Front. Comput. Sci., 2022, 16(6): 166335
https://doi.org/10.1007/s11704-021-0569-4

REVIEW ARTICLE

A survey of music emotion recognition

Donghong HAN (✉)[1], Yanru KONG[1], Jiayi HAN[2], Guoren WANG[3]

1 School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China
2 Institute of Science and Technology Brain-Inspired Intelligence, Fudan University, Shanghai 200082, China
3 School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100089, China

Abstract Music is the language of emotions. In recent years, music emotion recognition has attracted widespread attention in the academic and industrial community since it can be widely used in fields like recommendation systems, automatic music composing, psychotherapy, music visualization, and so on. Especially with the rapid development of artificial intelligence, deep learning-based music emotion recognition is gradually becoming mainstream. This paper gives a detailed survey of music emotion recognition. Starting with some preliminary knowledge of music emotion recognition, this paper first introduces some commonly used evaluation metrics. Then a three-part research framework is put forward. Based on this three-part research framework, the knowledge and algorithms involved in each part are introduced with detailed analysis, including some commonly used datasets, emotion models, feature extraction, and emotion recognition algorithms. After that, the challenging problems and development trends of music emotion recognition technology are proposed, and finally, the whole paper is summarized.

Keywords artificial intelligence, deep learning, music emotion recognition

1 Introduction

In recent years, the electronic music market has achieved rapid development, massive music resources can be obtained from various sources. These music resources need to be organized and managed based on label information such as emotion, genre, etc. so that listeners can obtain music works conveniently. Since music is the carrier of emotions, so it is particularly important to recognize the emotion labels of music works. Using manual methods to obtain the label information can be time-consuming, labor-intensive, and error-prone. Therefore, the research field of automatically recognizing emotion labels came into being.

Music emotion recognition (MER) constitutes a process of using computers to extract and analyze music features, form the mapping relations between music features and emotion space, and recognize the emotion that music expresses [1]. Music features are often extracted from the audio signal, symbolic music scores, lyrics texts, and even biological features like EEG (Electroencephalogram). Emotion space can be represented by a finite number of discrete categories or the infinite number of points in a continuous multidimensional space.

MER belongs to the interdisciplinary research field of music psychology, audio signal processing, and natural language processing (NLP), and MER is a sub-task of music information retrieval (MIR). MER can be wildly used in many fields, including music recommendation [2], music retrieval [3], music visualization, automatic music composing, psychotherapy, and so on. Therefore, MER has become a research hotspot in the academic and industrial community.

Since the 1930s, researchers have launched pioneering research on the relationships between music and emotions. At the beginning of this century, more researchers began to study how to automatically extract emotion from music data. In 2007, Music Information Retrieval Evaluation eXchange (MIREX), one of the most authoritative international audio retrieval and evaluation competition, added audio mood classification (AMC) to its competition tasks, showing the significance of MER. Later, Kim et al. [4] reviewed the works of MER with a comprehensive analysis in 2010. Now with the emergence and development of deep learning, MER is facing more challenges and opportunities.

Recently, with people's deep understanding of music characteristics and the growing maturity of artificial intelligence, MER has made great progress. It's necessary to summarize the research progress in time. This paper gives a detailed survey of music emotion recognition. Since there have been some high-quality reviews of MER works focusing on audio features and traditional machine learning algorithms [4,5], the focus of this paper will be on the MER researches that use deep learning algorithms in recent years. The contributions of this paper are as follows: (1) Briefly describe the preliminary knowledge, hand-crafted features, and traditional machine learning algorithms in the MER field, and conduct a detailed analysis of MER researches that use deep learning methods. (2) Put forward some opinions on the challenges faced by MER, and propose some development directions.

Received November 29, 2020; accepted June 25, 2021

E-mail: handonghong@cse.neu.edu.cn

The rest of this paper is organized as follows: Some preliminary knowledge and the research framework of MER are introduced in Section 2. Section 3 reviews and analyzes the algorithms involved in the framework. Section 4 summarizes the development status and trends of MER. Section 5 gives a summary of this paper.

2 Preliminary knowledge

Existing MER research papers can be roughly divided into two research directions, namely song-level MER (or static) and music emotion variation detection (MEVD, or dynamic). Song-level MER refers to assigning the overall emotion label (or labels if the task is considered a multi-classification or regression problem) to one song. While MEVD considers the emotion of music as a changing process, the dynamic changing process of emotion needs to be recognized when conducting MEVD research. The related information is summarized in Table 1.

2.1 Evaluation metrics

The MER model needs evaluation metrics to evaluate performance. The evaluation metrics commonly used in this field can be divided into metrics for classification and regression models respectively.

For classification problems, metrics like accuracy and precision are commonly used. Accuracy can calculate the proportion of correctly classified samples to the total number of samples, but it does not perform well on unbalanced data, so precision is introduced. Precision is the proportion of the real positive samples to the total number of samples predicted to be positive.

For regression problems, the evaluation metrics used in the MER field include R^2 and Root Mean Square Error (RMSE). R^2 is the coefficient of determination, it can evaluate how well the regression model fits the sample data. RMSE can calculate the error between the predicted value and the true value.

2.2 Research framework

Most existing MER works based on machine learning include three parts [1,6], namely domain definition, feature extraction and emotion recognition. The overall framework is shown in Fig. 1. It can be seen from Fig. 1 that emotion models and datasets are selected in the domain definition stage, useful features are extracted in the feature extraction stage, and the emotion label is predicted in the emotion recognition stage. Section 2.3 will summarize the knowledge involved in the domain definition stage, and Section 3 will give a detailed analysis of the models and algorithms used in the feature extraction and emotion recognition stage.

2.3 Emotion models and datasets

As shown in Fig. 1, in the domain definition stage, emotion models and datasets need to be selected to define the scope of the research. The following will introduce some preliminary knowledge as well as some commonly used emotion models and datasets.

2.3.1 Emotion model

Table 2 summarizes some commonly used emotion models in MER. In the "Application Domain" column, "General" refers to general emotion models, and "Music" refers to music emotion models. General emotion models can be used for sentiment analysis in various fields, which is good for multi-modal MER. Music emotion models are dedicated to the music domain, which can describe music emotions more accurately. In the "Emotion Conceptualization" column, "Categorical" refers to the categorical emotion model and "Dimensional" means the dimensional emotion model. Some scholars believe that the categorical emotion model is ambiguous [12,13], so dimensional emotion models are used more recently [12,14]. In the "Emotional Definition" column, "Perceived" refers to perceived emotion, and "Induced" is induced emotion. Perceived emotion means the emotion conveyed by the music itself, which usually needs music data

Table 1 MER research directions

Methodology	Emotion conceptualization	Description
Song-level MER	Categorical approach	Predict the categorical emotion labels of music pieces
	Dimensional approach	Predict the numerical emotion values of music pieces
MEVD	Categorical approach	Predict the dynamic categorical emotion variation within a music piece
	Dimensional approach	Predict the dynamic dimensional emotion variation within a music piece

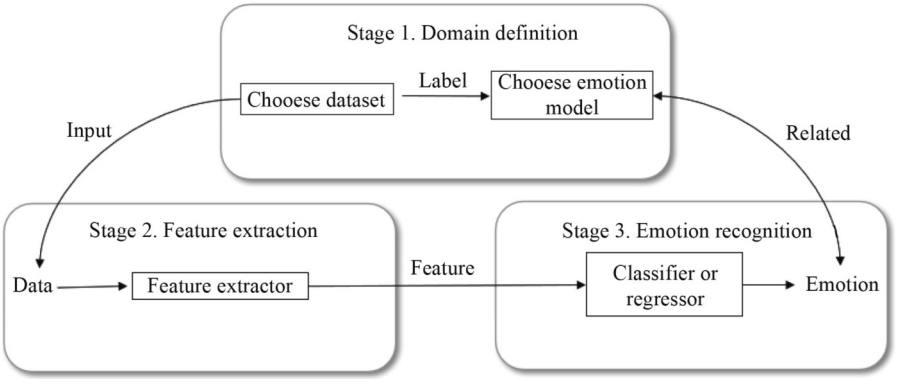


Fig. 1 MER framework

(like audio, symbolic music scores, and lyrics, etc.) to recognize it. Induced emotion is the emotion that the music provokes among the audience [15], which needs to be recognized by the physiological data (like EEG [16]) generated when the audiences listen to music.

Next, this paragraph will introduce the emotion models listed in Table 2. Hevner's affective ring [7] is one of the earliest and most influential music emotion models. In 1935, Hevner conducted extensive experiments and found 67 emotional adjectives to describe the emotional space expressed by music. These 67 emotional adjectives can be divided into eight categories, namely dignified, sad, dreamy, serene, graceful, happy, exciting, and vigorous. Russell's circumplex model of affect [8,9] is the most commonly used emotion model in current MER papers. This emotion model has two dimensions, which are valence and arousal. Valence reflects the degree of positive and negative emotion, ranging from unpleasant to pleasant. The intensity of emotion is reflected by arousal, ranging from passive to activated. GEMS (the Geneva Emotional Music Scales) [10] is considered to be the first emotion model designed for music-induced emotion, it contains 45 emotional tags, which are divided into nine categories, namely amazement, solemnity, tenderness, nostalgia, calmness, power, joyful activation, tension, sadness. The two-dimensional emotion model designed by Thayer [11] contains energetic arousal and tense arousal, and he believes that valence can be explained by the different combinations of energetic and tense arousal.

In addition to the emotion models listed in Table 2, some scholars also use ranking [15], probability distributions [17], and pairs of antonyms [18] to express music emotions. Ranking refers to rank the music works according to the intensity of a certain emotion, which can reduce the cognitive load caused by manually labeling continuous emotional polarities and inaccuracy caused by emotion subjectivity. Probability distribution refers to represent the emotion of a song as a probability distribution in the emotion space, which can also alleviate the problems caused by the subjectivity of emotion. The emotion space in literature [18] is composed of multiple pairs of antonyms, which can make emotion labels more objective.

2.3.2 Datasets

Due to music copyright restrictions, some MER researchers use self-built and unpublished datasets. Table 3 lists some commonly used public datasets. Million Song Dataset [27] (MSD) is not in Table 3, because is not more like a resource integration platform that collects authoritative music data from seven music communities. MSD does not provide the original audio and lyrics but provides processed features like Mel-frequency cepstral coefficient (MFCC) and bags-of-words (BOW). And MSD does not provide the emotion label for each song.

3 Feature extraction and emotion recognition

There are currently two main methods for feature extraction and emotion recognition steps. One is to complete the two steps separately, by extracting handcrafted features and using them to train the traditional machine learning (ML) model to predict emotion. The other one is to complete them together through deep learning (DL) models.

3.1 Handcrafted features and traditional machine learning model

3.1.1 Handcrafted features

Feature extraction is the core issue of MER, the quality of features directly affects the accuracy of emotion recognition. For MER task, features can be extracted from the audio signal, symbolic music scores, lyrics, and even physiological data generated by listeners. Therefore, MER tasks mainly contain four kinds of handcrafted features, namely audio feature, symbolic feature, lyric feature and biological feature. Some data formats, preprocessing methods, tools and results are summarized in Table 4.

- *Audio feature.* Audio features are the earliest and most widely studied features in the MER field, and mostly extracted from waveform files with the help of existing toolkits in literature [28−31].

  Audio features related to emotions can be divided into rhythmic features, timbre features, and spectral features

Table 2 A summary of emotion models

Model name	Application domain	Emotion conceptualization	Number of classes/dimensions	Emotional definition
Hevner affective ring [7]	Music	Categorical	67	Perceived
Russell's circumplex model of affect [8,9]	General	Dimensional	2	Perceived
GEMS [10]	Music	Categorical	45	Induced
Thayer [11]	General	Dimensional	2	Perceived

Table 3 A summary of datasets

Dataset name	Emotion conceptualization	Number of songs	Data type	Genres	Research directions
MediaEval emotion in music [19]	Dimensional	1000	MP3	Rock, pop, soul, blues, etc.	Dynamic
CAL500 [20]	Categorical	500	MP3	–	Static
CAL500exp [21]	Categorical	3223(segments)	MP3	–	Dynamic
AMG1608[22]	Dimensional	1608	WAV	Rock, metal, country, jazz, etc.	Static
DEAM [23]	Dimensional	1802	MP3	Rock, pop, electronic, etc.	Dynamic
MTurk [24]	Dimensional	240	–	–	Dynamic
Soundtracks [25]	Categorical and dimensional	360	MP3	Rap, R&B, electronic, etc.	Static
Emotify music database [26]	Categorical	400	MP3	Rock, classical, pop and electronic	Static

[32]. The most commonly used features of each category are summarized in Table 5. Rhythm refers to the changes in pitch, duration, speed, and severity of audio that are above semantic symbols, its existence determines whether a sentence sounds natural and mellow. The most commonly used rhythmic features include duration, pitch, and energy [33], among them pitch and energy are highly related to arousal detection [5]. Timbre describes the sound quality [1], and timbre features have been shown to provide the best performance in MER systems when used as individual features [34]. The most commonly used timbre feature is MFCC, it represents the format peaks of the spectrum [5], and reflects the nonlinear frequency sensitivity of the human auditory system [1]. Spectral features are considered to be the manifestation of the correlation between vocal tract shape changes and articulator movement. The emotional content in audio has a significant impact on the distribution of spectrum energy in each spectrum interval [33].

- *Symbolic feature.* Symbolic features refer to features extracted from symbolic music scores. These features are less studied compared with audio features. In the field of MER, symbolic music scores are usually represented by MIDI (Musical Instrument Digital Interface), MIDI file is a music file format that directly contains the exact sequences of pitches, intensity, etc. The most commonly used symbolic features are related to pitch, interval, loudness, and duration [35].
- *Lyric feature*. Lyrics are very important for conveying semantic information [36]. With the advancement of NLP technology, more and more researches are focusing on lyrics. When manually extracting lyric features, researchers mainly learn from NLP technologies such as BOW [37−39], *n*-grams [38], and part-of-speech (POS) [39]. Hu et al. [39] compare BOW, POS, and function words. Zaanen et al. [40] use the term frequency-inverse document frequency (TF-IDF) to measure the relevance of words and emotion categories. Based on [40], literature [41] adds rhyme information. However, Malheiro et al. [42] believe that features like BOW and POS are not enough for MER, so they propose three novel features, namely slang presence, structural analysis features, and semantic features. In addition, methods such as sentiment lexicon [43], statistical analysis tools [44], and Latent Dirichlet Allocation (LDA) [45] are also used to extract lyric features.
- *Biological feature.* Induced emotion emphasizes the listener's experiences, so it needs to collect physiological data from the listeners. EEG can effectively capture information about emotions from the brain with high temporal resolution and low cost [46], so the earliest and frequently used biological feature is EEG. Besides, with the development of medical technology and wearable devices, the collection of peripheral physiological signals such as heart rate (HR) and skin temperature (TEMP) has also become convenient [47]. And many researchers have made bold attempts to collect induced data, such as the functional magnetic resonance imaging (fMRI) technology [48]. However, compared with the perceived features like audio and lyrics, biological features are still in the early stage of development. Since the biological feature is very helpful for exploring the impact of music on the human brain and personalized music recommendations, it is worth investigating further.

3.1.2 Traditional machine learning models

The literature that uses traditional machine learning models to implement MER systems can be divided into four categories, namely song-level categorical MER, song-level dimensional MER, categorical MEVD, and dimensional MEVD. Since there are few studies on categorical MEVD (probably due to the lack of dynamically labeled categorical datasets), therefore the literature using MEVD ideas will be introduced together. The following will briefly describe the literature from three aspects.

- *Song-level categorical MER.* The representative works of song-level categorical MER are summarized in Table 6. Table 6 shows that the frequently used classification model is the support vector machines (SVM). In addition, *k*-nearest neighbors (KNN), decision tree (DT), random forests (RF), and naïve Bayes (NB) are also involved. In Table 6, Hu et al. [39] compare BOW, POS, and function words, and experiments show that the best performing feature is BOW with stemming. Hu et al. [47] collect peripheral physiological signals from subjects and input them into four kinds of classification

Table 4 Data format and processing information

Data format	Preprocessing method	Preprocessing tools	Preprocessing result
WAV, MP3	Framing, windowing, MFCC extraction, spectrogram extraction, etc.	Psysound (software), MIRtoolbox (MATLAB), Librosa (Python package), etc.	MFCC, spectrogram, etc.
MIDI	Main track extraction, etc.	pretty_music, music21 (all Python package)	Key, BPM, melody, etc.
Text	Segmentation, cleaning, normalization, etc.	NLTK, Gensim, Jieba, Stanford NLP (all Python package), etc.	BOW, TF-IDF, word embeddings, etc.

Table 5 Audio features

First level audio feature	Second level audio features
Rhythmic features	Duration, pitch, energy, etc. [33]
Timbre features	MFCC, zero crossing rate, chroma, etc. [32]
Spectral features	Spectral flatness measure, spectral centroid, etc. [32]

methods, KNN shows the best performance with accuracy being about 60%. Li et al. [49] use MARSYAS to extract timbre, rhythmic and pitch features, then input them into SVM for classification, but there is a large variation in the accuracy for different mood categories. Laurier et al. [50] use two fusion methods, and experiments show that fuse features first can have slightly better results. Literature [51] evaluates multiple features and fusion methods, with sufficient experiments, they conclude that late fusion by subtask merging can improve the classification accuracy, and lyrics carry semantic information which can complement the audio. Liu et al. [52] propose an algorithm called multi-emotion similarity preserving embedding (ME-SPE), and they combine ME-SPE with calibrated label ranking (CLR) to recognize the emotions of music.

- *Song-level dimensional MER*. The representative works in this domain are summarized in Table 7. The most used regression methods include support vector regression (SVR), linear regression (LR), multivariate linear regression (MLR), Gaussian process regression (GPR) and acoustic emotion Gaussian (AEG), etc. In Table 7, [13] is one of the earliest works to regard MER as a regression problem, existing toolkits are adopted to extract 114 kinds of audio features and input them into SVR. Malheiro et al. [42] use lyrics feature to conduct classification and regression experiments, and the uniqueness of [42] is that three novel lyrics features are proposed, namely slang presence, structural analysis features, and semantic features. [53] proposes a novel generative model name acoustic emotion Gaussians (AEG) to recognize emotion in music, which demonstrates superior accuracy over models like SVR and MLR. [54] explores how to adapt the AEG model in [53] to a personalized MER model with minimum user load. [55] also explores personalized MER with an LR-based model. Fukayama et al. [56] use GPR to take new acoustic signal inputs into account. [57] tries static and dynamic approaches using ML and DL methods, so a detailed analysis of [57] will be given in Section 3.2.3.
- *MEVD*. The representative research works on MEVD are summarized in Table 8. It can be seen that the most commonly used ML models are SVM-based models. Besides SVM, Gaussian mixture model (GMM), MLR and GPR are also adopted. The idea of music mood tracking was first proposed in [58]. Lu et al. [58] use a boundary detection algorithm to divide the entire song into several independent segments and assign an emotion label to each segment. And this paper is also the only one in Table 8 that belongs to the field of categorical MEVD. [13] shows that the regression approach can also be applied to MEVD, but their model lacks temporal information. [59] exploits multiple audio features for regression tasks and finds that the spectral contrast feature shows superior result. [60] proposes DS-SVR (Double-scale Support Vector Regression), which uses two independent SVR models. One identifies emotion changes among different songs, the other detects emotion changes within one song, then the results of the two SVRs are combined as the final result.

Table 6 Representative works of song-level categorical MER (ML)

Reference	Feature modalities	Machine learning model	Emotion model	Dataset
[39]	Lyric	SVM	18 classes	Self-built
[47]	Physiological signals	SVM, NB, KNN, DT	3 classes	Peripheral physiological signals data
[49]	Audio	SVM	13/6 classes	Self-built
[50]	Audio and lyric	SVM, RF	4 classes	Self-built
[51]	Audio and lyric	SVM	4 classes	Self-built
[52]	Audio	CLR	6/18 classes	EMOTIONS, CAL500

Table 7 Representative works of song-level dimensional MER (ML)

Reference	Feature modalities	Machine learning model	Emotion model	Dataset
[13]	Audio	SVR, MLR	VA model	Self-built
[42]	Lyric	SVM	VA model	Self-built
[53]	Audio	AEG	VA model	MTurk, MER60
[54]	Audio	AEG	VA model	AMG1608
[55]	Audio	LR	VA model	AMG240
[56]	Audio	GPR	VA model	MediaEval emotion in music
[57]	Audio	SVR, MLR, GPR	VA model	Self-built

Table 8 Representative works of MEVD (ML)

Reference	Feature modalities	Machine learning model	Emotion model	Dataset
[13]	Audio	SVR	VA model	Self-built
[57]	Audio	SVR, MLR, GPR	VA model	Self-built
[58]	Audio	GMM	4 classes	Self-built
[59]	Audio	SVM, SVR	VA model	Self-built
[60]	Audio	DS-SVR	VA model	MediaEval emotion in music

3.2 MER based on deep learning

Since formally proposed in 2006, deep learning has developed rapidly in recent years. DL models like convolutional neural network (CNN) or recurrent neural network (RNN) can be adopted as an end-to-end processing framework, the entire learning process is completely handed over to the DL framework and completes the mapping from the original data to the expected output. Compared with traditional machine learning models, DL-based MER models have two advantages. First, the performance of DL models will increase as the amount of training data increases. Second, DL-based models can automatically extract suitable features from data.

Since the commonly used public dynamically labeled datasets mostly are dimensional datasets [19,23,24], so the literature using the ideas of MEVD all belongs to dimensional MEVD. The following will analyze research papers in three parts, namely song-level categorical MER, song-level dimensional MER, and dimensional MEVD.

3.2.1 Song-level categorical MER

The representative works in this domain are summarized in Table 9. It can be seen that CNN-based models are common. CNN is one of the representative learning algorithms of DL, it mimics the visual perception of the living creature, and can learn feature representations from data effectively. Next, this paper will analyze the references listed in Table 9 in detail according to the publication time.

The model [61] proposed is called Bi-modal Deep Boltzmann Machine, it contains two 2-layer DBM (Deep Boltzmann Machine) networks, one for audio and one for lyrics, then an additional layer is added on top to join the two DBMS. Finally, the feature representation learned by their model is used as input to SVM for the final classification result. They compare their model with single modality, early fusion, and late fusion, and experiments show that their model can outperform other models in every mood category. In their experiments, all fusion methods can outperform single modality, affirming the effectiveness of multi-modal. But using lyrics alone can't achieve a satisfying result mainly due to the sparse representations of BOW features, so a higher-level lyric feature is needed.

[32] uses spectrogram computed via the Short-Time-Fourier-Transformation of audio signals as input, the spectrogram of each music goes through convolutional, pooling, and hidden layers, and predictions are made with a SOFTMAX at the end. The innovation of this paper is that it uses CNN to reduce the burden of manually extracting features. It also uses the convolution method on local time and frequency to make the spectrogram of each piece of music equal in length. But the disadvantage is that it is difficult to infer what feature contains the clues of musical emotion.

In addition to perceived features, [62] inputs EEG and other biological features into CNN for emotion recognition, and the results are superior to old methodology like SVM. The difference between [62] and other EEG-based MER works is that they focus on creating a model that is subject-independent. But they believe a more diverse dataset is needed.

Sarkar et al. [63] improve VGGNet (Visual Geometry Group Net) for MER. VGGNet is an improved version of CNN, which has fewer layers. Handcrafted features plus traditional ML model and VGGNet-based method are both tried by Sarkar et al. [63], and they find that the performance of handcrafted feature varies for different classifiers, and deep learning method improves the recognition accuracy considerably. But their model does not perform well on arousal, and the time series nature of audio cannot be emphasized by using VGGNet.

[64] proposes an emotion recognition method based on CNN. Yang et al. [64] try a variety of feature extraction methods to convert the original data into spectrograms and then input the spectrogram into CNN for emotion recognition. Results show that the spectrograms based on Constant-Q Transform can achieve the best performance.

3.2.2 Song-level dimensional MER

The representative works in the field of song-level dimensional MER are listed in Table 10, CNN and RNN-based DL frameworks are often used. RNN is another representative algorithm of DL, it is good at processing sequence data, so it is often used in the field of NLP. Bi-RNN and LSTM are two commonly used variants of RNN.

The model Ma et al. [65] propose is a multi-scale context based attention model (MCA). The innovation of this paper is that they use the attention mechanism to dynamically integrate different time scales to learn the temporal and hierocracy information of music. The attention mechanism is used twice in their paper. Firstly, they train an LSTM for each time scale to calculate its VA values, the attention mechanism is used to calculate the weight for each time step, and the weighted sum of all hidden layers is the final result of this time scale. Secondly, they apply the attention mechanism again to calculate weights for each time scale, and the weighted sum of all time scales is the final result of this song. And their experiments show that fusing different time scales with attention can learn the representation of music structure dynamically.

Liu et al. [18] adopt BiLSTM (Bi-directional Long Short-Term Memory) to extract features from audio, but their innovation is that in addition to LSTM, they also extract

Table 9 Representative works of song-level categorical MER (DL)

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[61]	2016	Audio and lyric	DBM	4 classes	MSD
[32]	2017	Audio	CNN	18 classes	CAL500, CAL500exp
[62]	2019	EEG	CNN	2 classes	EEG data collected from subjects
[63]	2020	Audio	VGGNet	4 classes	Soundtracks, Bi-Modal
[64]	2020	Audio	CNN	2 classes	EmoMusic

constant features and integrate them, then input the integrated features into the fully connected layer for emotion recognition. The constant features are extracted from tempo and energy information.

GARN in [66] refers to Genre-Affect Relationship Network, which is a multi-task learning framework. The main task is emotion regression, and the auxiliary task is genre classification. Since arousal emotion depends on the genre of music, age, and gender, so adding an auxiliary task to learn the relationship between emotion and genre can benefit the emotion recognition process.

Delbouys et al. [67] adopt audio and lyrics to test various learning models and fusion methods. After experiments, they conclude that two layers of CNN are better for processing audio features, and the lyric are better to be processed by one layer of CNN plus one layer of LSTM. The contribution of this paper is to compare the effects of multiple models.

BCRSN in [68] refers to the Bidirectional Convolutional Recurrent Sparse Network, which combines the advantages of CNN and RNN. CNN can learn features adaptively, and RNN is more suitable for processing sequence data. The uniqueness of this work is that they combine CNN and LSTM to strengthen the ability to learn features from the spectrogram. And they also optimize the emotion representation to make the training process faster.

[69] not only tries the VGG-based model with spectrograms but also exploits mid-level features. Mid-level features can represent some musical quality and are recognizable by listeners without music-theoretic knowledge. The VGG-based modal in [69] can learn two individual prediction tasks. One is to predict mid-level features from the audio, and the second is to predict emotions from mid-level features. The uniqueness of [69] is building explainable models with human-interpretable mid-level perceptual features. Pursuing interpretable models can shed some light on the nature of MER.

3.2.3 Dimensional MEVD

The representative works in this field using the DL technique are summarized in Table 11. In dynamic emotion recognition, models based on RNN are commonly used. The reason is that the contextual and structural information in music is very important for identifying emotional changes, and RNN is great at extracting these kinds of sequential information.

Weninger et al. [12] segment music to seconds and extract supra-segmental features for each segment, and then input them into LSTM to obtain the dynamic changing process of VA values.

[57] tries static and dynamic approaches using ML and DL models. Experimental results show that for the static MER task, GPR and SVR can outperform MLR. For the dynamic MER task, the combination of a large feature set and RNN model shows the best performance.

Li et al. [70] believe that the emotion at a certain point in music is not only related to the content before the point, but also after the point, so DBLSTM (Deep Bidirectional Long Short-Term Memory) is proposed to extract information in both directions. Their model has three parts, including DBLSTM, post-processing, and fusion. DBLSTM initially extracts temporal context and hierarchical structure information from two directions and gives a prediction of VA value. Post-processing and fusion can further improve the model's ability to extract temporal and hierarchical information. Post-processing is applied to make use of the temporal correlation, and fusion is to fuse the results of DBLSTM with multiple different time scales into one prediction. They carry out experiments on the Emotion in Music task at MediaEval 2015 dataset and found that using DBLSTM alone can be better than using MLR or SVR, indicating that DBLSTM has the ability to capture contextual information. Adding post-processing and fusion, especially post-processing after fusion, the effect is more significant, indicating the effectiveness of the combination of post-processing and fusion.

The attentive LSTM proposed by Chaki et al. [71] is an LSTM incorporated with the modified attention mechanism. They believe that the music's emotion at a certain moment is only dependent on the music preceding it. So, when they calculate the context vector for a certain moment, they only consider the hidden states before it. Experiments show that their results are significantly better than literature [12].

4 Development status and trends

MER has made great progress, and the number of literatures is increasing. The metrics on the performance on some common

Table 10 Representative works of song-level dimensional MER (DL)

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[65]	2017	Audio	LSTM, attention mechanism	VA model	Emotion in Music task at MediaEval 2015
[18]	2018	Audio	BiLSTM	Based on VA model	DEAM
[66]	2018	Audio	GARN	GEMS	Emotify music database
[67]	2018	Audio and lyric	CNN, RNN, etc.	VA model	MSD
[68]	2019	Audio	BCRSN	Based on VA model	DEAM, MTurk
[69]	2019	Audio	VGG-based	8 dimensions	Mid-level Perceptual Features dataset, Soundtracks

Table 11 Representative works of dimensional MEVD (DL)

Reference	Year	Feature modalities	Learning model	Emotion model	Dataset
[12]	2014	Audio	LSTM, SVR	VA model	MediaEval Emotion in Music
[57]	2014	Audio	SVR, MLR, GPR	VA model	Self-built
[70]	2016	Audio	DBLSTM	VA model	MediaEval Emotion in Music
[71]	2020	Audio	Attentive LSTM	VA model	MediaEval Emotion in Music

datasets are presented in Table 12 to comprehensively show the current performance of MER methods. MediaEval Emotion in Music and AMC task in MIREX are two international MER competitions which have witnessed the development of MER in recent years, and they are attracting more researchers' attention.

4.1 Current challenges

Currently, there are many challenging problems in the MER field, and the following gives four of them.

Firstly, emotions are subjective and difficult to quantify. Different people may have different emotional perceptions of the same music, even the same person is also inconsistent in different times and situations, which reflects the subjectivity of music. For categorical emotion models, a few adjectives are difficult to accurately quantify the richness of music emotions. For dimensional emotion models, such as the most commonly used VA model, one of its quadrants often contains multiple approximate emotions (such as the first quadrant contains glad, excited, etc.), but which numerical value corresponds to which emotion is vague and difficult to quantify. Therefore, since the current emotion models are having difficulty in accurately quantifying the rich emotions of music, some further research on new emotion models is needed.

Secondly, although other data modalities and features besides audio have been used to extract feature, but the research towards it is insufficient. For example, the emotional words in lyrics are sparse, and some songs will change the normal word order in order to cater to the melody and rhyme. So, it is necessary to study how to extract new features according to the characteristics of music.

Thirdly, some authoritative large-scale diversified emotion-labeled music datasets are needed for the MER field. The number of songs in the existing public dataset is generally around hundreds, with a few in the range of 1500-2000 (see Table 3). At present, the largest dataset in the MIR field is the MSD, which has the capacity of one million songs. But the facts are that MSD only provides features, the genre of most songs is pop and the label quality is worrisome. The above facts limit the ability of researchers to develop novel features, and the designed MER system may not be suitable for other genres. To sum up, large-scale diversified emotion-labeled music datasets are urgently needed.

Fourthly, music concepts and theory knowledge are very important for MER, but there are few quantitative works on them. Some high-level music concepts (key, melody progression, etc.) can reflect music emotions to a certain extent. For example, the major key is usually bright and pleasant, while the minor key is soft and gloomy. The upward melody progression sounds positive and powerful, while the continuous downward melody progression gives people a dark feeling. Therefore, it is a feasible development direction to conduct MER research with reference to high-level music concepts. [5,72] state the same.

4.2 Development trends

By analyzing the research works in recent years, it is found that methods used in each part in the MER framework show the following development trends. (1) Domain definition. New datasets and emotion models have emerged, such as the dynamically annotated dataset DEAM, and the induced emotion model GEMS. Two reasons can account for these changes. First, dynamic processing is more in line with the characteristics of music. Music emotion will change dynamically within a music piece, so static processing is not detailed and accurate. Especially with the emergence of sequence models like RNN, dynamically recognizing continuous emotions has become more convenient. Some papers also demonstrate the necessity of shifting from static to dynamic processing [23,65]. The second reason is that multi-modal processing is superior. The performance of using audio data alone has reach glass-ceiling, so adding other information like induced data is necessary. Also, some papers have already shown that multi-modal can achieve better results [50,51,61,67]. (2) Feature extraction and emotion recognition. Methods for these two steps are shifting from manual extraction and traditional ML models, to using DL frameworks for end-to-end processing. Table 13 demonstrates the year, method and accuracy information for the AMC task in MIREX, the trend from ML to DL is obvious. This is because DL frameworks are simple but powerful. By combining feature extraction and emotion recognition stages, DL frameworks simplify the learning and training process. In addition, handcrafted features are difficult to maintain good performance across different classifiers and datasets [63], while DL frameworks can automatically extract useful features from raw data and maintain good performance. In summary, the MER field presents three development trends, namely from static processing to dynamic processing, from single modal to multi-modal and from traditional ML models to DL models.

At present, many new technologies have been added to the MER field, such as VGGNet [63,69], attention mechanism [65], etc. However, technologies that have achieved remarkable results in other fields, such as transfer learning and knowledge graphs, have yet to be explored. Since audio emotion recognition and text sentiment analysis have been well studied, therefore conduct transfer learning from the above fields to MER may show some promising results. In addition to audio and lyrics, other information such as singing

Table 12 Metrics on some common datasets

Reference	Method	Dataset	Performance
[61]	Bi-modal deep boltzmann machine	MSD	78.5% (Accuracy)
[67]	CNN, LSTM	MSD	0.219 for valence, 0.232 for arousal (R2)
[65]	MCA	MediaEval dataset	0.291 for valence, 0.241 for arousal (RMSE)
[70]	DBLSTM	MediaEval dataset	0.285 for valence, 0.225 for arousal (RMSE)
[32]	CNN	CAL500	42.6% (Marco average precision)
[52]	CLR	CAL500	48.8% (Marco average precision)

Table 13 Results of AMC task in MIREX

Year	Method	Accuracy/%
2020	Mel spectrogram + CNN	69.5
2019	-	68
2018	STFT + CNN	61.17
2017	Mel spectrogram + DCNN+SVM	69.83
2016	FFT, MFCC + CNN	63.33
2015	-	66.17
2014	MFCC + SVM	66.33
2013	Visual and acoustic features + SVM	68.33
2012	Audio features + SVM based models	67.83
2011	Audio features + SRC	69.5

voice, social tags, music video, and album cover data may also be helpful for the emotion recognition process, since knowledge graph can reveal the relationship between entities, therefore knowledge graph can be tried in MER field. Lastly, as mentioned above, the emotion of the music is related to melody progression and key, and literature [66] shows that arousal emotion is related to genre, therefore research ideas in MER-related fields can also be borrowed, such as automatic melody extraction (AME), chord recognition (CR) and music genre classification (MGC), etc.

5 Conclusions

This paper reviews the current research on MER. Firstly, it introduces the research background, gives the definition, summarizes the significance of MER, and gives a brief introduction of MER history. Then the current research framework is introduced, and the knowledge and algorithms involved in each part are elaborated. Lastly, the challenging problems and future development trends of MER are pointed out.

There are two main contributions of this paper. The first is to give a detailed analysis of research papers that use the DL technique, the uniqueness, models, and experiments of each paper are elaborated. Secondly, the challenging problems faced by MER and future development trends are pointed out in Section 4. Currently, in the field of MER, there are urgent needs for authoritative large-scale diversified datasets and more accurate emotion models. Music concepts and carefully designed features are also needed. Generally speaking, the MER field is shifting from static processing to dynamic process, from single modal to multi-modal and from traditional ML models to DL models, more technologies such as transfer learning and knowledge graphs, more information like the singing voice, social tags, album cover data, and MV data can be explored, and ideas from related fields including AME, CR, and MGC can be borrowed.

MER is still in the early stage of development since after ten years the performance of current MER systems seems to be stuck at 69% (see Table 13). But MER has great significance, such as it can be wildly used in many fields like automatic music emotion recognition, music recommendation [73], automatic music composing [74], psychotherapy, etc. Therefore, MER has become popular in the academic and industrial community and has huge development potential. It is believed that with the addition of researchers, MER has a bright development prospect.

Acknowledgements This work was supported by the National Nature Science Foundation of China (Grant Nos. 61672144, 61872072, 61173029) and the National Key R&D Program of China (2019YFB1405302)

References

1. Yang X Y, Dong Y Z, Li J. Review of data features-based music emotion recognition methods. Multimedia System, 2018, 24(4): 365–389
2. Cheng Z Y, Shen J L, Zhu L, Kankanhalli M, Nie L Q. Exploiting music play sequence for music recommendation. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017, 3654−3660
3. Cheng Z Y, Shen J L, Nie L Q, Chua T S, Kankanhalli M. Exploring user-specific information in music retrieval. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017, 655–664
4. Kim Y E, Schmidt E M, Migneco R, Morton B G, Richardson P, Scott J, Speck J A, Turnbull D. Music emotion recognition: a state of the art review. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. 2010, 255–266
5. Yang Y H, Chen H H. Machine recognition of music emotion: a review. ACM Transactions on Intelligent Systems and Technology. 2011, 3(3): 1−30
6. Bartoszewski M, Kwasnicka H, Kaczmar M U, Myszkowski P B. Extraction of emotional content from music data. In: Proceedings of the 7th International Conference on Computer Information Systems and Industrial Management Applications. 2008, 293–299
7. Hevner K. Experimental studies of the elements of expression in music. The American Journal of Psychology, 1936, 48(2): 246–268
8. Russell J A. A circumplex model of affect. Journal of Personality and Social Psychology, 1980, 39(6): 1161–1178
9. Posner J, Russell J A, Peterson B S. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychology. Development and Psychopathology, 2005, 17(3): 715–734
10. Chekowska-Zacharewicz M, Janowski M. Polish adaptation of the geneva emotional music scale (GEMS): factor structure and reliability. Psychology of Music, 2020, 57(6): 427−438
11. Thayer R. The Biopsychology of Mood and Arousal. 1st ed. Oxford: Oxford University Press, 1989
12. Weninger F, Eyben F, Schuller B W. On-line continuous-time music mood regression with deep recurrent neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2014, 5412−5416
13. Yang Y H, Lin Y C, Su Y F, Chen H H. A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(2): 448–457
14. Li X X, Xianyu H S, Tian J S, Chen W X, Meng F H, Xu M X, Cai L H. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In: Proceedings of the

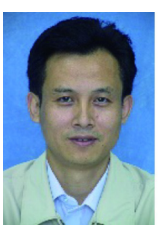
IEEE International Conference on Acoustics, Speech and Signal. 2016, 544–548
15. Fan J Y, Tatar K, Thorogood M, Pasquier P. Ranking-based emotion recognition for experimental music. In: Proceedings of the 18th International Society for Music Information Retrieval Conference. 2017, 368–375
16. Thammasan N, Fukui K I, Numao M. Multimodal fusion of EEG and musical features in music-emotion recognition. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017, 4991−4992
17. Yang Y H, Chen H H. Prediction of the distribution of perceived music emotions using discrete samples. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(7): 2184–2196
18. Liu H P, Fang Y, Huang Q H. Music emotion recognition using a variant of recurrent neural network. In: Proceedings of the International Conference on Matheatics, Modeling, Simulation and Statistics Application. 2018, 15−18
19. Soleymani M, Caro M N, Schmidt E M, Sha C Y, Yang Y H. 1000 songs for emotional analysis of music. In: Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia. 2013, 1–6
20. Turnbull D, Barrington L, Torres D, Lanckriet G. Towards musical query-by-semantic-description using the CAL500 data set. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007, 439–446
21. Wang S Y, Wang J C, Yang Y H, Wang H M. Towards time-varying music auto-tagging on CAL500 expansion. In: Proceedings of the IEEE International Conference on Multimedia and Expo. 2014, 1–6
22. Chen Y A, Yang Y H, Wang J C, Chen H. The AMG1608 dataset for music emotion recognition. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. 2015, 693–697
23. Aljanaki A, Yang Y H, Soleymani M. Developing a benchmark for emotional analysis of music. PLoS ONE, 2017, 12(3): e0173392
24. Speck J A, Schmidt E M, Morton B G, Kim Y E. A comparative study of collaborative vs. traditional musical mood annotation. In: Proceedings of the 12th International Society for Music Informational Retrieval Conference. 2011, 549–554
25. Eerola T, Vuoskoski J K. A comparison of the discrete and dimensional models of emotion in music. Psychology Music, 2011, 39(1): 18–49
26. Zentner M, Grandjean D, Scherer K R. Emotions evoked by the sound of music: characterization, classification, and measurement. Emotion, 2008, 8(4): 494–521
27. Mahieux T B, Ellis D P W, Whitman B, Lamere P. The million songs dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference. 2011, 591–596
28. Tzanetakis G, Cook P. MARSYAS: a framework for audio analysis. Organised Sound, 2000, 4(3): 169–175
29. Mathieu B, Essid S, Fillon T, Prado J, Richard G. YAAFE, an easy to use and efficient audio feature extraction software. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. 2010, 441–446
30. Lartillot O, Toiviainen P. MIR in MATLAB (Ⅱ)A toolbox for musical feature extraction from audio. In: Proceedings of the 8th International Conference on Music Information Retrieval. 2007, 127–130
31. McEnnis D, Mckay C, Fujinaga I, Depalle P. jAudio: a feature extraction library. In: Proceedings of the 6th International Conference on Music Information Retrieval. 2005, 600–603
32. Liu X, Chen Q C, Wu X P, Liu Y, Liu Y. CNN based music emotion classification. 2017, arXiv preprint arXiv: 1704.5665
33. Han W J, Li H F, Ruan H B, Ma Lin. Review on speech emotion recognition (In Chinese). Journal of Software, 2014, 25(1): 37–50
34. Barthet M, Fazekas G, Sandler M. Multidisciplinary perspectives on music emotion recognition: implications for content and context-based model. In: Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval. 2012, 492–507
35. Chen P L, Zhao L, Xin Z Y, Qiang Y M, Zhang M, Li T M. A scheme of MIDI music emotion classification based on fuzzy theme extraction and neural network. In: Proceedings of the 12th International Conference on Computational Intelligence and Security. 2016, 323–326
36. Juslin P N, Laukka P. Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. Journal of New Music Research, 2004, 33(3): 217–238
37. Yang D, Lee W S. Disambiguating music emotion using software agents. In: Proceedings of the 5th International Conference on Music Information Retrieval. 2004, 218–223
38. He H, Jin J M, Xiong Y H, Chen B, Zhao L. Language feature mining for music emotion classification via supervised learning from lyrics. In: Proceedings of International Symposium on Intelligence Computation and Applications. 2008, 426–435
39. Hu X, Downie J S, Ehmann A F. Lyric text mining in music mood classification. In: Proceedings of the 10th International Society for Music Information Retrieval Conference. 2009, 411–416
40. Zaanen M V, Kanters P. Automatic mood classification using TF*IDF based on lyrics. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. 2010, 75–80
41. Wang X, Chen X O, Yang D S, Wu Y Q. Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In: Proceedings of the 12th International Society for Music Information Retrieval Conference. 2011, 765–770
42. Malheiro R, Panda R, Gomes P, Paiva R P. Emotionally-relevant features for classification and regression of music lyrics. IEEE Transactions on Affective Computing, 2018, 9(2): 240–254
43. Hu Y J, Chen X O, Yang D S. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: Proceedings of the 10th International Society for Music Information Retrieval Conference. 2009, 123–128
44. Yang D, Lee W S. Music emotion identification from lyrics. In: Proceedings of the 11th IEEE International Symposium on Multimedia. 2009, 624–629
45. Dakshina K, Sridhar R. LDA based emotion recognition from lyrics. Advanced Computing, Networking and Informatics, 2014, 27(1): 187–194
46. Thammasan N, Fukui K I, Numao M. Application of deep belief networks in EEG-based dynamic music-emotion recognition. In: Proceedings of the 2016 International Joint Conference on Neural Networks. 2016, 881–888
47. Hu X, Li F J, Ng D T J. On the relationships between music-induced emotion and physiological signals. In: Proceedings of the 19th International Society for Music Information Retrieval Conference. 2018, 362–369
48. Nawa N E, Callan D E, Mokhtari P, Ando H, Iversen J. Decoding music-induced experienced emotions using functional magnetic resonance imaging- Preliminary result. In: Proceedings of the 2018 International Joint Conference on Neural Networks. 2018, 1–7
49. Li T, Ogihara M. Detecting emotion in music. In: Proceedings of the 4th International Conference on Music Information Retrieval. 2003, 239–240
50. Laurier C, Grivolla J, Herrera P. Multimodal music mood classification using audio and lyrics. In: Proceedings of the 7th International Conference on Machine Learning and Applications. 2008, 688–693
51. Yang Y H, Lin Y C, Cheng H T, Liao I B, Ho Y C, Chen H. Toward multi-modal music emotion classification. In: Proceedings of the 9th Pacific Rim Conference on Multimedia. 2008, 70–79
52. Liu Y, Liu Y, Zhao Y, Hua K A. What strikes the strings of your heart? – feature mining for music emotion analysis. IEEE Transactions on Affective Computing, 2015, 6(3): 247–260
53. Wang J C, Yang Y H, Wang H M, Jeng S K. The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In: Proceedings of the 20th ACM Multimedia Conference. 2012, 89–98
54. Chen Y A, Wang J C, Yang Y H, Chen H. Component tying for mixture model adaptation in personalization of music emotion recognition. IEEE ACM Transactions on Audio, Speech and Language Processing, 2017,

25(7): 1409–1420
55. Chen Y A, Wang J C, Yang Y H, Chen H. Linear regression-based adaptation of music emotion recognition models for personalization. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. 2014, 2149−2153
56. Fukayama S, Goto M. Music emotion recognition with adaptive aggregation of Gaussian process regressors. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. 2016, 71–75
57. Soleymani M, Aljanaki A, Yang Y H, Caro M N, Eyben F, Markov K, Schuller B, Veltkamp R C, Weninger F, Wiering F. Emotional analysis of music: a comparison of methods. In: Proceedings of the ACM International Conference on Multimedia. 2014, 1161−1164
58. Lu L, Liu D, Zhang H J. Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(1): 5–18
59. Schmidt E M, Turnbull D, Kim Y E. Feature selection for content-based, time-varying musical emotion regression. In: Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval. 2010, 267–274
60. Xianyu H S, Li X X, Chen W S, Meng F H, Tian J S, Xu M X, Cai L H. SVR based double-scale regression for dynamic emotion prediction in music. In: Proceedings of the 2016 IEEE International Conference on Acoustic, Speech and Signal Processing. 2016, 549–553
61. Huang M Y, Rong W G, Arjannikov T, Nan J, Xiong Z. Bi-modal deep Boltzmann machine based musical emotion classification. In: Proceedings of the 25th International Conference on Artificial Neural Network. 2016, 199–207
62. Keelawat P, Thammasan N, Kijsirikul B, Numao M. Subject-independent emotion recognition during music listening based on EEG using deep convolutional neural networks. In: Proceedings of the 2019 the 15th IEEE International Colloquium on Signal Processing & Its Application. 2019, 21–26
63. Sarkar R, Choudhury S, Dutta S, Roy A, Saha S K. Recognition of emotion in music based on deep convolutional neural network. Multimedia Tools and Application, 2020, 79(9): 765–783
64. Yang P T, Kuang S M, Wu C C, Hsu J L. Predicting music emotion by using convolutional neural network. In: Proceedings of the 22nd HCI International Conference. 2020, 266–275
65. Ma Y, Li X X, Xu M X, Jia J, Cai L H. Multi-scale context based attention for dynamic music emotion prediction. In: Proceedings of the 25th ACM International Conference on Multimedia Conference. 2017, 1443−1450
66. Chang W H, Li J L, Lin Y S, Lee C C. A genre-affect relationship network with task-specific uncertainty weighting for recognizing induced emotion in music. In: Proceedings of the 2018 IEEE International Conference on Multimedia and Expo. 2018, 1–8
67. Delbouys R, Hennequin R, Piccoli F, Letelier J R, Moussallam M. Music mood detection based on audio and lyrics with deep neural net. In: Proceedings of the 19th International Society for Music Information Retrieval Conference. 2018, 370–375
68. Dong Y Z, Yang X Y, Zhao X, Li J. Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. IEEE Transactions on Multimedia, 2019, 21(12): 3150–3163
69. Chowdhury S, Vall A, Haunscmid V, Widmer G. Towards explainable music emotion recognition: the route via mid-level features. In: Proceedings of the 20th International Society for Music Information Retrieval Conference. 2019, 237–243
70. Li X X, Tian J S, Xu M X, Ning Y S, Cai L H. DBLSTM-based multi-scale fusion for dynamic emotion prediction in music. In: Proceedings of the IEEE International Conference on Multimedia and Expo. 2016, 1–6
71. Chaki S, Doshi P, Patnaik P, Bhattacharya S. Attentive RNNs for continuous-time emotion prediction in music clips. In: Proceedings of the 3rd Workshop in Affective Content Analysis co-located with 34th AAAI Conference on Artificial Intelligence. 2020, 36–45
72. Panda R, Malheiro R, Paiva R P. Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing, 2020, 11(4): 614–626
73. Deng S G, Wang D J, Li X T, Xu G D. Exploring user emotion in microblogs for music recommendation. Expert System with Applications, 2015, 42(1): 9284–9293
74. Ferreira L N, Whitehead J. Learning to generate music with sentiment. In: Proceedings of the 20th International Society for Music Information Retrieval Conference. 2019, 384–390

Donghong Han received the PhD degree from Northeastern University, China in 2007. She is currently an associate professor with the School of Computer Science and Engineering, Northeastern University, China. She is the reviewer of Applied Intelligence, IEEE Transaction on Cybernetics, Frontiers of Information Technology & Electronic Engineering, etc. She has in total more than 40 publications till date. Her current research interests include data flow management, uncertain data flow analysis and social network sentiment analysis. She is a member of China Computer Federation (CCF), and a member of Chinese Information Processing Society, Social Media Processing.

Yanru Kong received the BS degree from Shandong University of Science and Technology, China in 2018. She is working toward the MS degree in computer science from Northeastern University, China. Her current research interests include natural language processing, sentiment analysis and music emotion recognition.

Jiayi Han received the BS degree from Northeastern University, China in 2018. He is currently pursuing a PhD degree in the Institute of Science and Technology for Brain-Inspired Intelligence at Fudan University, China. His research interests focus on facial expression recognition and medical imaging. He published paper on ICBEB.



Guoren Wang received the PhD degree in computer science from Northeastern University, China in 1996. He is currently a professor with the School of Computer Science & Technology, Beijing Institute of Technology, China. He has published about 300 journal and conference papers. He received The National Science Fund for Distinguished Young Scholars in 2010. His current research interests include uncertain data management, data intensive computing, visual media data management and analysis, unstructured data management, distributed query processing and optimization technology, bioinformatics. He is the vice chairman of China Computer Federation, Technical Committee on Databases (CCF TCDB). And he is an expert review member of National Nature Science Foundation of China, Information Science Department.