EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Music Emotion Recognition Based on Long Short-Term Memory and Forward Neural Network

Aizhen Liu[1,*]

[1]College of Art and Design, Luoyang Vocational College of Science and Technology, Luoyang 471000 China

Abstract

In this paper, we propose a new music emotion recognition method based on long short-term memory and forward neural network. First, Mel Frequency Cepstral Coefficient (MFCC) and Residual Phase (RP) are weighted to extract music emotion features, which improves the recognition efficiency of music emotion features. Meanwhile, in order to improve the classification accuracy of music emotion and shorten the training time of the new model, Long short-term Memory network (LSTM) and forward neural network (FNN) are combined. Using LSTM as the feature mapping node of FNN, a new deep learning network (LSTM-FNN) is proposed for music emotion recognition and classification training. Finally, we conduct the experiments on the emotion data set. The results show that the proposed algorithm achieves higher recognition accuracy than other state-of-the-art complex networks.

Keywords: music emotion recognition; long short-term memory; forward neural network; MFCC; RP.

Received on 13 January 2022, accepted on 19 January 2022, published on 27 January 2022

doi: 10.4108/eai.27-1-2022.173162

*Corresponding author. Email: [illegible]

1. Introduction

Music has always been an indispensable part of human activities. It can not only represent the author to express his/her inner emotional activities, but also make the listener accept the power of music, so as to achieve some positive spiritual guidance [1,2]. In this era of pursuing intelligence, many films, television works and multimedia videos emerge in an endless stream. Music emotion recognition can also perform real-time soundtrack according to the emotion conveyed by voice and video content [3].

At present, the research on musical emotion recognition is mainly divided into two aspects. One is how to better extract the emotion features of music. One is how to improve the classifier effect of emotion recognition. Chen et al. [4] adopted Deep Pitch Class Profile (DPCP) feature based on deep learning in the stage of audio feature extraction to ensure the robustness and generalization ability of audio feature extraction and improve the feature performance of nonlinear deep semantics of music. Weninger et al. [5] input the underlying features of music into the recurrent neural network for training, so as to complete music emotion recognition. Markov et al. [6] used Gaussian Process (GP) and Support Vector Machines (SVM) to research different features, including MFCC, Linear Prediction Coefficient (LPC), timbre features and their various combined features. Then they were used for music style classification and VA (Valency arousal) emotion estimation. It can be seen from their experiments that the classification result of GP method is indeed better than that of SVM method. However, the algorithm complexity of GP method is higher than SVM

method. So it is very difficult to apply in a large scale mission. Chen et al. [7] spliced the features related to rhythm, intensity, timbre, and pitch into 38-dimensional music features, and used the Deep Gaussian Process (DGP) method for music emotion recognition. They built a GP regressor for each emotion category and used regression to classify music emotions. Although this method achieved a good effect on emotion classification, music samples could not be expanded after the model training is completed. Li et al. [8] proposed a method based on Deep Bidirectional Long Short Term Memory (DBLSTM) to dynamically predict music emotion, which trained multiple DBLSTM based on time series of different scales. Then Extreme Learning Machine (ELM) was used to fuse the results of multi-scale DBLSTM to get the final result. Wei Xiang et al. [9] used Convolutional Neural Network (CNN) and its variants in deep learning to automatically extract abstract features of emotion samples, eliminating the process of artificial feature selection and dimension reduction. Sarkar et al. [10] proposed a convolutional neural network built around VGGNet and a novel post-processing technology to improve the performance of music emotion recognition in accordance with the method based on deep learning. Orjesek et al. [11] proposed a deep learning model, which used the feature spectrogram of music signals as the input of music features, and used the combination of convolutional neural network and recurrent neural network to extract features and classify emotions from the spectrogram. Issa et al. [12] introduced a new architecture to extract MFCC, chromatogram, Meyers scale spectrum, Tonnetz representation and spectral contrast features from sound files and then input them into a one-dimensional convolutional neural network. An incremental approach was then used to modify the initial model to improve classification accuracy. Unlike some previous approaches, all models can work directly with raw sound data without having to be converted to a visual representation. Nalini et al. [13] extracted music emotional features by combining MFCC and RP, and applied the Auto-associative Neural Network. The results showed that the recognition results of fusion feature were consistently better than that of single music emotion feature. However, training models in traditional deep learning were time-consuming and inefficient, especially in dynamically increasing the number of samples. Most music algorithms for emotion recognition work in two ways. The first is feature extraction. It tries to extract the emotion feature information contained in the music signal as the model input. The second is classifier design. In order to maximize the accuracy of music emotion recognition and classification, a better learning model is designed.

Although these algorithms have achieved good recognition effect, there are still areas for improvement:

1) there are many kinds of extracted music emotion features, but the flexibility of the algorithm is not enough to adapt to various features.
2) The deep learning network is simple to build, but its internal structure is very complex and the number of hyperparameters is huge, which makes it difficult to modify. It is very difficult to analyze its internal structure theoretically.
3) Emotion is subjective, so it is not easy to grasp how to better extract its music features and which aspects to start with for innovation.

Forward neural network (FNN) provides an alternative to deep learning network, which has simple structure and fast data processing [14]. Tang et al. [15] used the random convolutional neural network to extract features of audio, and then used the FNN to predict labels. Deep learning and FNN were sequentially spliced, effectively improving the classification accuracy and training efficiency of the model. In order to take into account the advantages of deep learning and FNN at the same time, Chen et al. [16] proposed a cascade FNN based on convolutional feature mapping nodes, and the experiment proved that the network greatly exceeded the traditional deep learning network in feature extraction and training efficiency. Serhat et al. [17] proposed an approach for music emotion recognition based on convolutional long short term memory deep neural network architecture. It utilized features obtained by feeding convolutional neural network layers with log-mel filterbank energies and MFCCs in addition to standard acoustic features. Madeline et al. [18] used machine learning techniques to classify which genre of music was being listen to using physiological responses. Both Long Short Term Memory Networks and Convolutional Neural Networks could be used for making predictions from sequence data. It trained and compared two networks which attempted to classify the genre of music a participant was listening to from their electrodermal activity. Benito-Gorron et al. [19] proposed a hybrid convolutional-LSTM model which achieved the better overall results.

In this paper, LSTM and FNN are combined. LSTM is used as the feature mapping node of FNN to build a new Long short-term memory-FNN (LSTM-FNN) to improve the accuracy of music emotion classification. LSTM-FNN uses an incremental learning algorithm to process the training of new nodes without reprocessing all data, which greatly reduces the running time of the model. Firstly, in the stage of music feature extraction, content-based acoustic feature MFCC is used to increase emotion sensitivity. Residual phase and bit are derived from music signals to extract specific music emotion information, and the weighted combination of the two is used as model input. Secondly, LSTM model training is performed on the input data to extract the contextual relationship of music, and the feature node set is generated as the input of

FNN enhanced node. The enhanced layer output is generated through mapping, and the combination of feature node and enhanced node set is used to obtain the final output by global violation. Finally, the trained model is used to predict the types of musical emotions. Experimental results show that the proposed algorithm can extract audio information more effectively by adding music features, and the constructed LSTM-FNN can effectively improve the accuracy and efficiency of music emotion recognition.

This paper is organized as follows. In sections 2, we detailed introduce the proposed the music emotion recognition model. Section 3 gives the experiments to verify the effectiveness of the proposed method. Finally, a conclusion is conducted in section 4.

2. Proposed music emotion recognition model

2.1. Feature extraction

At present, content-based acoustic features are mainly divided into timbre, rhythm, pitch, harmony and temporal characteristics. Timbre features include cepstrum features such as MFCC. The features of rhythm content mainly include cadence number, rhythm histogram and so on. The content features of pitch are mainly frequency information. Harmonic characteristics include chromaticity diagram [20,21]. The time feature includes the center of time mass. Where, MFCC makes use of the principle of hearing and the declination characteristics of cepstrum, which is as one of the most successful spectral features in speech and music related recognition tasks. In order to extract the feature, firstly, the audio signal is preprocessed, and the frame is segmented and windowed. The original signal with a sampling rate of 44.1khz is segmented into frames with 2048 samples by blackman-Harris window. After the audio signal is windowed, the ends of each frame will fade to 0. As a result, both ends of the signal are weakened. In order to overcome this problem, the adjacent frames will overlap in the frame splitting process. Generally, half of the frame length is taken or the frame length is fixed at 10ms. In this paper, adjacent frames overlap by 50%, which can not only reduce spectrum leakage, but also reduce unnecessary workload. Then, the discrete STFT is applied to each frame to obtain the spectral energy, which is weighted by the frequency response of k_1 Mayer filters and further filtered to generate Mayer spectra. Its center frequency and bandwidth roughly match that of an auditory critical band filter. Finally, the whole Merle spectrum sequence is divided into L blocks with the size of k_2 frames, represented as $I_q, q = 1,2,\cdots,L$ along the time axis. Therefore, each block has a size of $k_1 \times k_2$.

RP is defined as the cosine of the phase function of the analytic signal derived from the Linear Predictive (LP) residual of the musical signal. At time t, the music sample s(t) can be estimated as a linear combination of p samples in the past, so the predicted music sample can be expressed as:

$$\hat{s}(t) = \sum_{k=1}^{p} a_k s(t-k) \tag{1}$$

Where p is the sequence of predicted time. $\{a_k\}, k = 1,2,\cdots,p$ is the set of Linear Prediction Coefficients (LPCs). The prediction error e(t) is defined as the difference between the actual value s(t) and the predicted value. The formula is as follows:

$$e(t) = s(t) - \hat{s}(t) = s(t) - \sum_{k=1}^{p} a_k s(t-k) \tag{2}$$

LPCs, the LP residual r(t) of the music signal, are obtained by minimizing the prediction error e(t). Analytic signal $r_a(t)$ can be obtained from r(t):

$$r_a(t) = r(t) + j r_h(t) \tag{3}$$

$$r_h(t) = IFT[R_h(w)] \tag{4}$$

Where $r_h(t)$ is the Hilbert transform of r(t). $R_h(w) = \begin{cases} -jR(w), 0 \le w < \pi \\ jR(w), -\pi \le w < 0 \end{cases}$. $R(w)$ is the Fourier transform of r(t). IFT stands for the inverse Fourier transform. Therefore, the analytic signal $r_a(t)$ can be expressed as:

$$h_e(t) = | r_a(t) | = \sqrt{r^2(t) + r_h^2(t)} \tag{5}$$

A lot of emotion information about music exists in LP residual. Calculating the residual phase can help to extract emotion specific information in music signal. Residual phase is the cosine of the phase of the analytic signal, and the calculation formula is as follows:

$$\cos(\theta(t)) = \frac{R_e(r_a(t))}{| r_a(t) |} = \frac{r(t)}{h_e(t)} \tag{6}$$

Marius et al. [22] had proved that RP contained audio-specific information that was complementary to MFCC features. RP is defined as the cosine of the phase function of the analytic signal derived from the LP residual of the musical signal. The recognition rate in the deep learning model indicates that there is specific emotion information in the music signal, and RP can extract these specific information. Weighted combination of MFCC features and RP features to get the final output can improve the model's ability to extract emotional features contained in

music signals. The flow chart of feature extraction is shown in figure 1.

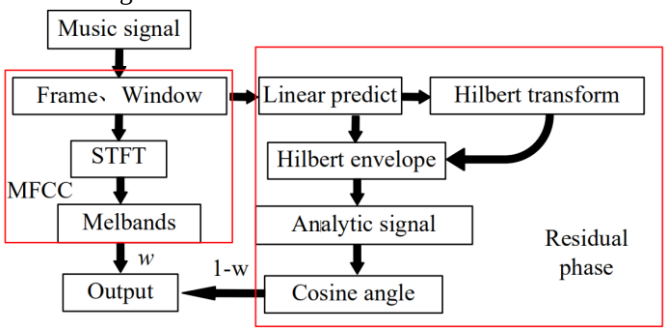

Figure 1. Flow diagram of feature extraction

2.2. Model designing

Forward neural network (FNN) is the basic network architecture of neural network, which is different from recurrent neural network (RNN). No connections are formed between units of the forward neural network. The data is fitted after various nonlinear changes of the input. The classification of music emotion in this experiment can be regarded as a multi-classification problem, which is expressed by the following formula.

$$z = \delta(w^h + b) \tag{7}$$

Where h represents the input of the neural network, z represents the output of the neural network. w and b are the parameter weights of the input and output layer.

The propagation modes of neural network include forward propagation and back propagation. Forward propagation means that the model propagates from the bottom up, performing calculations based on a given input. The loss value is calculated according to the calculation result of forward propagation. Back propagation errors use gradient descent algorithm to calculate and train the parameters of each neuron. The forward propagation formula of the neural network can be represented by the following set of recursive formulas, i.e.,

$$a_k^t = \sum_i w_{ik} x_i^t + \sum_{k'} w_{k'k} b_{k'}^{t-1} \tag{8}$$

$$b_k^t = f(b_k^t) \tag{9}$$

Where, the input of the hidden layer k at time t is marked as a_k^t. The output of the hidden layer of the k-th neuron at time t is marked as b_k^t. The input of the k-th neuron is labeled as x_k^t. w_{ik} represents the parameter weight of the input layer and the hidden layer. Function f calculates the excitation function of the hidden layer.

In this paper, Softmax is used as the activation function for the training of FNN [23]. Softmax has a good performance in the multi-classification problem. Because the output of each neuron of Sofimax is positive and the sum is 1. The output of Softmax layer can be regarded as a probability distribution.

Suppose the output of Softmax is $f(z_j)$. The output is Z. The i-th element in array Z is denoted by Zi, then, the judgment formula of Softmax is:

$$f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_j}} \tag{10}$$

The loss function of Softmax is:

$$loss = -\log f(z_k) \tag{11}$$

Where Z_k is the correct input for the sample. However, since more than one data is input during training, assuming that the input is $f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}}$, so $loss = \log \sum_{i}^{n} e^{z_i} - z_j$. In the case of using the chain rule, the partial derivative of Loss to the weight matrix is obtained, then,

$$\frac{\partial loss}{\partial z_i} = \begin{cases} f(z_i) - 1, z_i = z_j \\ f(z_i), z_i \neq z_j \end{cases} \tag{12}$$

Its main steps include:

1) calculate the activation value of each node in the network;

2) The gradient is propagated through the back propagation algorithm to obtain the gradient value of each parameter;

3) The gradient descent algorithm is used to update the model parameters;

4) Iterate the above process until convergence.

The block diagram of LSTM-FNN is shown in figure 2.

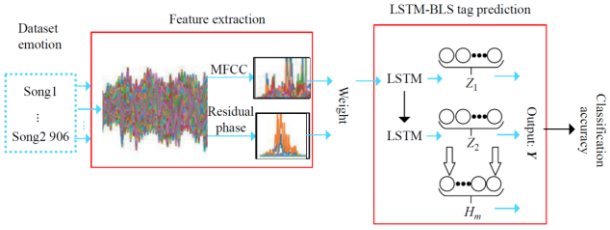

Figure 2. Block diagram of LSTM-FNN model

3. Experiments and analysis

3.1. Data set

This paper uses the Emotion music data set to test and evaluate the performance of complex models of deep learning and forward learning networks in Emotion classification. The dataset consists of 2906 songs in four emotion categories: 639 angry songs, 753 happy songs, 750 relaxed songs and 764 sad songs [24]. For the convenience and tidiness of the experiment, only the first 30s of each song are used, and the zero filling operation is carried out if the song is less than 30s. The data set is randomly divided into three parts in the ratio of 8:1:1, which are training, verification and testing sets respectively to maximize the fairness of the experiment.

3.2. Parameter setting

In order to verify the validity of model classification based on LSTM-FNN learning network, four complex networks are selected for comparison including RCNNBL [15], MCC-3 model [25], MCCLSTM model [16], Cascade of Convolution Feature Mapping Nodes (CCFBLS) [26], and two basic deep Learning networks CNN and LSTM. Among them, the parameter settings of RCNNBL model, MCC-3 model and MCCLSTM model are referred to their respective references, and the parameter settings of LSTM-FNN and CCFBLS are shown in Table 1. The experiment is carried out on Nvidia Tiatitan 1060 GPU with 64GB memory.

Table 1. Parameter setting

Model	Parameter
LSTM-FNN	$k_1=40,\ k_2=80$ OutputDim_L1=400 OutputDim_L2=400 $N_1=10,\ N_2=10,$ $N_3=100$
LSTM	OutputDim_L1=400 OutputDim_L2=200 OutputDim_L3=100
FNN	$N_1=10,\ N_2=10,$ $N_3=500$
CCFBLS	F=3×3 $N_1=10,\ N_2=10,$ $N_3=500$

In the preprocessing stage, MFCC features are extracted with 40 Meir filter banks and 80 frame lengths. LP residuals are derived with 16-order LP. By using the first-order digital filter and 20ms frame size, the overlap between adjacent frames is 50%. LP residual is extracted from the emotion music signal by pre-emphasizing the input music data, and the highest Hilbert envelope of each frame is extracted to generate RP features. The feature sequence diagram is extracted by combining the two features weighted. Feature extraction is carried out for each type of music signal, and the obtained sequence feature diagram is shown in figure 3. Figure 3 shows the timing features of three frames extracted from the audio signals of four emotion types. The input parameters of the network are in the form of Batch size, height, width, channels]. According to the computer memory size and the complexity of the classification model, batch-size is 128, that is, 128 sequence diagrams are input at one time. In LSTM-FNN network, 3-layer LSTM is used for node mapping, and the output dimensions are 400, 200 and 100 respectively. The model structure with the best effect is selected through experiment comparison, and then the output of LSTM is mapped to the enhancement layer. There are four convolutional blocks in CCFBLS network, and each CNN block contains a convolutional layer, a pooling layer, and a dropout layer. The number of filters in the convolutional layer is 64, the shape is fixed at 3×3, the step is 1, the pooling mode is selected as maximum pooling, and the dropout parameter is 0.5, in which the four convolutional outputs are all connected to the output node of CCFBLS.

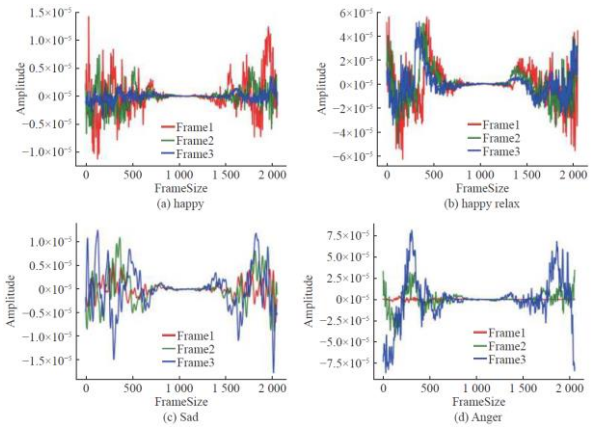

Figure 3. Timing features extracted from four music emotions

3.3. Experiment result

It is uncertain which LSTM model structure combined with FNN can achieve higher accuracy of music emotion classification. In this paper, the LSTM structure experiment is carried out to select the number of LSTM nodes in the mapping layer. The experiment compares the LSTM and FNN combined models of 1-3 layers respectively [27], and tries to find out the influence of the number of LSTM nodes on the classification accuracy of the overall model. The classification results are shown in figure 4. It can be seen that the classification accuracy of the two-layer LSTM model is higher than that of the other two models, and increasing the number of layers does not make the result more excellent but increases the training time [28-31]. Therefore, the output of the two-layer LSTM model is selected as the input of the mapping layer and combined with FNN for music emotion classification training.

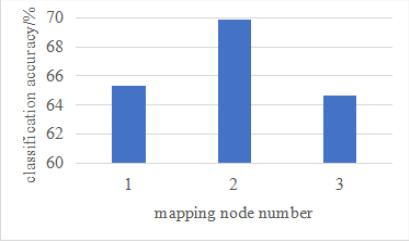

Figure 4. Classification accuracy comparison of different numbers .of LSTM mapping nodes

The accuracy and effectiveness of the proposed classification model are evaluated by comparing the proposed model with the four classification models in [12]. In order to make a fair comparison, each method was cross-validated by 10 times to obtain the classification accuracy, and the classification results are shown in Table 2. The experimental environments are same for the compared methods. Experimental results show that the proposed model is far superior to the model based on deep learning for music emotion classification, and also better than the RCNNBL model. Because the emotional analysis of music is very subjective, it is very difficult to use physical parameters to describe music emotion from the characteristics of audio signals in the recognition of musical emotion [32-34]. Moreover, the current research results are not satisfactory for the classification of music emotion, and only the possible direction can be identified in the slight advantages.

Table 2. Classification accuracy comparison with different models

Model	Accuracy/%
CNN	54.67
LSTM	59.71
MCCLSTM	58.48
MCCBL	58.45
RCNNLSTM	60.36
RCNNBL	61.72
MCCLSTM+FNN	62.11
CCFBLS	64.36
LSTM-FNN	70.25

As can be seen from Table 2, LSTM has a slight advantage in music emotion classification, while MCCLSTM uses multi-channel CNN and LSTM to perform music emotion recognition and classification task. Although the recognition accuracy is a little more stable than LSTM, the complex model will increase the model training time, so this paper chooses to use LSTM and FNN.

The width learning system using the cascaded convolutional neural network has shown obvious advantages in music emotion classification. Compared with other complex models, the accuracy of music emotion classification has been greatly improved, thus proving the superiority of FNN network. The network structure proposed in this paper makes full use of the ability of advanced network to deal with complex data quickly. Its advantages lie in its simple structure and short training time, thus improving the recognition efficiency. LSTM has excellent performance in extracting time sequence features from time series data. It can extract the time sequence relationship of music, so as to retain music emotion features to the greatest extent. Combining the advantages of the above, LSTM-FNN network model is obtained for music emotion classification task. Figure 5 shows the comparison results of classification accuracy with different models. The recognition accuracy of LSTM-FNN model is 13% higher than MCCLSTM, 10.2% higher than RCNNBL, and 9.5% higher than CCFBLS, which proves that LSTM-FNN model can achieve music emotion classification more accurately.

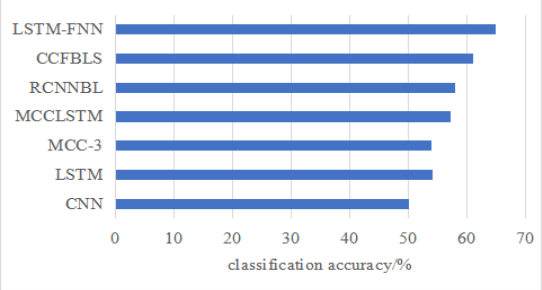

Figure 5. classification accuracy comparison distribution with different methods

Performance of the proposed method is compared with other methods for the standard feature set [17] in Table 3. Results in table 3 display that, for the standard feature set, LSTM+FNN produces greatly improvements in music emotion recognition in terms of F-measure compared to that of other six methods, respectively. But it slightly lower than LSTM-DNN [17]. In table 4, we can see that the running time is faster than LSTM-DNN due to the employ of three GPUs. It reflects the effectiveness of the proposed method.

Table 3. Performance comparison of proposed methods with other methods using standard features/%

Method	LSTM-FNN	CCFBLS	RCNNBL	MCCLSTM	MCC-3	LSTM	CNN	LSTM-DNN
Precision	91.7	90.9	89.6	87.1	84.5	87.3	83.9	92.5
Recall	91.5	90.7	88.8	86.2	87.0	86.5	84.1	92.7
F-measure	91.6	90.8	89.2	86.6	84.3	87.1	83.9	92.6

3.4. Comparison of training efficiency

In order to verify the training efficiency of LSTM-FNN model and other complex prediction models, the features extracted from all models are predicted respectively. For the same feature, table 4 shows the average 10-fold cross-validation training time required by these models. It can be seen that the training efficiency based on FNN model is much higher than that based on LSTM model for Emotion data set. LSTM-FNN model also has higher training efficiency than MCCLSTM+FNN model. Comparing CCFBLS model with LSTM-BLS model, the former has higher training efficiency than the latter, because LSTM model itself is more complex than CNN, so it is not abnormal to have a slightly lower training efficiency.

Table 4. Training efficiency comparison with different methods

Method	Training time/s
MCCLSTM	236.85
RCNNLSTM	604.46
MCCLSTM+FNN	274.54
CCFBLS	112.28
LSTM-DNN	163.74
LSTM-FNN	158.21

5. Conclusion

In this paper, the LSTM-FNN network model based on FNN and deep learning is proposed for music emotion recognition and classification. In the stage of audio preprocessing, MFCC feature and RP feature can be used to extract more comprehensive music emotion features. In the emotion prediction stage, FNN in the cascaded LSTM network mapping nodes is used for model training. The network structure makes full use of the fast processing ability in FNN. Its advantages are that the structure is simple and it has short model training time, thus improving the recognition efficiency. LSTM has excellent performance in extracting time sequence features from time series data. It can extract the time sequence relationship of music, so as to retain the emotional

features of music to the greatest extent. Combining the advantages of LSTM and FNN, LSTM-FNN network model is obtained to perform the task of music emotion classification. Experiment results show that the LSTM-BLS network model has higher recognition accuracy than the single deep learning model, and realizes lower time complexity than the complex model based on LSTM, and effectively realizes the emotion classification of music. As the emotions expressed in different "paragraphs" of complex music may not be consistent with the overall emotions, which brings difficulties in recognition, resulting in low accuracy in this type of music by the proposed method in this paper. If the same audio data is divided into multiple segments, and multiple segments are fed into trained networks to "vote", the judgment of musical emotion may be more accurate and scientific. At the same time, experiments using more complex neural network architectures, such as the excellent recurrent neural network for time-related classification of music emotion classification, it may be able to obtain better results than the presented method in this paper.

Acknowledgements.

The author greatly appreciates the anonymous comments of the reviewers.

References

[1] Kim JY, Belkin NJ. (2002) Categories of mus[illegible] descri[illegible] and search terms and phrases used by n[illegible]-mus[illegible] xperts. Proc. of the Third International Co[illegible]rence on Music Information Retrieval (ISMIR2002), Paris, [illegible]rance.

[2] Zhang WN, Ming ZY, Zhang Y, et al. (2016) Capturing the Semantics of Key Phra[illegible]s Usi[illegible] Multiple Languages for Question Retrieval. IEEE [illegible]nsaction[illegible]on Knowledge & Data Fngine[illegible]ing [illegible](4): 888-900. doi:10.1109/TKDE.2015.2[illegible]02944

[3] Suwicha J, Setha PN, [illegible]sin I. (2014) EEG-Based Emotion Recognition Using Deep [illegible]earning Network with Principal Component Based Covariate Shift Adaptation. The Scientific World Journal 2014:627892. doi:10.1155/2014/627892

[4] Chen X, Li J, Zhang Y, et al. (2020) Automatic Feature Extraction in X-ray Image Based on Deep Learning Approach for Determination of Bone Age. Future Generation Computer Systems 110: 795-801. https://doi.org/10.1016/j.future.2019.10.032

[5] Weninger F, Eyben F, Schuller B. (2014) On-line continuous-time music mood regression with deep recurrent neural networks. ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 5412-5416. doi:10.1109/ICASSP.2014.6854637

[6] Markov K, Matsui T. (2014) Music Genre and Emotion Recognition Using Gaussian Processes. IEEE Access 2:688-697. doi:10.1109/ACCESS.2014.2333095

[7] Chen S, Lee Y, Hsieh W and Wang J. (2015) Music emotion recognition using deep Gaussian process. 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) 495-498. doi: 10.1109/APSIPA.2015.7415321.

[8] Li XX, Xianyu HS, Tian JS, Chen WX et al. (2016) A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 544-548. doi: 10.1109/ICASSP.2016.7471734.

[9] Li X, Song D, Zhang[illegible], et al.[illegible](2017) Emotion recognition from multi-chann[illegible] EEG da[illegible] through Convolutional Recurrent Neu[illegible]l Net[illegible]ork. IEE[illegible] International Conference on Bioinf[illegible]matics &[illegible] Bi[illegible]medicine 352-359. doi: 10.1109/[illegible]BM.20[illegible]6.7822545

[10] Sarkar R, Cho[illegible]dhury[illegible], Dutta S, et al. (2020) Recognition [illegible]emo[illegible]tion in m[illegible]c based on deep convolutional neural network. Multimedia Tools and Applications 79(10): 765-783. doi: 10.[illegible]007/s11042-019-08192-x

[11] Orje[illegible]k R, Jarina R, Chmulik M and Kuba M. (2019) DNN Bas[illegible] Music Emotion Recognition from Raw Audio [illegible]gnal. 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA) 1-4. doi: 10.1109/RADIOELEK.2019.8733572.

[12] Issa D, Demirci M F, Yazici A. (2020) Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control 59:101894. https://doi.org/10.1016/j.bspc.2020.101894

[13] Nalini NJ, Palanivel S. (2016) Music emotion recognition: The combined evidence of MFCC and residual phase. Egyptian Informatics Journal 17(1):1-10. doi: 10.1016/j.eij.2015.05.004

[14] Chias-Palacios C, Vargas-Salgado C, Aguila-Leon J, et al. (2021) A cascade hybrid PSO feed-forward neural network model of a biomass gasification plant for covering the energy demand in an AC microgrid. Energy Conversion and Management 232.

[15] Tang N, Chen N. (2020) Combining CNN and broad learning for music classification. IEICE Transactions on Information and Systems E103. D(3): 695-703.

[16] Chen CLP, Liu Z and Feng S. (2019) Universal Approximation Capability of Broad Learning System and Its Structural Variations. IEEE Transactions on Neural Networks and Learning Systems 30(4): 1191-1204. doi: 10.1109/TNNLS.2018.2866622.

[17] Hizlisoy S, Yildirim S, Tufekci Z. (2020) Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and

Technology an International Journal, 2020, 24(3):760-767. DOI: 10.1016/j.jestch.2020.10.009

[18] Brewer M., Rahman J.S. (2020) Pruning Long Short Term Memory Networks and Convolutional Neural Networks for Music Emotion Recognition. Neural Information Processing. ICONIP 2020. Lecture Notes in Computer Science, vol 12534. Springer, Cham.

https://doi.org/10.1007/978-3-030-63836-8_29

[19] Benito-Gorron D D, Lozano-Diez A, Toledano D T, et al. Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset[J]. EURASIP Journal on Audio Speech and Music Processing, 2019, 2019(1).

[20] Yin SL, Li H, Laghari AA, et al. (2021) A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection. EAI Endorsed Transactions on Scalable Information Systems 21(33), e8. http://dx.doi.org/10.4108/eai.6-10-2021.171247

[21] Wang DL, Wang XW, Yin SL. (2021) A New Recursive Neural Network and Center Loss for Expression Recognition. International Journal of Electronics and Information Engineering 13(3): 97-104. DOI: 10.6636/IJEIE.202109_13(3).02

[22] Kaminskas M, Ricci F. (2012) Contextual music information retrieval and recommendation: State of the art and challenges. Computer Science Review 6(2-3):89-119. https://doi.org/10.1016/j.cosrev.2012.04.002

[23] Yin SL and Li H. (2020) Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13: 5862-5871. doi: 10.1109/JSTARS.2020.3025582.

[24] Wang JJ, Huang R. (2021) Music Emotion Recognition Based on the Broad and Deep Learning Network. Journal of East China University of Science and Technology. doi: 10.14135/j.cnki.1006-3080.20210225007

[25] Chen C, Hua Z, Zhang R, et al. (2020) Automated arrhythmia classification based on a combination network of CNN and LSTM. Biomedical Signal Processing and Control 57:101819.

https://doi.org/10.1016/j.bspc.2019.101819

[26] Pons J, Lidy T and Serra X. Experimenting with musically motivated convolutional neural networks. 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), 1-6. doi: 10.1109/CBMI.2016.7500246.

[27] Yin SL, Li H, Teng L, et al. (2020) An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images [J]. International Journal of Image and Data Fusion 11(2): 201-214. DOI: 10.1080/19479832.2020.1727573

[28] Yin, S., Li, H. GSAPSO-MQC:medical image encryption based on genetic simulated annealing particle swarm optimization and modified quantum chaos system. Evolutionary Intelligence, 14: 1817-1829, 2021. doi: 10.1007/s12065-020-00440-6

[29] Desheng Liu, Linna Shan, Lei Wang, Shoulin Yin, et al. P3OI-MELSH: Privacy Protection Point of Interest Recommendation Algorithm Based on Multi-exploring Locality Sensitive Hashing[J]. Frontiers in Neurorobotics, 2021. doi: 10.3389/fnbot.2021.660304.

[30] Jisi A and Shoulin Yin. A New Feature Fusion Network for Student Behavior Recognition in Education [J]. Journal of Applied Science and Engineering. vol. 24, no. 2, pp.133-140, 2021.

[31] Shoulin Yin, Hang Li, Desheng Liu and Shahid Karim. Active Contour Model Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation [J]. Multimedia Tools and Applications. Vol. 79, pp. 31049-31068, 2020.

[32] Laghari, A.A., Wu, K., Laghari, R.A. et al. A Review and State of Art of Internet of Things (IoT). Arch Computat Methods Eng (2021). https://doi.org/10.1007/s11831-021-09622-6

[33] Laghari A A, Laghari M A. Quality of experience assessment of calling services in social network[J]. ICT Express, 2021, 7(2): 158-161. doi: 10.1016/j.icte.2021.04.011

[34] A. A. Laghari, H. He, A. Khan, N. Kumar and R. Kharel, "Quality of Experience Framework for Cloud Computing (QoC)," in IEEE Access, vol. 6, pp. 64876-64890, 2018, doi: 10.1109/ACCESS.2018.2865967.