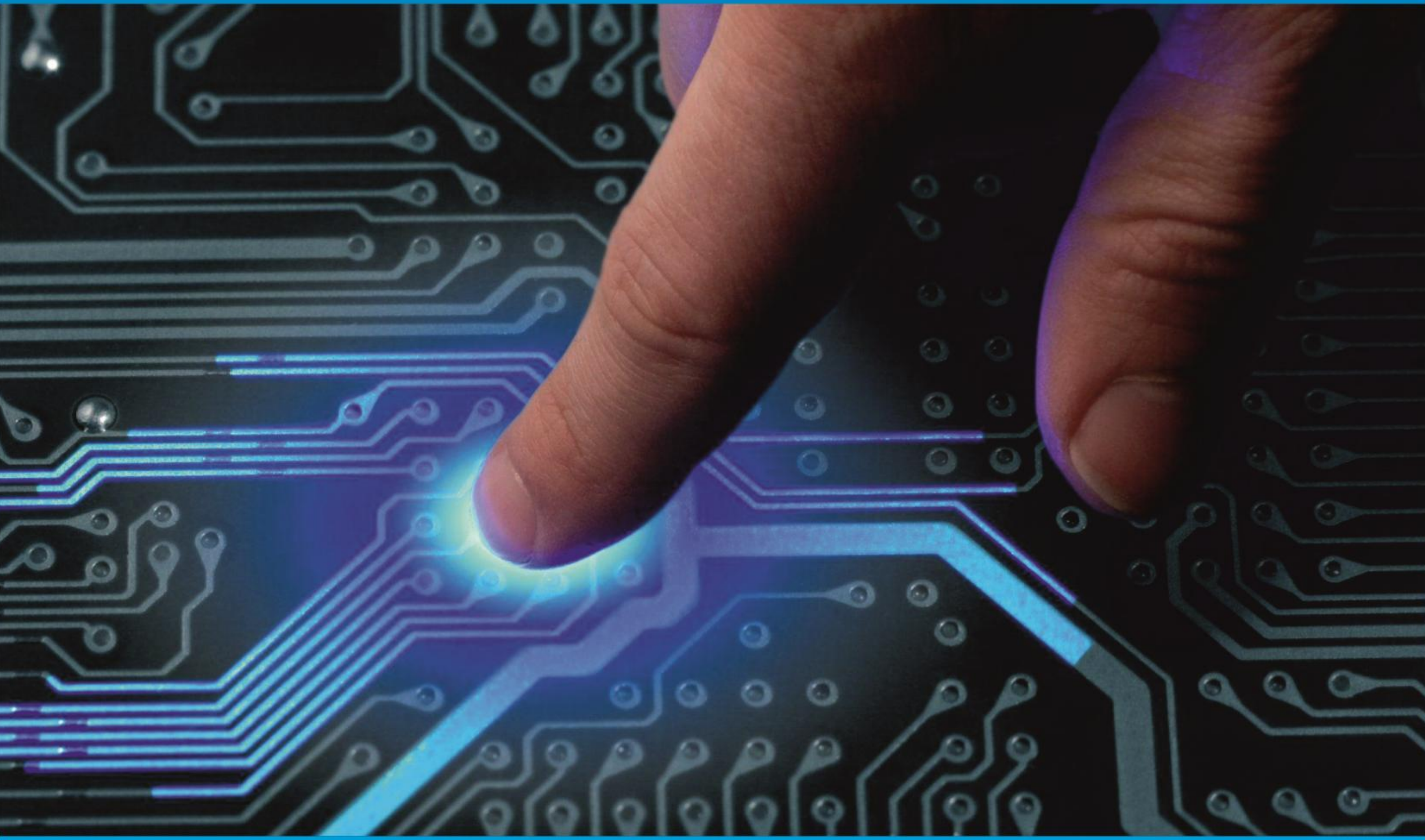
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| DOI: 10.15680/IJIRCCE.2022.1006131 |

Music Recommendation Model Based on Speech Emotion

Harshita Saxena[1], Mihir Aryan[2], Anurag kumar Singh[3], Arjita Agarwal[4] , Mr. Vitesh Babu[5]

BE Student, Department of Information Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[1]

BE Student, Department of Information Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[2]

BE Student, Department of Information Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[3]

BE Student, Department of Information Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[4]

Assistant Professor, Department of Information Science, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[5]

ABSTRACT: Working on the process of music emotion classification, a music emotion recognition system based on the convolutional neural network is proposed. We start with, the Mel-frequency cepstral coefficient (MFCC), and residual phase (RP) are weighted and combined to extract features from the low-level audio, which improves the accuracy of data mining. We use speech and audio to determine the emotion of a person. We make use of the vocal characteristics of the user for the task. Which makes it a text-independent system of speech recognition. The goal of this paper is to create a smart music player based on the emotion and vocal characteristics of the user that recommends and creates music related to that emotional state. The major focus of the work consists of identifying the emotion of a given user and recommending songs that are related to the mood of the user, also advising the songs that the user should listen to elate their mood. We take a live audio sample of the user as input and analyze it to predict the emotional state. The mod is linked to a few related songs which are listed in the console for the user to select from.

KEYWORDS: Convolutional Neural Network, Librosa, Mel-Frequency Cepstral Coefficient, Speech Emotion, Music classification, Music Recommendation

I. INTRODUCTION

With the advanced development of the computer network and multimedia technology, more and more multimedia data such as text, images, audio, and video emerge in the internet. Handling and Analytics of media data have also become a important issue. As an important part of multimedia data, music shows tremendous growth in both quantity and type. Music emotion recognition (MER) has significant research value in the fields of music database management, music retrieval, music recommendation, and music therapy, and it has attracted extensive attention of experts. As an important emotional carrier, music is filled with rich emotional information. Emotional words are the most used words in retrieving and describing music. Therefore, classifying music with emotions a label or category of emotions can really improve the efficiency of music retrieval and has slowly became a research trend. At present, the research on music emotion recognition is mainly divided into two aspects. The first is how to better extract and find the emotional features of music. The precise extraction of music emotional features is directly related to the accuracy and efficiency of classification results second is how to improve the classifier performance of emotion recognition. Because MER involves both music psychology and computer science, it is somewhat difficult to implement music emotion classification.

1. Song Concept

In a song, it can be analyzed again to find out the melody of the music. The melody can be recognized from the way a person mutters or whistles. From the melody, we can find about the genre of a song.

Here are the characteristics of the song:

1. Sound

The voice in the song describes how the sound is notated or written. Sound waves are commonly identified in terms of frequency. The sound fundamental components are described in terms of pitch, duration, intensity, and timbre.

2. Tone

The sound is divided into tones that have a specific pitch or tuning. The difference in tuning between two notes is called an interval. The tone can be arranged on different scales. The root tone of a song tells us about the frequency of each letter in the song.

3. Melody

In its most simple way, the melody is a sequence of notes and duration of notes. Another definition, in another reason, the term includes a succession of other musical elements such as tonal color, a series of letters in time. The circuit can sound alone.

4. Notation

Musical notation is a written representation of music. In block notation, the pitch is represented vertically while the time (rhythm) is represented horizontally. These two elements make up the tone stick.

5. Rhythm

The time signature indicates the number of beats in the measurement and notes counting and it counts as one beat. Specific notes can be accentuated by applying pressure.

2. *Mood*

The mood is a form of emotional state. Mood is different from simple emotions that has less specific and less intense feeling and tend not to be triggered by certain reaction or events. A person's mood can last only a few hours. Moods can be affected by many unexpected event.

Moods are different from emotions because emotions do not have to be triggered by anything. A person feels happy after getting married or receiving a gift, while a happy mood tends to react to external stimuli. A mood last longer than feelings. Personal characteristics to a person affect mood such as optimism and neuroticism influence certain types of moods. Long-term mood disorders, such as clinical depression and bipolar disorder, are mood disorders. Mood can change in both internal and subjective circumstances but can often be inferred from other postures and behaviors.

3. *Speech Emotion Recognition*

Speech emotion recognition, one of the latest processing challenges, is a technology for identifying emotions from a human speech. Apart from facial expressions, speech has also proven to be one of the most promising modalities for the automatic recognition of human emotions. In the field of security and monitoring systems, particularly, the development can be observed over the past year. In realistic view, the classification is considered as the technical approaches that only rely on pragmatic decisions about the type, the extent, and the number of emotions according to the situation.

There is a study that sound often reflects the emotion through tone and pitch. To implement a speech emotion recognition system, we need to define and model emotions carefully. However, there is no consensus on the definition of emotion, and it is still an open problem in psychology. The discrete emotional and dimensional emotional models are two models have become common in introducing speech emotions.

Microphone dramatically affects the quality of the analyzed sound. Using multiple microphones is more accurate because information about the speaker's possible location from a multi-microphone signal usually helps separate speech. One of the crucial points in emotion recognition is what features are used. In recent research, many standard features were extracted, such as energy, pitch, formant, and some spectral features such as linear prediction coefficient (LPC), Mel-frequency cestrum coefficient (MFCC), and spectral modulation features.

II. ALGORITHM MODEL

1. *Overall Framework.*

After taking audio low-level features from audio signals, machine learning method is used for speech emotion classification. For now, the performance of the speech emotion recognition system based on audio meets the “ceiling,” which is because the low-level features of audio lack emotional relevance, and the high-level features are closer to emotion, and the manual design cost is high. By using the deep learning method to audio emotion classification helps to bridge the semantic gap between audio low-level features and music high-level emotion concepts. The overall architecture of the model is shown in Figure 1. Compared with the low-level features of audio, the spectrogram contains more audio information. Therefore, the model combines the spectrogram and the low-level features of audio as the input sequence to implement information complementarity.

The input of the CNN is sound spectrogram, in which the CNN uses two different convolution kernels to extract the time-domain and frequency-domain features of audio.

2. Audio Low-Level Feature Extraction

1. *Mel-Frequency Cepstral Coefficient (MFCC). For now*, content-based acoustic features are mainly divided into timbre, rhythm, pitch, harmony, and time features. Timbre features encompass cestrum features, such as MFCC; rhythm features mainly point to beat number and rhythm histogram; pitch features are important frequency information; harmony features are chromaticity diagram; time characteristics are time centroid. MFCC uses auditory principle and the decorrelation characteristics of cestrum.

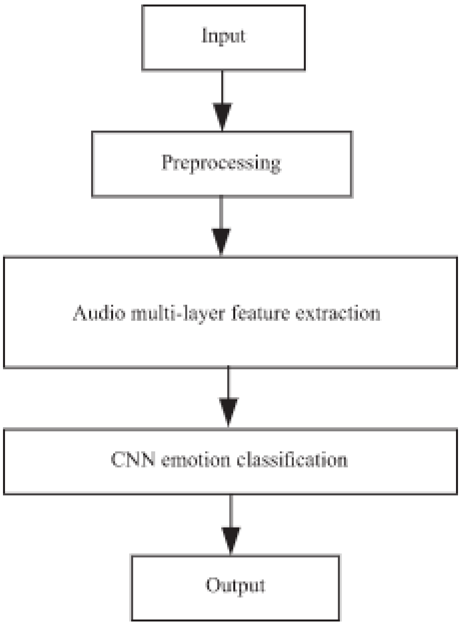

Input
Preprocessing
Audio multi-layer feature extraction
CNN emotion classification
Output

Figure 1: Overall framework of the proposed method.

To extract MFCC features, firstly, the audio signal is preprocessed and windowed. The Blackman–Harris window is used to divide the original signal with sampling rate of 44.1 kHz into 2048 sample frames. After windowing the audio, the two ends of each frame signal will slowly become 0, so the two ends of the signal will be weaker. To overcome this issue, adjacent frames can be made to overlap during frame division. The overlap length is half of the frame length or fixed as 10 ms. In this method, the adjacent frames overlap by 50%, which can not only reduce the spectrum leakage but also cut the unnecessary workload.

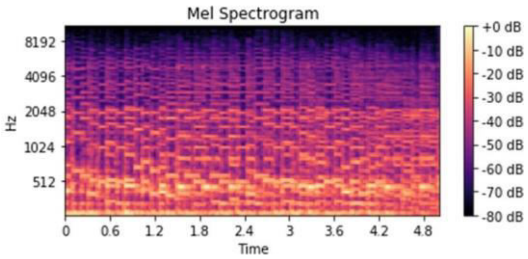


Figure 2: Mel-Spectrogram

3. CNN Emotion Classification

Audio is an important component of music. Most researchers analyze music emotion by audio, generally extract time-domain and frequency-domain features from audio and classify music emotion using machine learning algorithms like *K*-nearest neighbor, SVM, and Gaussian mixture model. Using the existing machine learning methods of frequency-

domain and time-domain features makes it difficult to improve the performance of music emotion recognition. Because the manually extracted features usually belong to the low-level features of audio,
and these are not related to the audio emotion state and contain no significant information. The high-level features of music are closer to emotion, and the amount of information is significant. But the cost of manual feature extraction is higher, and professional knowledge is needed. Few researchers apply the deep learning method to music emotion classification, and the results are obviously better than the traditional machine learning method. Therefore, using the deep learning method can extract more accurate emotion-related music feature.

III. EXPERIMENTS AND ANALYSIS

The experiments use the emotion dataset to test and analyze the performance of the proposed method in emotion recognition. The dataset has 2906 songs, which contain 4 emotional classes, including 639 anger songs, 753 happy songs, 750 relaxing songs, and 764 sad songs. For the ease and consistency of the experiment, only
have good identification. With the huge amount of data, the improvement effect is restricted to certain limitations.
The idea of center SoftMax is to reduce the intraclass spacing and control the feature center by introducing center loss.
the first 30 s of each song is used, and those less than 30 s are filled with zeros. The dataset is divided into three parts in the ratio of 7 : 2 : 1, which are training, verification, and test sets, so as to maximize and bring in accurate results of the experiment.

1. Data Set.
We focused our work on the benchmark RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) for this speech emotion recognition. You can find details about the dataset on its Kaggle page: RAVDESS Emotional speech audio | Kaggle.

The data contains 3-second audio clips spoken of the same two sentences by 24 different actors over an emotional range of 7 emotions. In addition to that 12 male and 12 female actors give the data a more diverse and challenging range. Making total of 1440 samples.

Upon downloading the data, we can see a particular naming style for the audio files. This nomenclature is detailed on the Kaggle page as follows.

- Modality (01: full-AV, 02: video-only, 03: audio- only).
- Vocal channel (01: speech, 02: song).
- Emotion (01: neutral, 02: calm, 03: happy, 04: sad, 05: angry, 06: fearful, 07: disgust, 08: surprised).
- Emotional intensity (01: normal, 02: strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01: "Kids are talking by the door", 02: "Dogs are sitting by the door").
- Repetition (01: 1st repetition, 02: 2nd repetition).
- Actor (01 to 24. Odd-numbered actors are male, even-numbered actors are female).

For this experiment, we have considered only 5 of those emotions - happy, sad, calm, angry and neutral. We used a list of 900 songs with emotion tags for all the five emotions used in our project to recommend songs to the user fitting to mood.
2. Data Preprocessing.
Mel spectrum is a common signal representation method in audio classification tasks. Mel spectrum retains the features of music signals more accurately and also, Mel spectrum is more in line with human auditory characteristics. Therefore, Mel spectrum is chosen as the input data of music audio analysis.
Voice activity detection (VAD) is to detect whether there is a mute frame in the music signal. These parts of the mute frame hinders the recognition result. The flow of music audio signal representation and preprocessing is given below Figure 3.

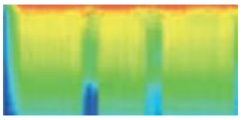


Figure 3: Preprocessing flow

3. *CNN Model*

The data set is divided into two parts one is training sample and other is testing sample. The training sample and testing sample both go through preprocessing of data and feature extraction. While training the model emotion feature vectors are extracted from training sample and given to the classifier. So, after evaluating the model the testing sample is fed through the classifier and use that emotion feature vectors and output the result.

4. *Audio Time Period Selection.*

Different researchers takes different lengths of audio clips. So the experimental comparison will be carried out for different time periods, to choose the most suitable audio time period and carry out further experiments on this basis. Experiments are done on the CNN model and the improved CNN model. Audio clips of 3 s, 5 s, 10 s, and 16 s are taken for comparison from the emotion dataset. The relationship between time period and number of samples is shown in Table 1.

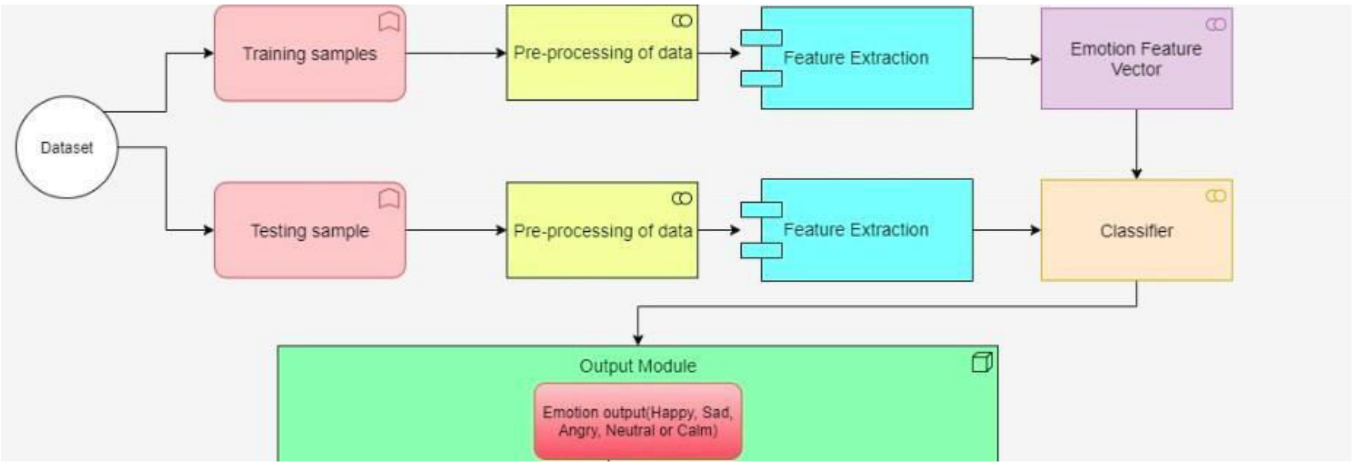


Figure 4: Preprocessing flow.

5. *Emotion Classification Confusion Matrix*

In order to carry out the classification of the music emotion state by the improved CNN model and gauge the performance of the classification model, the emotion 4-class confusion matrices of the original CNN model and the improved CNN model are calculated and tabulated, as shown in Table 4.

Table 1: Relationship between time period length and sample number.

Length of the time period (s)	3	5	10	16
Number of samples	23000	14000	7000	3000

Table 2: Classification accuracy in different time periods.

Length of the time period (s)	3	5	10	16
CNN model (%)	69.19	75.58	80.36	83.71

Table 3: Experimental results using convolutional networks.

Model	Recognition accuracy (%)	
	Training set	Test set
CNN	92.65	81.08

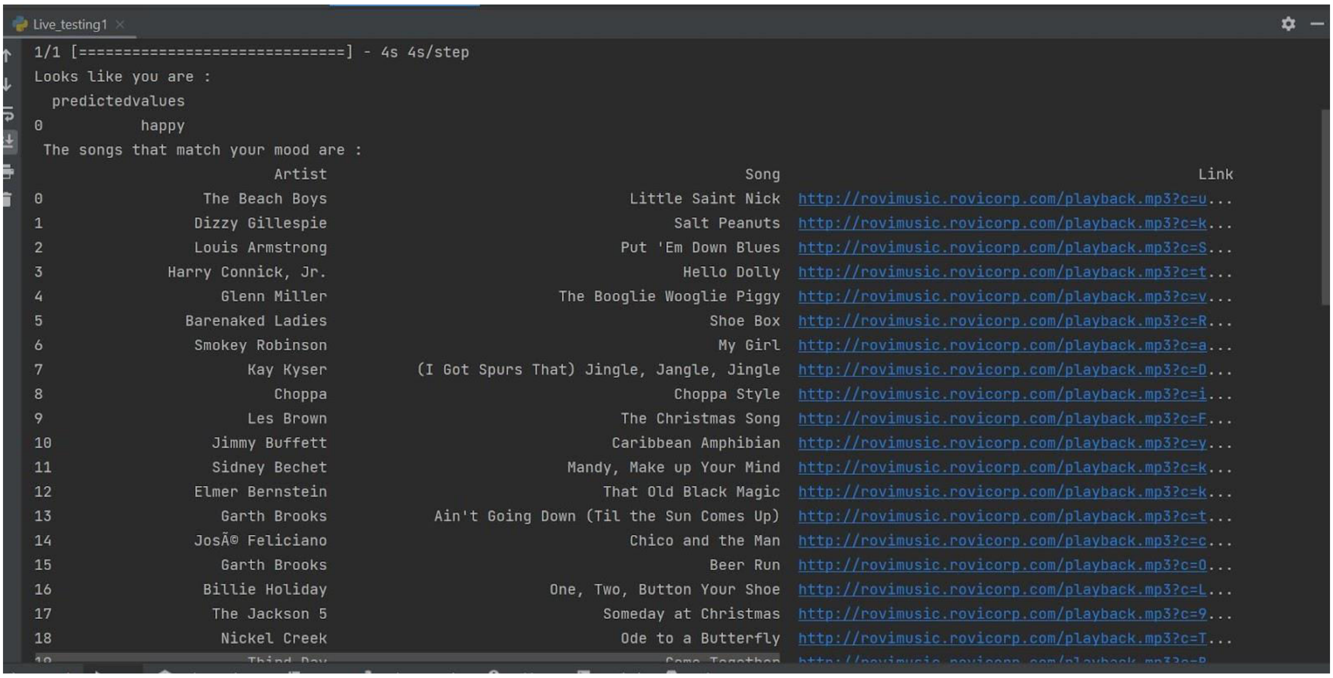
Table 4 : Confusion matrix of emotion classification.

Model	Emotion	Sad (%)	Happy (%)	Anger (%)	Quiet (%)
CNN	Sad	86.29	6.13	4.57	3.01
	Happy	7.05	76.43	9.28	7.24
	Anger	6.64	8.45	74.06	10.85
	Quiet	5.86	4.18	9.67	80.29

6. *Implementation*

The recommender system work calculating cosine similarity of extraction features (equation 1) from one music to another music. The extracted features are in vector form; therefore, it is possible to calculate their distance. First, we take one music for each genre as the basis for the recommender system. Next the prediction of the basis music genre is calculated by help of neural networks. The feature vectors that produce before the classification layer are taken into consideration for recommendations.

Our music recommender application is in Python programming language and using key libraries, that is: Tensorflow, Keras, Librosa. To get recommendations from the music that is playing, user can click the corresponding three-button dot button to the right of the music title and "Get relevant" button. The output of recommendations are given below Figure 6.



```
1/1 [==============================] - 4s 4s/step
Looks like you are :
  predictedvalues
0          happy
 The songs that match your mood are :
                           Artist                                            Song                                                 Link
0                  The Beach Boys                                Little Saint Nick  http://rovimusic.rovicorp.com/playback.mp3?c=u...
1                 Dizzy Gillespie                                     Salt Peanuts  http://rovimusic.rovicorp.com/playback.mp3?c=k...
2                 Louis Armstrong                               Put 'Em Down Blues  http://rovimusic.rovicorp.com/playback.mp3?c=S...
3              Harry Connick, Jr.                                      Hello Dolly  http://rovimusic.rovicorp.com/playback.mp3?c=t...
4                    Glenn Miller                       The Booglie Wooglie Piggy  http://rovimusic.rovicorp.com/playback.mp3?c=v...
5                Barenaked Ladies                                         Shoe Box  http://rovimusic.rovicorp.com/playback.mp3?c=R...
6                 Smokey Robinson                                          My Girl  http://rovimusic.rovicorp.com/playback.mp3?c=a...
7                       Kay Kyser          (I Got Spurs That) Jingle, Jangle, Jingle  http://rovimusic.rovicorp.com/playback.mp3?c=D...
8                          Choppa                                     Choppa Style  http://rovimusic.rovicorp.com/playback.mp3?c=i...
9                       Les Brown                               The Christmas Song  http://rovimusic.rovicorp.com/playback.mp3?c=F...
10                  Jimmy Buffett                               Caribbean Amphibian  http://rovimusic.rovicorp.com/playback.mp3?c=y...
11                  Sidney Bechet                            Mandy, Make up Your Mind  http://rovimusic.rovicorp.com/playback.mp3?c=k...
12                 Elmer Bernstein                              That Old Black Magic  http://rovimusic.rovicorp.com/playback.mp3?c=k...
13                   Garth Brooks            Ain't Going Down (Til the Sun Comes Up)  http://rovimusic.rovicorp.com/playback.mp3?c=t...
14                 JosÃ© Feliciano                                Chico and the Man  http://rovimusic.rovicorp.com/playback.mp3?c=c...
15                   Garth Brooks                                         Beer Run  http://rovimusic.rovicorp.com/playback.mp3?c=0...
16                 Billie Holiday                          One, Two, Button Your Shoe  http://rovimusic.rovicorp.com/playback.mp3?c=L...
17                  The Jackson 5                              Someday at Christmas  http://rovimusic.rovicorp.com/playback.mp3?c=9...
18                   Nickel Creek                                Ode to a Butterfly  http://rovimusic.rovicorp.com/playback.mp3?c=T...
19                      Third Day                                    Come Together  http://rovimusic.rovicorp.com/playback.mp3?c=B
```

Figure 6: Music Recommendation Application

IV. CONCLUSION

With the development of artificial intelligence and digital audio technology, music information retrieval has gradually become a research hotspot. Music emotion recognition has great research value for video music, but there is little research on it at present. Therefore, a music emotion recognition method using the convolutional neural network is proposed. The audio time- domain features, frequency-domain features, and sequence features extracted by the CNN and the audio sequence context information of the Bi-LSTM extractor are fused and sent to the Emotion feature vector and then to Emotion Classification Confusion Matrix for analysis to identify four music emotion types.

The experimental results based on the emotion music dataset show the following:

1. The audio time period, the number of iterations, and the convolutional network structure will have a certain effect on the recognition model. When the time period of 10 s is selected, the number of iterations is set to 7000, and the lightweight network is adopted, the recognition performance of the model has been significantly improved.

2. The proposed CNN model recognition accuracy is as high as 81%, for test set in Table 3.

3. At present, deep learning method has gradually become the mainstream method of text emotion classification. Therefore, the deep learning method is considered to fully extract the emotional information of songs.

REFERENCES

1. K. Zhao, S. Li, J. Cai, H. Wang and J. Wang, "An Emotional Symbolic Music Generation System based on LSTM Networks," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC),Chengdu,China,2019
2. S. Lukose and S. S. Upadhya, "Music player based on emotion recognition of voice signals," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, 2017
3. P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018
4. Livingstone SR, Russo FA (2018) The Ryerson Audio- Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLos ONE 13(5): e0196391.
5. Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset." In International Conference on Learning Representations,2019.
6. Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, Yanhui Tu, Attention Based Fully Convolutional Network for Speech Emotion Recognition , arXiv:1806.01506v2 May 2 2019
7. Saikat Basu ,Jaybrata Chakraborty , Arnab Bag , Md. Aftabuddin, A review on emotion recognition using speech, 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)
8. M. Kattel, A. Nepal, A. K. Shah, D. Shrestha Department of Computer Science and Engineering, School of Engineering Kathmandu University, Nepal : Chroma Feature Extraction
9. Cyril Laurier and Perfecto Herrera, Mood Cloud: ARealTime Music Mood Visualization Tool, Music Technology Group Universitat Pompeu Fabra Ocata 1, 08003 Barcelona, Spain
10. Aayush Bhardwaj; Ankit Gupta ; Pallav Jain ; Asha Rani ; Jyoti Yadav "Classification of human emotions from EEG signals using SVM and LDA Classifiers", 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi® cross ref

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com