Research on Musical Sentiment Classification Model Based on Joint Representation Structure

Chunjun Zheng¹², Ning Jia²

¹Dalian Maritime University, Liaoning Dalian, China
²Dalian Neusoft University of Information, Liaoning Dalian, China
zhengchunjun@neusoft.edu.cn

Abstract. In the traditional music emotion classification process, there are problems such as low classification accuracy rate, long period, and difficulty in satisfying the individualized needs of the theme music in people's lives. Based on this, a neural network model based on joint representation structure is designed. The model uses low-level descriptors and spectrograms to construct a joint representation of the characteristics of the manual and convolutional recurrent neural network, thus realizing the discrimination of music emotion subclasses. At the time of the experiment, the model was designed and the CRNN traditional model was used as the baseline. The experimental results show that this model can improve the classification accuracy of music emotions compared with the traditional single model.

1. Introduction
With the development of music streaming and self-media, various forms of massive audio data are popular on the Internet. How to quickly search for the user's favorite music is one of the hot issues at present. In the field of music classification, music classification according to related topics is the basis of all research. Due to the variety and form of music and the differences between music, music classification is an extremely complicated task, and the problem of low classification accuracy and long period of time often occurs. Based on this, this paper selects the basic characteristics of the speech signal, namely emotion, and based on this, completes the emotional classification of music.

The emotions for speech signals mainly include two areas of research: music feature extraction and classification model construction. Many scholars look for effective features to characterize the emotions of music itself. G. Tzanetakis et al. use the Mel Frequency Cepstral Coefficient (MFCC) in the audio signal extraction process. Lee.K.K et al. proposed a robust feature extraction method combined with SFS feature selection, namely multi-feature clustering (MFC). Du.W et al. extracted spectrogram features from audio signals and layered them to improve the accuracy of classification.

There are problems such as low precision, large noise impact, and poor robustness in the practical application. A mature deep learning framework is generally selected and used in combination. Conversely, if a multi-dimensional feature combination is used, it will lead to a very large amount of computation. At this time, a mature deep learning framework is generally selected and used in combination. Mirsamadi S et al. added the local attention mechanism to the Recurrent Neural Network (RNN) to focus on extracting short-term frame-level acoustic features related to emotions to achieve automatic recognition of speaker emotions.

On the basis of the decomposition of the original acoustic signal, the three methods of feature extraction related to emotion recognition are summarized, which are: extracting signal features from...
the original audio file\cite{10}; deep learning model running on the original audio waveform; Automatic Speech Recognition (ASR) technology is used to convert speech into text and then analyze text, but the success rate of this method is low.

Based on this, this paper proposes a neural network model based on joint representation structure for music classification. This network model uses the first two methods of fusion feature extraction to perform user's emotion recognition, forming a joint representation of low-level descriptor features and CRNN features.

2. Music emotion classification model based on joint representation structure

The design process of music emotion classification model based on joint representation structure is shown in Figure 1. For the existing training set, multi-dimensional speech related features are extracted and fed into the HSF and CRNN networks to complete the construction of this model.

2.1. Music data crawling

For the massive music files in the Internet, the strategy adopted in this paper is to automatically crawl new music resources in the music resource website within the specified range, and complete the data storage on the basis of data inspection and analysis. The data crawling process is implemented based on the Hadoop framework. In the case of a known URL address, the organization of the music resources in the website is judged. It is worth noting that the completeness and correctness of the data must be ensured during the crawling process to avoid data acquisition failure due to instability of the web server.

In the data preparation process, how to design a crawling strategy is particularly important. In this paper, the data is obtained in a depth-first form. According to this strategy, when crawling, it is first necessary to summarize the rules of the URL address of the website, extract the path of the difference, split the page into multiple sets of sub-pages, and put them into different threads.

2.2. Music emotion feature extraction

On the basis of obtaining music data, it is first necessary to extract the emotional characteristics of music, and then judge the music emotion category. It can be seen that selecting the appropriate feature descriptor is very important for the classification result.
The speech signal is converted into a speech feature vector, which combines LLD and High-level Statistic Functional (HSF). Common LLDs can be roughly classified into three types: prosodic features, spectral features, and sound quality features. Common extraction methods are: linear prediction coefficient (LPC), MFCC and linear predictive cepstral coefficient (LPCC). Sound quality features mainly include breathing sounds, throat sounds, phonemes, word boundaries, brightness, etc.

2.2.1 MFCC characteristics
From the perspective of human ear perception loudness and amplitude, the MFCC feature embodies the envelope characteristics of the channel shape in the short-term power spectrum of speech.

The calculation process of the MFCC feature is shown in Fig. 2. On the basis of framing and windowing, a discrete Fourier transform is performed on each frame signal to calculate a logarithmic amplitude spectrum. It is filtered by an equal-bandwidth Mel filter bank and passed through discrete cosine transform to obtain the MFCC feature.

![Fig. 2 the extraction method of MFCC](image)

For music signals, MFCC is mainly embodied as static features. The dynamic characteristics can be expressed by the delta and delta-delta of the MFCC feature. Through convergence, the accuracy is effectively improved.

2.2.2 The frequency of the pitch
The frequency of the pitch is generally the representation of the lowest frequency of the sound vibration, and the variation can be understood as the pitch. Therefore, the frequency of the pitch generally contains discriminating information of the music. In this paper, the fundamental frequency of the current frame is mainly extracted, and 8-dimensional features such as maximum value, mean value, variation range, standard deviation, median, and quarterback number are recorded.

2.2.3 Formant
When a glottal pulse passes through the channel, the formant is the frequency at which resonance occurs inside. It is generally considered that the maximum value of the spectral envelope is the formant, which is one of the important parameters of tone and sound quality. Based on this, the cepstrum method is used to calculate the mean value, standard deviation, median and bandwidth of the first, second and third formants, including 12-dimensional features.

2.2.4 Band energy
Since the distribution of band energy reflects the degree of change of music intensity, and the emotional characteristics of music are often related to energy intensity, the distribution of band energy has guiding significance for the judgment of music type. In this paper, 500 Hz is used as the energy band interval difference, and the mean values of the eight band energy intervals in the 0-4000Hz interval are extracted respectively, and the 8-dimensional features are obtained.

2.2.5 Beat
Rhythm features typically include beats, tempo, rhythm, etc., and features can be extracted using a Beat Histogram, which is the result of several filtering of the time domain signal. From this, six values,
such as the value, period, ratio and amplitude, of the two peaks are extracted. The calculation process of the beat histogram is shown in Figure 3.

![Fig.3 Calculation method of beat histogram](image)

Since the above features are closely related to the music data, the above 76-dimensional features are integrated, and the labelling process is carried out. Considering the characteristics of music timing, this paper divides the music into several segments with a duration of 3s, and extracts a 76-dimensional segment for each segment.

2.3. Music sentiment classification based on joint representation structure

This section intends to realize the emotional classification of music. In this section, combining the learning features, a two-channel emotion recognition model based on joint representation structure is proposed. The features are connected through the hidden layer and are projected into the same space. Compared with the traditional way, the discrimination of emotion-related features is enhanced.

As one of the channels, the HSF calculates the statistics of the LLD to represent the global dynamics of the LLD. This section will focus on the second channel CRNN and the combination of the two channels.

2.3.1 Spectrogram

As an input to the CRNN model, the abscissa of the spectrogram is time, the ordinate is frequency, and the coordinate point value is speech data energy. Since it uses a two-dimensional plane to express three-dimensional information, the magnitude of the energy value is represented.

In this paper, we will use a Hamming window for windowing with a length of 20ms. Each frame will use the fast Fourier transform to calculate the energy value of each frequency, and slide it in steps of 10ms to generate the spectrum of each frame. In the chronological order, all the spectrograms are spliced into a spectrogram corresponding to the speech.

2.3.2 CRNN model

The CRNN network architecture consists of two parts. The first part is a convolution feature extractor that takes a spectrogram as input and then convolves the input image. For pre-segmented speech, CNN can obtain learning characteristics for each segment.

The second part is the LSTM, with each time step corresponding to a segment of the original audio input, without the need to clip or fill the audio. LSTM can preserve long-term dependencies between segments. The output statistics are calculated by the maximum pooling layer, the minimum pooling layer, and the average pooling layer, and the obtained pooling vectors are connected into one.

2.3.3 Joint representation of HSF and CRNN

The design process represented by HSF and CRNN is shown in Figure 4. The entire process is trained in an end-to-end manner while processing a given speech in two parallel channels.

In one channel, these segments are input to the CRNN network based on the spectrogram segmentation. The output size of the CRNN is a 384-dimensional vector, which in turn is added to an attention layer and a 128-dimensional hidden layer. In another channel, the original waveform is segmented into frames. The LLD is extracted therefrom to calculate the HSF. Then an LSTM layer, an attention layer and a hidden layer are added, the high-dimensional HSF is continuously mapped to the low-dimensional feature space, and finally the feature vector is output.
The two feature vectors output from the two channels are joined as the input to the next hidden layer, which projects them into the feature space and passes them to the Softmax layer for classification.

In this paper, the output is the emotion category to which the music belongs. The Softmax function is used to map the input into the [0,1] space, and the normalization operation is performed to ensure that the sum is 1, and the maximum probability value is used as the music emotion category. According to the classification idea of valence/activation, the identifiable emotion categories are divided into the following four items: sad, happy, anger and neutral.

### 3. Experiment and analysis of results

In order to ensure the uniqueness of the downloaded music, the audio title, singer, author, duration, album and other information will also be stored while the audio file is being acquired. When the retrieved new music is exactly the same as the information in the database, it is determined that the music already exists and does not need to be acquired again.

A spectrum segment has a size of 50 frames and an overlap of 25 frames. The segment length is 515 milliseconds. The CRNN infrastructure consists of two convolutional layers, the first of which has 30 filters of size 40x5. The second layer has 30 filters of size 1 x 3. In the HSF channel, the Softmax has 4 nodes, and the previous hidden layer has 32 nodes. All hidden layers use the sigmoid activation and use cross entropy as the objective function for training.

The data categories of the training set and the test set are relatively balanced. The 5-fold cross-test is used, and the test set is divided into 5 parts, one of which is used as a test sample and the remaining 4 are used as a training set.

This paper uses the Tensorflow framework to build the network model structure to achieve emotional classification of music. The CRNN model is used as a baseline to compare with Model 1: HSF + Standard CRNN and Model 2: Current System (HSF + CRNN (with Attention Mechanism)). The accuracy of emotion recognition is calculated separately for each model.

Table 1 shows the accuracy of different musical sentiment classifications after experimental verification.

| Model                                      | average accuracy |
|--------------------------------------------|------------------|
| baseline: CRNN                             | 64%              |
| Model 1: HSF + Standard CRNN               | 68.5%            |
| Current System (HSF + CRNN (with Attention Mechanism)) | 72.6%            |

It can be seen from Table 1 that the proposed strategy has the best classification in all models and has the best accuracy. It can be determined that the combined attention mechanism and diversified features have a greater impact on the model.
Table 2 is the result of the confusion matrix for the specific sentiment categories in the current system. According to the results, the recognition accuracy is higher, such as happy, anger, etc., and conversely, for the categories with low arousal such as neutral and sad, the recognition accuracy is low.

| Recognition rate | sad  | happy | anger | neutral |
|------------------|------|-------|-------|---------|
| sad              | 71.5%| 5%    | 5.5%  | 18%     |
| happy            | 6%   | 75.2% | 6.8%  | 12%     |
| anger            | 10.3%| 8.9%  | 74.3% | 6.5%    |
| neutral          | 12.2%| 10.9% | 13.4% | 63.5%   |

4. Conclusion

For the problems of low classification accuracy and low recommendation accuracy of music themes, data collection, processing, transfer, and emotional analysis of emotions were designed. Among them, the fusion of two channels is designed in the music sentiment analysis. The low-level descriptors and the spectral maps are used as input to construct the joint representation of the features. The experimental results show that the proposed model has higher classification accuracy.

In the future, music emotions and theme classification will continue, a music emotion classification database will be set up, and corresponding interfaces will be provided. At the same time, model optimization will be further completed for the existing neural network to improve the classification accuracy.

Acknowledgments

This paper is funded by the Natural Science Foundation of Liaoning Province, "Research on Emotional Analysis and Evaluation Model of Speech Reading Based on Machine Learning".

References

[1] Velarde G, Chacón C C, Meredith D, et al. Convolution-based classification of audio and symbolic representations of music[J]. Journal of New Music Research, 2018:1-15.
[2] Lakomkin E, Zamani M A, Weber C, et al. EmoRL: Continuous Acoustic Emotion Classification using Deep Reinforcement Learning[J]. 2018.
[3] He H, Xia R. Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification[J]. 2018.
[4] Rajanna A R , Aryafar K. Deep Neural Networks: A Case Study for Music Genre Classification[C]/ IEEE International Conference on Machine Learning & Applications. IEEE, 2015.
[5] Trabelsi I , Ayed D B . On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels[J]. 2014.
[6] Li T , Ogihara M , Li Q . A Comparative Study on Content-Based Music Genre Classification[C]/ International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2003.
[7] Lee K K , Park K S . Robust Feature Extraction for Automatic Classification of Korean Traditional Music in Digital Library[C]/ International Conference on Asian Digital Libraries: Implementing Strategies & Sharing Experiences. Springer-Verlag, 2005.
[8] Du W , Lin H , Sun J , et al. A new hierarchical method for music genre classification[C]/ International Congress on Image & Signal Processing. IEEE, 2017.
[9] Deb S, Dundapat S. Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification.[J]. IEEE Transactions on Cybernetics, 2018, PP(99):1-14.
[10] Huang Y S , Chou S Y , Yang Y H . Pop Music Highlighter: Marking the Emotion Keypoints[J]. 2018.