Emotion recognition of musical instruments based on convolution long short time memory depth neural network

Jing Wang¹*, Qingqing Wang ² and Hongyan Liu³

¹ Changchun College of Electronic Technology, Changchun 130000, China
² Jilin Animation Institute, Changchun 130012, China
³ Network management center of China Mobile Communication Group Jilin Co., Ltd, Changchun 130012, China

*Corresponding author. Email: 8624949@qq.com

Abstract. In this paper, a method of emotion recognition for musical instruments based on convolution long short time memory depth neural network is proposed, and an emotion recognition music database composed of four musical instruments is established, including keyboard instruments, wind instruments, string instruments and percussion instruments. The emotional types of these four instruments are divided into happiness, anger, sadness and fear. Through the establishment of CLDNN model based musical instrument emotion recognition architecture, MFCCs, CNN and CFS are used for feature extraction and training. The experimental structure shows that the best classification effect is to use long-term memory (LSTM) and deep neural network (DNN) to extract and combine the emotional feature sets of musical instruments, which has the highest accuracy. Considering the dynamic changes of musical features in musical instruments, the modeling method is used to predict the emotional changes of musical instruments.

1. Introduction

Music emotion recognition is a research field of artificial intelligence and pattern recognition. Music can effectively convey people's emotions and help people relax. Through music, a variety of emotions can be identified, such as happiness, sadness and anger. At the same time, music emotion recognition has many applications, such as music information retrieval, automatic generation of music play list, music recommendation system, music therapy and health care. In the field of music therapy and health care, music therapy can effectively identify the emotion of autistic children [1], and can also reduce the anxiety and pain of severe patients. Emotional music database can have two kinds of data, one is singing music signal, the other is musical instrument signal. At present, a large number of existing researches are basically on the emotion recognition of singing music, but there are few researches on the emotion recognition of musical instruments [2]. Therefore, this paper focuses on the clue of emotion recognition in musical instruments, and the difficulty is to extract the instrumental timbre effectively. At the same time, most of the previous music emotion recognition in terms of emotion is to identify the main emotion in music [3]. In this study, the study of fear is added to expand the scope of emotion recognition. An architecture based on CLDNN is proposed, and MFFCs and log Mel are used as the energy input of the filter [4]. This architecture has been used in speech recognition before. In this study, a technique composed of depth convolution layer is proposed for emotion recognition of musical instruments.
The structure of this paper is as follows: Section 2 describes the database of musical instrument emotion recognition, section 3 describes the architecture of musical instrument emotion recognition, section 4 describes how to carry out the experiment, section 5 introduces the experimental results and discussion, and section 6 summarizes the future work.

2. Database
Database in this study, a new emotional music database composed of different musical instruments is established. This database contains 276 pieces of music samples composed of various musical instrument solo excerpts, including keyboard instruments, wind instruments, string instruments and percussion instruments. Each piece of music is 40 seconds long, representing the climax of the whole music. There are four kinds of emotions including happiness, anger, sadness and fear. Music samples corresponding to four kinds of musical instruments are collected for each emotion. Table 1 shows the types of musical instruments used in this study and the related emotions.

| Types of musical instruments | Musical Instruments | Emotions |
|-----------------------------|---------------------|----------|
| keyboard instrument         | Piano               | Anger    |
| Wind instrument             | Flute               | Fear     |
| String instrument           | Violin              | Sad      |
| Percussion instruments      | Xylophone            | Happy    |

3. Musical instrument emotion recognition based on CLDNN
CLDNN includes CNN, LSTM and DNN, in which CNN is used as the output of feature vector, LSTM and DNN are used as classifiers, and the instrument emotion recognition system is completed by combining CNN, LSTM and DNN into a unified architecture. Figure 1 is the architecture diagram of CLDNN.

3.1. CNN feature extraction
CNN is a feedforward neural network composed of multiple convolution layers. Each convolution layer has many filters to generate new eigenvalues. In musical instrument emotion recognition, CNN takes log Mel filter energy and MFCCs features as input and output. A sample is generated every 20 seconds, and MFCCs features and log Mel energy are calculated every 20 seconds. Therefore, 552 CNN based MFCCs features and 552 CNN based log Mel filter energy features are obtained.

3.2. MFCCs
MFCCs are the most commonly used feature vectors in speech recognition and music recognition, and the most prominent performance is in the aspect of speech emotion recognition. However, there are few researches in the aspect of musical instrument emotion recognition. The feature vector of MFCCs is very close to the feature value of human ear perception. When calculating the feature value of MFCCs, the music clips of each instrument are segmented, every 30ms is a frame, and feature extraction is carried out from each frame. The extracted content includes derivative coefficient, acceleration coefficient and static coefficient, as shown in Figure 2, which shows four different types of instruments displayed by MFCCs feature vector Images of related emotions.
3.3. Feature selection

In this study, Correlation based feature selection (CFS) is used to select features. CFS is the most effective method to reduce the feature size, and the method is to select the most characteristic of all features. CFS is an algorithm, which can use the objective function to evaluate each feature subset and classify, so as to find the relevant feature subset function. The formula (1) is as follows:

$$ F = \frac{xk_{clf}}{x+x(x-1)k_{ff}} $$  \hspace{1cm} (1)

Where \( f \) is the sub feature space, \( X \) is the number of features in the sub feature space \( f \), \( k_{clf} \) represents the correlation of average class features, \( k_{ff} \) represents the correlation between average features.

**Figure 2.** MFCC features.
4. Experimental process
Convolutional neural network is introduced to process the data of sequence and time series, and a bidirectional LSTM with long-term and short-term memory is proposed, which can effectively process the forward and backward information flow in the network, and improve the recall ability of nodes by using DNN.

4.1. Musical instrument emotion recognition
The deep neural network uses MFCCs and CFS for feature selection training, and each instrument establishes four emotion models, including ZCR, attenuation, spectral centroid and bandwidth. After training, the instrument emotion model is tested, and the recognition rate is evaluated, compared and analyzed. The feature matrix of emotion recognition for four musical instruments is shown in Figure 3.

![Figure 3. Feature matrix for emotion recognition of four musical instruments.](image)

From the data in the figure, we can see that piano is the best at identifying angry emotions, flute is the best at identifying fear emotions, violin is the best at identifying sad emotions, and xylophone is the best at identifying happy emotions. Therefore, the control instruments and emotions can be divided into angry piano, afraid flute, sad violin and happy xylophone. Next, we use the features extracted from the training data set to evaluate the instrument emotion model, as shown in Table 2.

![Table 2. Evaluation of musical instrument emotion model.](table)

5. Experimental results and discussion
This study discusses the task of deep learning technology and neural network in the emotional recognition of musical instruments. Musical instruments are divided into different types. Based on convolution neural network and traditional machine learning algorithm, four different emotions are expressed by four kinds of musical instruments. The experimental results show that the deep neural network has certain advantages in the emotional recognition of musical instruments, and different kinds of musical instruments have their own advantages in different emotions. Violin is more accurate in expressing sadness, piano is more accurate in expressing anger, flute is more accurate in expressing fear, and xylophone is more accurate in expressing joy.
6. Conclusion
In this paper, a method of emotion recognition for musical instruments based on convolution long-term and short-term memory depth neural network is proposed. The music samples of four kinds of musical instruments are included in the database, including keyboard instrument, wind instrument, string instrument and percussion instrument. The music samples of these four kinds of musical instruments express four common emotions. Based on CLDNN model, the emotion recognition architecture of musical instruments is used to extract the emotion features of musical instruments. Four emotion models are established, which are angry piano, scared flute, sad violin and happy xylophone. The experimental structure shows that MFCCs and CFS are combined for feature extraction training, and the recognition rate is 88.6%. In the future work, we will use deep learning algorithm to continue to study the emotion recognition of ensemble and ensemble instruments, in order to get more rich emotion categories.

References
[1] Li Gaoling. Research on algorithm composition and emotion recognition based on neural network [D]. Zhengzhou University, 2019.
[2] Sun Xiaohu, Li Hongjun. Overview of speech emotion recognition [J]. Computer engineering and application, 2020,56 (11): 1-9.
[3] Liu Jingjing, Wu Xiaofeng. Multimodal emotion recognition and spatial annotation based on long-term and short-term memory networks [J]. Journal of Fudan University (NATURAL SCIENCE EDITION), 2020,59 (05): 565-574.
[4] Zhong Qi, Feng Yaqin, Wang Wei. A comparative study of speech emotion recognition based on cross language corpus [J]. Journal of Nanjing University (NATURAL SCIENCE), 2019,55 (05): 765-773.