# Research on Multi-modal Music Emotion Classification Based on Audio and Lyirc

Gaojun Liu[1],Zhiyuan Tan[1]
1. North China University of Technology, BeiJing, China
Email: lgj@ncut.edu.cn, 906208803@qq.com

*Abstract*— **To solve the problem of low accuracy of emotion classification, this paper proposes a new multi-modal fusion emotion classification method based on audio and lyrics. Firstly, Mel Frequency Cepstrum Coefficient, spectrum centroid and frequency-band energy distribution are used as feature data in audio, and LSTM in deep learning is applied to music emotion classification; In terms of lyrics, the Bert model is used to classify the lyrics, and the sentiment dictionary is used to perform LFSM-based equalization on the lyrics emotion classification results. Finally, a new fusion method is proposed on the traditional fusion method. The experimental results show that the new fusion method has 5.77% and 4.03% improvement over the linear weighted multimodal fusion and LFSM fusion methods.**

***Keywords—music emotion classification; Bert; MFCC; Multimodal***

## I. Introduction

With the rapid development and popularization of Internet technology, more and more people choose to listen to songs on the Internet, and expect to search or retrieve music in a manner related to music content, such as emotion, genre, lyrics, style, etc. Among them, the automatic recognition of music emotion has caused widespread attention, and has been widely used in music information retrieval system and music recommendation system, such as advanced semantic retrieval based on emotional music, recommendation based on user's emotion and song's emotion [1].

After Krumhansl used mathematical models to describe the problem of music sentiment classification in 1997 [2], people began to focus on music sentiment classification and research, mainly including two aspects of music audio content and music lyrics text; In terms of audio, the commonly used method is to analyze the acoustic features extracted from the music audio to get the classification result of music emotion. However, using music audio as a classification basis, the effect is often not ideal; Meanwhile, the music lyrics text itself contains emotional information, so the classification effect based on the fusion of music audio content and music lyrics text is better than the classification effect of single audio features and single lyrics features [3-4]. This article uses emotional audio and lyrics to classify emotions separately, and then uses improved fusion methods to improve the accuracy of emotion classification.

## II. Music Emotion Model

There are two main emotional models of music: Thayer's emotional model and Hevner's emotional model. Thayer emotional model is a two-dimensional emotional model [5] (Fig. 1). The vertical and horizontal coordinates represent energy and pressure respectively. Energy and pressure divide the plane into four areas, representing four types of music emotion.

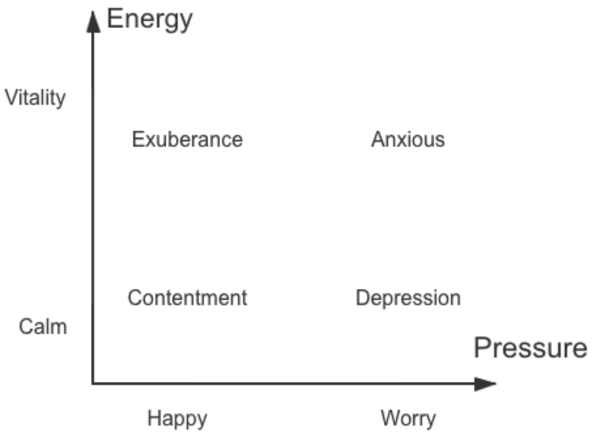

Fig. 1. Thayer's two-dimensional emotional model.

Hevner's emotion model is a discrete category model [6]. This model uses dozens of adjectives to describe music emotion. These adjectives are divided into eight categories according to the difference between emotions. Thayer emotional model has better emotional smoothness than Hevner emotional model, especially in the two factors of energy and pressure, which can better correspond to acoustic characteristics; and also has higher differentiation between emotions, which is closer to human emotions. Therefore, the musical emotion classification studied in this paper is based on Thayer two-dimensional emotional model.

## III. Emotional Classification of Audio Content

### A. Characteristic Parameters

According to different hierarchy, music audio content can be divided into three levels: bottom level, middle level and high level. The high-level label is the information specifically marked by human senses and cognition, the middle level and bottom level features represent the sign features that exist objectively. The main features extracted in this paper are middle level and bottom level features: Mel frequency cepstrum coefficient (MFCC), spectrum centroid and frequency-band energy distribution.

Mel Frequency Cepstrum Coefficient (MFCC)

MFCC[8,9] is the most commonly used feature parameter in the field of music emotion recognition. It is proposed by

**978-1-7281-4390-3/20/$31.00 ©2020 IEEE**

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 24,2020 at 15:42:21 UTC from IEEE Xplore. Restrictions apply.

Mermel and others based on human language and hearing, and has the characteristics of high recognition and strong noise resistance. MFCC have pre-weighting, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank, Discrete Cosine Transform (DCT) and other processes, especially FFT and Mel filter banks, which implement dimension reduction operations respectively. This paper use Librosa : a special library for audio data processing, to extract MFCC features from audio data. The result is a two-dimensional array, one dimension represents time, and the other dimension represents frequency.

Spectrum Centroid

The spectrum centroid is to normalize each frame of the amplitude spectrum graph and extract the average value from each frame, which is the important information of the energy and frequency distribution of the sound signal; The low-frequency content reflects the gloomy and dim quality sound, with a low spectrum centroid; The high-frequency content mostly reflects the bright and happy quality sound, with a high spectrum centroid. Therefore, the spectrum centroid is also an important characteristic parameter of audio classification.

Frequency Band Energy Distribution

The frequency band energy distribution reflects the law that the energy of the audio signal changes with the frequency, which can explain the distribution of the energy contained in the different frequency ranges in the audio. The songs with anxiety and depression are generally slow and less energy, while the songs with strong energy are generally happy and excited, so the frequency band energy distribution is also an important feature of music emotion. In this paper, the average frequency band energy values of different frequencies are selected as parameters, and finally 5-Dimensional parameters are obtained.

### B. Network Model

In the field of deep learning, there are two kinds of neural networks: Convolutional neural network (CNN) and Recurrent neural network (RNN). CNN is mostly used to process image type data, RNN is mostly used to process data with continuous time series. As a special type of RNN, LSTM is often used to deal with the problem that RNN cannot be relied on long time. LSTM was proposed by Hochreiter & Schmidhuber [10] (1997) , and recently improved and promoted by Alex graves [11]. On many issues, LSTM has made great achievements and has been widely used in various fields. The structure of LSTM network model (Fig. 2).

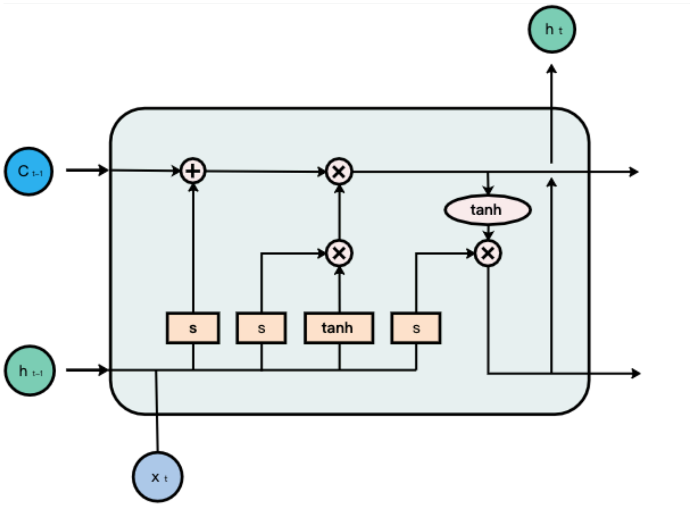

Fig. 2. LSTM network model diagram.

In this study, according to the characteristics of the original training data, the input layer of the designed network structure model is the extracted feature data MFCC (0-13), spectral centroid (14-26), frequency band energy distribution (27-31) The number of neurons in the LSTM hidden layer is 128, the number of neurons in the second LSTM hidden layer is 32, and the final output layer is 4 neurons, which represent the 4 emotional probabilities of music classification.

## IV. Emotional Classification of Lyrics

### A. Bert

In 2018, Google AI team released a new language model Bert [12], which brought a breakthrough progress to the pre-training model in natural language processing. Bert is a multi-layer bidirectional transformer encoder, whose structure is shown in Fig. 3.

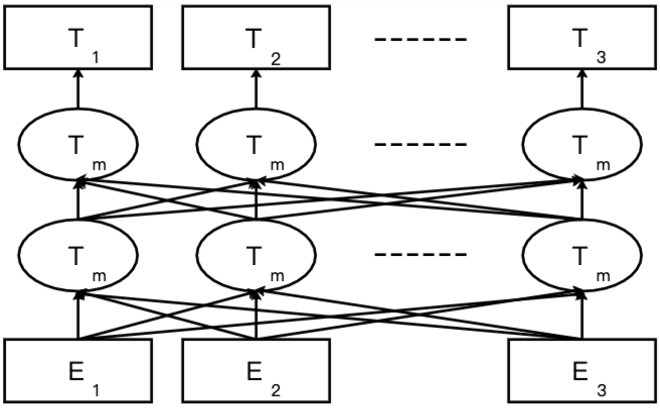

Fig. 3. Bert Structure.

Bert follows the language encoder in attention is all you need [13], puts forward the concept of bidirectional, and uses the masked language model to realize bidirectional. Bert proposes two kinds of pre-training methods: Masked LM (masked language model) and Next Sentence Prediction. The first method is to use Masked LM to realize the pre-training of two-way language model. Unlike other language models, it is to predict

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 24,2020 at 15:42:21 UTC from IEEE Xplore. Restrictions apply.

all words in the input sequence. Masked LM randomly selects 15% of the words in the input data for masked operation through context. The second method is Next Sentence Predicton, which is used to judge whether two sentences are two classification tasks. In the training data, 50% of the data are trained with real up and down sentences as positive examples, and the rest are randomly selected 50% of the data as negative examples Training. The final pre-training result of this task can reach 97% ~ 98% accuracy.

In this experiment, Chinese 's pre-trained model "BERT-Base, Uncased" was used. This model uses a 12-layer Transformer and outputs a dimensional vector with a size of 768. The multi-head attention parameter is 12, and the total model size is 110 MB.

## V. Fusion Method

At present, according to the music audio content and music lyrics text, the multimodal music emotion classification method can be roughly divided into two categories. The first category is to combine the features of music audio content and music lyrics text to pass the classifier together. The second category is to classify the music audio content and music lyrics text respectively, and then combine the two classification results to determine the music emotion again Sense category. This paper adopts the second type, which mainly includes the following fusion methods:

### *A. Subtask Combined with Late Fusion Method (LFSM)*

LFSM method [14] is based on the fusion method of two-dimensional emotional model, and proposes that the music content and lyrics have better differentiation in energy and pressure [15-16]; The combination method is shown in TableI.

TABLE I. LFSM Fusion Method

| *Energy* | *Pressure* | *Emotion* |
|---|---|---|
| vitality | worry | anxious |
| vitality | happy | exuberance |
| calm | worry | depression |
| calm | happy | contentment |

### *B. Improved Fusion Method*

In this paper, a new fusion method is proposed: because audio has a better discrimination in energy and lyrics has a better effect in stress, if the result of audio content classification and lyrics classification belong to the same category: anxiety and depression, exuberance and contentment, then decision-making fusion is carried out; if the two classification results do not belong to the same category, because on the whole The accuracy of the classification results of audio content is higher than that of lyrics. In terms of lyrics, emotional dictionaries are used to analyze the lyrics. Finally, the lyrics are divided into two categories: happiness and worry, and the scores from emotional dictionaries are mapped to [- 1,1] in proportion, then the classification weight of lyrics is adjusted, and then the decision-making level is made with the classification results of audio content after adjustment, Finally, we get the result of music emotion classification.

In the decision level fusion method, the linear weighted decision level fusion is not used, but it is improved. The neural network is introduced, which can give different weights to each mode and each category respectively. After the audio classification results and lyrics classification results are obtained, they are spliced and combined, and then used for the input of the full connection layer, and finally the classification results are obtained.

For example, the order of classification results is exuberance, anxious, contentment and depression. The audio classification results are {0.1, 0.3, 0.5, 0.1}, and the lyrics classification results are {0.3, 0, 0.6, 0.1} from a song. LFSM can know that the results of audio content and lyrics classification do not belong to the same category in terms of pressure. Then these two results are used for the input of full connection layer to get the final classification Results: if the result of lyrics classification is {0, 0.6, 0.3, 0.1}, LFSM can know that the audio content and lyrics do not belong to the same category in terms of pressure, then the emotional dictionary is introduced, if the score is positive and 0.4, the adjusted lyrics classification result is {0.2, 0.6, 0.5, 0.1}, and then the audio content classification result is the input of the full connection layer, and the final classification result is obtained.

## VI. Experiment

### *A. Date Set*

The data set of this paper is 'Music Mood Classification Data Sets' [17] in reference. There are 777 songs in the data set, including audio data and text data, and four emotional categories: exuberance, anxious, contentment and depression, the number of songs is 171, 201, 206 and 199 respectively. The number of songs was increased to 1200 : download from LAST.FM music website according to English song list of emotion tag .300 songs of each category, 1000 of which were used as training samples and 200 as test samples. The audio adopts the same format: sampling frequency is 16000 Hz, WAV, mono channel, and the middle 30s music segment of the song is intercepted as the emotional classification.

### *B. Experiment Process*

The experimental flow of this paper is shown in the figure below (Fig. 4).

The specific steps are as follows:

Step 1: the processing of audio data and the extraction of audio content features, in which the middle 30s music segment is intercepted.

Step 2: the lyrics data processing, the lyrics into the appropriate format.

Step 3: training; grouping the data and training the lyrics and audio models at the same time.

Step 4: according to the result of training, improve the decision-making and integration.

Step 5: fine tune the test; adjust the weight and other parameters according to the results, and evaluate the experimental results.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 24,2020 at 15:42:21 UTC from IEEE Xplore. Restrictions apply.

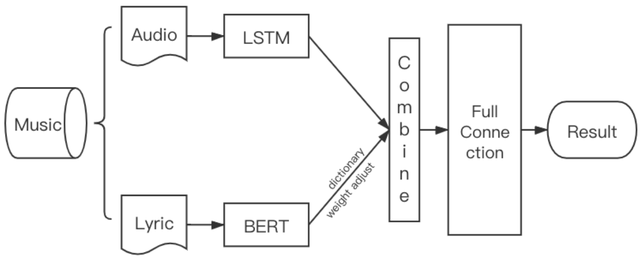

Fig. 4. Experiment Flow Chart.

### C. Annlysis of Experimental Resultss

In this study, 1000 songs in the training set were randomly divided into 4 groups of the same number, and classified training was carried out respectively. The following table (TableII) shows the accuracy rate after training, with an average of 79.70%.

TABLE II. GROUP CLASSIFICATION RESULT

| *Group* | *1* | *2* | *3* | *4* | *AVG.* |
|---|---|---|---|---|---|
| Accuracy(%) | 81.01 | 78.26 | 79.10 | 80.43 | 79.70 |

In addition, the classification results of LFSM and improved fusion method on four types of emotions are shown in the table (TableIII). According to the analysis in the table, it can be concluded that the classification effect of depression is better, the accuracy is 86.10% and 90.81%, respectively, and the classification effect of anxiety is poor, the accuracy is 64.21% and 74.02%, and the improved LSFM classification can be seen that the effect of the method has a higher improvement in anxiety and depression. According to the LSFM, it can be seen that there is a better improvement in stress (Lyrics). Among them, the improvement of emotion as anxiety is more significant. It also verifies the emotion dictionary It has a positive impact on the classification results, but the improvement of exuberance and contentment classification is not very obvious.

TABLE III. CLASSIFICATION RESULTS OF IMPROVED METHOD

| | *Anxious* | *Exuberance* | *Depression* | *Contentment* |
|---|---|---|---|---|
| LFSM | 64.21 | 77.32 | 86.10 | 74.73 |
| Improved LFSM | 74.02 | 77.64 | 90.81 | 76.33 |

### D. Method Comparison

Under the same experimental environment, the methods of only audio, only lyric, linear weighted fusion, LFSM, improved fusion are designed respectively. The experimental results are shown in the table(TABLEIV).

TABLE IV. METHOD COMPARISON

| *Method* | *Accuracy(%)* |
|---|---|
| Vocal+RNN | 67.02 |
| Vocal+LSTM | 70.63 |
| Lyric+LSTM | 59.80 |
| Lyric+BERT | 62.97 |
| Linear Weighted | 73.85 |
| LFSM | 75.59 |
| Improved LFSM | 79.62 |

It can be seen from the table that the accuracy rate of only audio and only lyrics is lower than that of audio + lyrics, with better results of 70.6% and 62.9% respectively. At the same time, the accuracy of the fusion method has been improved. The accuracy of linear weighted fusion is 73.65%, LFSM and improved LFSM is 75.59% and 79.62 respectively. Because the linear weighted fusion only calculates the weight and does not consider the performance of the song in terms of energy and pressure. The improved LFSM not only utilizes the emotional dictionary balance and enhances the emotion classification in lyrics, but also introduces a neural network to improve the overall emotion classification situation.

## VII. CONCLUSION

This paper introduces a new multi-modal music emotion classification method based on music audio content and music lyrics text. In terms of audio, it is proposed to use the LSTM network for classification, and the classification effect is significantly improved compared to other machine learning methods. In terms of lyrics, it is proposed to use Bert to classify lyrics emotions, which effectively solves the problem of long-term dependence. In terms of multi-modal fusion, LSFM is proposed in lyrics, emotion dictionary is used to adjust the emotional classification results of lyrics, and neural network is introduced on the basis of linear weighted decision-making level fusion. It can be seen from the experimental results that the decision-level fusion effect has been improved, and the accuracy of each category classification has also been improved. The next work focuses on how to further improve the classification accuracy, especially the classification accuracy based on the content of lyrics, and try other fusion methods to improve the fusion effect.

## ACKNOWLEDGMENT

This work was supported by CNONIX National Standard Application and Promotion Lab (4020548418X8) of North China University of Technology.

## REFERENCES

[1] SCHEDL M, ZAMANI H, CHEN C W, et al. Curent chalenges and visions in music recommender systems research [J]. International Journal of Multimedia Information Retrieval, 2018:1-22.

[2] Krumhansl CL. An exploratory study of musical emotions and psychophysiology. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expéri- mentale, 1997, 51(4): 336–353. [doi: 10.1037/1196-1961. 51.4.336] .

[3] Yang Dan, Lee W S. Music emotion identification from lyrics [C]/ /Proc of the 11th IEEE international symposium on mul- timedia. Washington D C, USA: IEEE Computer Society, 2009: 624-629.

[4] Hu Xiao, Downie J S. Improving mood classification in music digital libraries by combining lyrics and audio[C]/ / Proc of the 10th annual joint

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 24,2020 at 15:42:21 UTC from IEEE Xplore. Restrictions apply.

conference on digital libraries. New York，USA: ACM Press，2010: 159-168.

[5] Taylor J G，Fellenz W A，Cowie R，et al. Towards a neural-based heory of emotional dispositions [C]. Proc of IMACS IEEE CSCC’ 99, 1999.

[6] Hevner K. Experimental studies of the elementsof expression in mu- sic[J]. American Journal of Psychology，1936，48 ( 2 ) :246 - 268.

[7] Cuthbert MS, Ariza C. music21: A toolkit for computer- aided musicology and symbolic music data. Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010). Netherlands. 2010. 637–642.

[8] Loughran R, Walker J, O'Neill M, et al. The use of mel-frequency cepstral coefficients in musical instrument identification[C] //International Computer Music Conference. International Computer Music Association:Belfast,N.Ireland,2008.

[9] LeeCH,ShihJL,YuKM,etal. Automatic music genre classification using-modulation spectral contrast feature[C]//Multimedia and Expo, 2007 IEEE International Conference on. IEEE:Piscataway, NJ, 2007.

[10] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[11] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[12] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[13] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. 2017.

[14] Yang Yihsuan，Lin Yuching，Cheng Hengtze，et al. Toward multi-modal music emotion classification[C]/ / Proceeding of pacific rim conference on multimedia. Tainan，Taiwan: [s. n. ]，2008: 70-79.

[15] Lu Lie，Liu Dan，Zhang Hongjiang. Automatic mood detection and tracking of music audio signals[J]. IEEE Trans on Audi- o，Speech and Language Processing，2006，14( 1) : 5-18.

[16] Yang Y H，Lin Y C，Su Y F，et al. A regression approach to music emotion recognition[J]. IEEE Trans on Audio，Speech and Language Processing，2008，16( 2) : 448-457.

[17] XUE H，XUE L，SU F. Multimodal music mood clas sification by fusion of audio and lyrics [C ]. International Conference on Multimedia Modeling. Springer，Cham，2015:26-37.

Y.L. Jiang, C.F. Xu, “Advances in Set Pair Analysis Theory and its Applications,” Computer Science, vol. 33, issue 1, pp. 205-209, 2006.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 24,2020 at 15:42:21 UTC from IEEE Xplore.  Restrictions apply.