classification discrimination. The emotional categories are still happy, fresh, relaxed and sad.

3.4 Specific steps of music recommendation model

The specific steps of the music recommendation model are as follows: obtain the historical listening records of users and filter the music in the records. For the selected music, the 50-dimensional original audio signal features of each music are extracted, and the feature sequences of these music are weighted and averaged to get a new feature sequence. The new feature sequence passes through the three-layer LSTM and AM structure in turn, and a new feature vector is obtained, and the user's current emotion category is analyzed and determined. Users can choose whether to accept this recommendation. If it is accepted, the recommendation is successful, and the music under the emotion tag will continue to be recommended to users in the future; If you don't accept it, you need to go back to step 3 and re-analyze the current emotional state of the user in order to recommend other music next. Through this processing, the original audio information can be analyzed, and at the same time, the user's emotional change can be considered more, and the music resources such as the user's historical data can be fully utilized, and all aspects of music characteristic information can be utilized more comprehensively. In addition, this method has a short cycle and is closer to the current emotional state of users, which can greatly improve the efficiency of music recommendation.

4. experimental design

4.1 data selection

In this experiment, the data set of Netease cloud music is selected as the music source in the training set and the comparative experiment. This data set contains 10,000 songs, each with emotional labels, including happiness, freshness, relaxation and sadness. Each song also carries information such as song name, singer name and album name. In the training set, music with emotional tags is used as the data set. In order to avoid recommending the same song repeatedly, the song name, singer name and album name are used to judge whether a song is repeatedly recommended. Compared with other websites, such as Last.FM, QQ Music, Xiami Music and Cool Dog Music, the data set of Netease Cloud Music is easier to obtain the types and frequency of songs played by other users in the past week.

4.2 Parameter design

In this experiment, 10,000 pieces of music were crawled from Netease Cloud Music, each with more than one emotional tag, and the duration was less than 6 minutes. Divide each piece of music into 3s segments to get audio segments arranged in chronological order. In the feature extraction stage, 50-dimensional features are extracted, including MFCC 18, formant 8, short-term average energy 10, pitch frequency 6 and short-term zero-crossing rate 8. When the feature sequence is mapped to the range of [0,1], it is normalized by max-min. In the emotion classification stage, the three-layer LSTM network and AM are used, the learning rate is set at 0.002, the number of Epoch is set at 1000, the value of Dropout is set at 0.7, the tanh activation function is selected, the Batchsize is set at 128, and SGD is used in the optimizer. In the music recommendation model, the top 10 songs played most frequently by each user in the last week are selected, and 2s is taken as a segment, and each segment contains 50-dimensional audio features of the same category in the classification model. Sigmoid function is selected as the activation function of hidden layer, and the number of nodes of softmax is 6, the number of hidden layer nodes before sigmoid is 64, and the loss function of softmax is CE.

4.3 Experimental results

In the experimental environment of this paper, the CPU configuration used is Intel i7 8th Gen, and the framework is TensorFlow. The writing environment uses Jupyter Notebook and PyCharm. When the data set is divided into training set and test set, the 50% cross-validation method is used in the music emotion classification model, and the accuracy of the model which combines 50-dimensional features of audio, three-layer LSTM and AM is compared with other models. The specific comparison models include: 50-dimensional feature +LSTM model, 50-dimensional feature +LSTM+AM model, 50-dimensional feature+three-layer LSTM+AM model, classification model of support vector machine SVM, and the comparison results of classification models of LDA model with hidden Dirichlet distribution are shown in Table 1, which shows the accuracy of emotion classification of each model.

Table 1 Accuracy of Emotion Classification of Music

model	Average accuracy/%
50-dimensional feature +LSTM	65.8
50-dimensional feature +LSTM+AM	68.2
50-dimensional feature+3-layer LSTM+AM	71.6
SVM	69.2
LDA	66.3

Table 1 shows the average accuracy of music emotion classification using different models.

The results show that only using audio features and traditional LSTM model has the lowest classification accuracy, but after adding attention mechanism, the accuracy and concentration of emotion categories are improved. The model framework using audio features, three-layer LSTM network and AM has the highest classification accuracy.

Table 2 Emotion recognition rate of music

emotion	happy	pure and fresh	relax	sentimental
happy	74.5	11.2	9.6	4.7
pure and fresh	10.1	68.3	16.4	5.2
relax	4.6	11.5	78.2	5.7
sentimental	6.1	10.3	10.9	72.7

Table 2 shows the recognition rate of four musical emotions (which should be the result of a data set). These results show that the music emotion classification model using attention mechanism and multi-layer LSTM network can improve the classification accuracy and achieve good results in music emotion recognition. This table shows the recognition rate between different types of musical emotions. For example, when music is happy, 74.5% of the accuracy is recognized as happy, 11.2% is wrongly recognized as fresh, 9.6% is wrongly recognized as relaxed, and 4.7% is wrongly recognized as sad. Other emotional categories also have corresponding recognition rates. These results can help us understand the performance of the music emotion classification model, as well as the discrimination and accuracy between different emotion categories.

5. tag

For the research on the emotion recognition algorithm of piano music based on multimodal learning, the conclusion can be summarized according to the contents of specific papers. In this study, we propose a piano music emotion recognition algorithm based on multimodal learning, and verify it by experiments. By combining audio and visual features, we can identify the emotion of piano music more accurately. The experimental results show that the multi-modal learning method has a significant improvement in the emotional recognition of piano music compared with the single-modal method. By considering both audio and visual information, we can understand the emotions conveyed by piano music more comprehensively and extract richer features from it. In addition, we also introduce attention mechanism to improve the model's attention to key features. The experimental results show that attention mechanism can help us better understand the relationship between audio and visual information, and extract the most useful features for emotion recognition. Generally speaking, our research shows that the piano music emotion recognition algorithm based on multimodal learning and attention mechanism shows advantages in accuracy and effect. This study provides useful reference and enlightenment for further exploring the field of musical emotion recognition. It should be noted that this is only an example and does not represent the conclusion of a specific paper. In fact, for a specific research paper, the conclusion should be summarized and described according to the specific content and results of the research.

References

- [1] Sheng Tingyu, Feng Qiansheng, Yin Erwei, et al. An emotion recognition device based on multimodal emotion model: CN202210265773.7 [P]. CN202210265773.7 [2024-01-12].
- [2] Zhou Ping. Music Emotion Classification Algorithm Based on Multimodal Deep Learning [J]. Intelligent Computers and Applications, 2022, 12(9):5.
- [3] Cheng Min. Classification method of music emotion appreciation based on multimodal deep learning [J]. Journal of Anyang Institute of Technology, 2023, 22(5):113-117.
- [4] Zhang Xiao, Wang Tianli, Yu Dongxiao, et al. An emotion recognition method based on multimodal clustering federated learning.cn CN202211309647.3[2024-01-12] 01-12].
- [5] Jiang Yanshuang, Cui Can, Yun Xing, et al. Emotional analysis of multimodal learning in double-teacher classroom: key issues, logical reasoning and implementation route [J]. Modern Educational Technology, 2022, 32(4):8.
- [6] Liu Ying, Ai Hao, Zhang Weidong. Summary of multimodal emotion recognition based on deep learning [J]. Journal of Xi 'an University of Posts and Telecommunications, 2022(027-001).
- [7] Mei Shuzheng, Shi Xinxin, Wu Weihua, et al. Emotion recognition method and related equipment based on multimodal: CN202211500520. X [P]. CN115775565a [2024-01-12].
- [8] Wang Huahua, Zhang Ruizhe, Huang Yonghong. Identification method of spread spectrum and conventional modulation signals based on generative countermeasure network and multimode attention mechanism [J]. Journal of Electronics and Information, 2024,44 (Yu): 1-10. DOI: 10.11999/JEIT 230518.