Multimedia Systems (2018) 24:365–389
https://doi.org/10.1007/s00530-017-0559-4

REGULAR PAPER

Review of data features-based music emotion recognition methods

Xinyu Yang[1] · Yizhuo Dong[1] · Juan Li[1]

Received: 2 November 2016 / Accepted: 27 June 2017 / Published online: 3 August 2017

Abstract The ability of music to induce or convey emotions ensures the importance of its role in human life. Consequently, research on methods for identifying the high-level emotion states of a music segment from its low-level features has attracted attention. This paper offers new insights on music emotion recognition methods based on different combinations of data features that they use during the modeling phase from three aspects, music features only, ground-truth data only, and their combination, and provides a comprehensive review of them. Then, focusing on the relatively popular methods in which the two types of data, music features and ground-truth data, are combined, we further subdivide the methods in the literature according to the label- and numerical-type ground-truth data, and analyze the development of music emotion recognition with the cue of modeling methods and time sequence. Three current important research directions are then summarized. Although much has been achieved in the area of music emotion recognition, many issues remain. We review these issues and put forward some suggestions for future work.

Keywords Music emotion recognition methods · Music features · Ground-truth data · Emotion models · Review

Communicated by B. Prabhakaran.

✉ Juan Li
lijuan@mail.xjtu.edu.cn

Xinyu Yang
yxyphd@mail.xjtu.edu.cn

Yizhuo Dong
dyzhuo@stu.xjtu.edu.cn

[1] Xi’an Jiaotong University, Xi’an, Shaanxi, People’s Republic of China

1 Introduction

In recent years, the demand for techniques and products that provide affective computing and perceptual interaction capabilities has rendered affective computing more important than ever before. As a branch of affective computing, since music can convey emotion-related information, the study of music emotion has become the focus of researchers and considerable effort has been expended on it during the past decade.

Emotion is the essence of music, and the emotion information of music can be widely used in music retrieval and recommendation. Almost all music pieces are created to convey feelings; composers create music to resonate with their listeners; and performers use the language of music to elicit the emotional responses of audiences [23]. Meanwhile, massive cross-cultural studies of the power of music have indicated that common music psychological and emotional cues exit in music that can transcend the limits of language and achieve cultural infiltration simultaneously [2, 24, 32, 33, 106]. For these reasons, a technique for organizing and retrieving music using an emotion-based approach is feasible, and the core of such technique is the automatic recognition of music emotion information. At present, methods for using the physical, audial, and semantic features of music signals to achieve the automatic recognition of music emotion information have become an important part of the research on digital music applications [17, 112]. The study of music emotion recognition has thus become an extremely urgent issue.

The research area that investigates computation models for detecting the emotion of songs is known as music emotion recognition (MER)[1] [8]. As shown in Fig. 1, MER con-

[1] According to the psychologists’ definition [78], we use the term “music emotion” to represent the audience’s perceptual emotion of music.

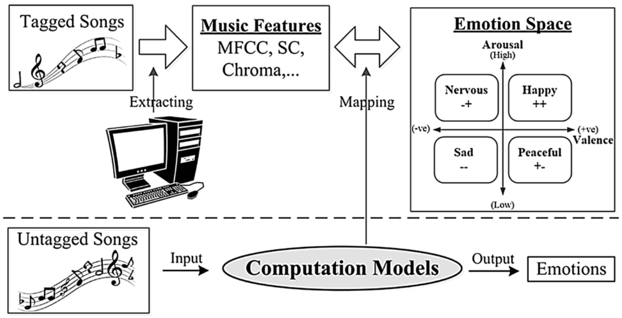


Fig. 1 Music emotion recognition process

stitutes 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; so that music databases can be organized and managed based on emotion [43].

MER models can be used in music information retrieval [19, 91, 104], music recommendation [14, 67], music therapy [6, 16], etc. Since 2007, increasingly successful music emotion classification (MEC) results have emerged in the Music Information Retrieval Evaluation eXchange (MIREX),[2] an annual evaluation of music information retrieval (MIR) algorithms. At the same time, many music Websites and social networks, such as Stereomood[3] and Sensbeat,[4] have used emotion as a cue for music recommendation and user interactions. Moreover, MER has also been applied to devices, such as mobile music emotion players [18, 22], robust personalized affective music players (AMP) [35] and a prototype smart house system [113], so that they can identify the emotion of music automatically and promote human-computer interaction.

A typical MER method comprises three steps [5, 107]. First, in the domain definition step, the target and extent of the problem are defined by choosing the different formats of music records and emotion models. Second, in the feature extraction step, the two types of data features that are commonly used in MER, music features and ground-truth data, are extracted. The ground-truth data are obtained by means of the annotations of the music by subjects based on emotion models, and the music features consist of information that can be extracted from the music. Finally, in the model training process step, machine learning methods are used to establish the mapping relations between the music features and emotions. Thus, automatic MER can be implemented.

MER requires that the mapping relations between music features and emotion space be established and, therefore, some common problems exist [79]. First, it is difficult to determine an appropriate feature extraction standard because of the differences between music records. Second, MER is affected by strong subjectivity, because it involves emotions and has strong correlations with the human character, preference, and other factors. Third, the expressivity and perceptibility of emotions make their classification difficult. The key to solving these three problems is the choice of data features, that is, the choice of music features and ground-truth data.

- First, as the input of MER models, music features are the features extracted from music records that can reflect the style, structure, and emotions of a song, and are directly related to the format of the music records. McKay [58, 59] noted that the formats of music records were divided into two categories: audio and symbol files. Audio files, which store sound content in the form of waves, are called the "low-level" representation of music; while symbol files, which describe sound information in bytes, are called the "high-level" representation of music. We can directly extract low-level and high-level features from audio and symbol files, which represent the specific descriptive and abstraction level characteristics of music, respectively.
- Second, the ground-truth data, which reflect the perceived emotions of human beings [79], are derived from the emotion labels assigned by subjects to the music. We can distinguish the differences between the basic emotion categories. However, experiments have shown that for different people listening to the same songs may produce different emotion perceptions in many cases [26]. Therefore, the ground-truth data are the main factor that affects the results of MER, and the methods of extracting and selecting ground-truth data are the key for reducing the subjectivity of the results. At the same time, ground-truth data are divided into two types, the label- and numerical-type, and the choice of the type of ground-truth data plays an important role in the selection of machine learning methods.
- Finally, a large number of empirical studies showed that the emotion models can be divided into two categories [21], categorical models (such as the typical Hevner model [31] and the updated Hevner model (UHM) [76]), which operate according to the definition of adjectives, and dimensional models (such as the Russell's valence–arousal (VA) model [69, 71], Thayer model [81], and the famous PAD model [60]), which operate according to the definition of dimensions. There is no consensus on the correct choice of emotion models, but it is related to the types of ground-truth data used in experi-

[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME.

[3] https://iTunes.Apple.com/HK/app/sensbeat/id725472587?Mt=8.

[4] https://iTunes.Apple.com/HK/app/stereomood-tuning-my-emotions/id524634435?Mt=8.

ments. We usually choose categorical models to define emotion categories while using label-type ground-truth data as the output of the MER models. However, when numerical-type ground-truth data are used, dimensional models are chosen to represent the music emotion as a point or continuous probability distribution in the emotion space.

In summary, the choice of data features plays a crucial role in MER. However, the complexity of MER and the diversity of research methods make it difficult for researchers to gain a clear understanding of the MER process and, therefore, it is necessary to classify and compare the different methods used in MER based on the different combinations of data features.

Kim et al. [43] conducted extensive research on MER. They presented an overview of the methods based on the context-based text information and the content-based audio information of music features as well as the combinations of them. Similarly, Barthet et al. [4] focused on reviewing MER methods from the content- to the context-based models, and divided music features into the same two categories, namely, content-based features (i.e., various music features such as pitch, timbre, and lyrics) and context-based features (i.e., music metadata, social tags, and others). However, neither of them divided the types of ground-truth data, the second important feature of music, into categories nor investigated their effects on MER methods. In addition, Yang et al. [105] summarized different methods of machine learning, but their description of the features was much shorter. Although it is very important to choose an appropriate machine learning method for MER, the emphasis of MER must remain on the choice of the appropriate features.

Consequently, in this paper, to give researchers a clear understanding of the MER process, our review of the current MER studies is based mainly on the different combinations of data features that they use during the modeling phase, that is, on three aspects: music features only, ground-truth data only, and their combination. More importantly, focusing on a more frequently used type of method in which the two kinds of data are combined, we further subdivide the methods in the literature into those using label- or numerical-type ground-truth data, and analyze the development of MER according to the time sequence and modeling methods.

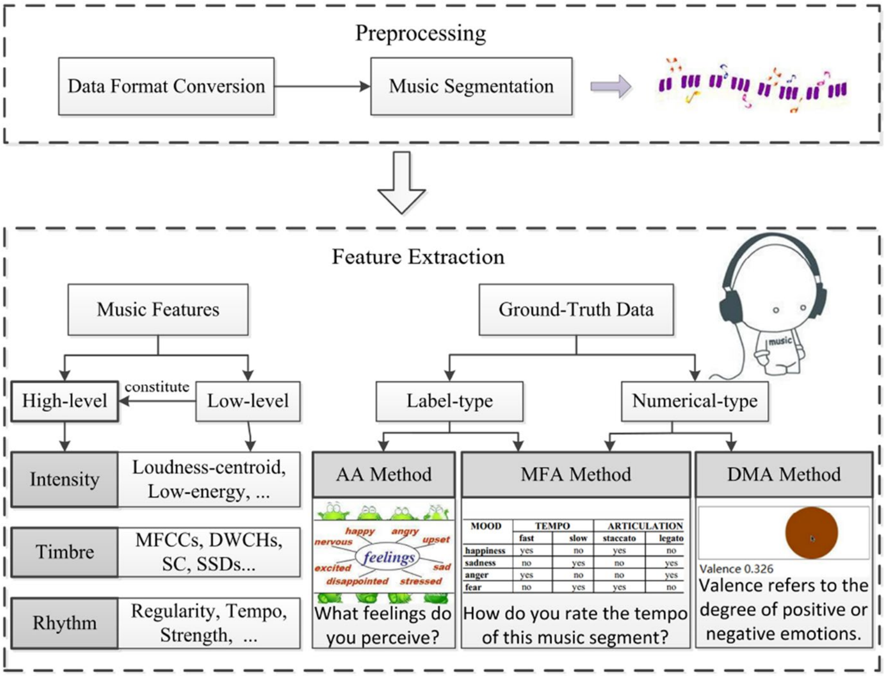
The remainder of this paper is organized as follows. In Sect. 2, we introduce the methods for extracting and selecting music features and ground-truth data. In Sect. 3, the methods in the literature are divided based on the different combinations of data features that they use during the modeling phase and we introduce examples of them. Furthermore, we subdivide the examples of the combination of these two types of data features according to the two types of ground-truth data, classify and compare them with the cue of the time sequence and modeling methods, and discuss the transition, development, and current situation of each stage in MER. Third, the research directions in the field of MER are classified in detail in Sect. 4, and then, we discuss the existing problems and put forward some suggestions for future research directions. Finally, we summarize the studies discussed in this review and the contributions of this paper.

2 Data feature extraction

The goal of data feature extraction is to reduce the information of songs to descriptors that can fully describe them. As mentioned, the data features commonly used in MER can be divided into two categories. One is called music features, which are the acoustic features and the lyrics of the music that can be extracted from music records using the corresponding feature extraction software. The second is called ground-truth data, which are the emotion labels assigned by subjects to the music that truly reflect the perceived emotions of human beings [79]. This section describes the extraction and selection methods of these two types of data features.

As shown in Fig. 2, before the feature extraction process, music records should be converted into a uniform format that can easily be processed, and then, the two types of data features are extracted using specific feature extraction software or annotation methods. The music data preprocessing comprises two steps: data format conversion and music segmentation.

- First, the goal of data format conversion is to convert the audio or symbol files that are used to store music into a consistent format that contains the required information. For example, audio files were converted to a uniform format (22,050 Hz, 16 bits, and mono channel PCM WAV) [107, 110, 114], and Bartoszewski converted MIDI files to a uniform format that contains the pitch, duration, velocity of the voice [5], etc.
- Second, although the music emotion can be predicted by the overall estimation of the entire song, the accuracy of the results is low. The reason for this is that the emotional content of a song fluctuates, and the direct evaluation of the song in its entirety will lead to a large error. Further, the emphasis of subjects when they are listening to music, as well as their familiarity with the music, differ. Therefore, to achieve more accurate results, we need first to split music into small segments, and then detect the emotion separately.



MOOD	TEMPO		ARTICULATION	
	fast	slow	staccato	legato
happiness	yes	no	yes	no
sadness	no	yes	no	yes
anger	yes	no	no	yes
fear	no	yes	yes	no

Fig. 2 The data feature extraction process comprises two steps: (1) preprocessing. (2) Extracting music features from music records and collecting the ground-truth data from subjects. The structure illustrated in the *lower-right* of this figure are three ground-truth data annotation methods: select the list of adjectives (AA method), mark the level of music features (MFA method), and directly label the valence/arousal value (DMA method)

The different music segmentation methods have their own advantages and disadvantages, and the length of the music segments also differs. After considering the influence of the length of segments on emotion recognition, MacDorman and Ho gave a more reasonable criterion [55]: in principle, we hope that the length of segments is as short as possible so that our analysis of the dynamics of the song is as granular as possible. The shorter the segment, the more homogeneous the emotion will be, thus making the evaluation results of subjects more consistent. However, unfortunately, too short a segment is not sufficient to allow the subjects to assess its emotional content accurately. In addition, very short segments in general are separated from their surrounding environment, resulting in a lack of evaluation of their ecological validity.

According to this trade-off criterion, the results of "Appendix" of this paper indicate that the length of the segment for popular music is usually 25–30 s, which corresponds to the typical length of the chorus part of it [91]. Xiao et al. [98] conducted an empirical study on the optimal length of a classical music segment that can express its stable emotional state. In the experiments, 60 unique classical music pieces were selected, which were divided into segments of 4, 8, 16, and 32 s, and each segment was marked by two subjects. The results showed that the classification performance of the system was optimal when the length of the segment was 8–16 s.

In addition to the segmentation method that uses the time length as the basic unit, other methods, such as those based on the structure of the music and those that use per-lyric [95], per-second [79], or frequency [53] segmentation, have been widely used in MER.

2.1 Music features

2.1.1 Music features classification

A large number of previous studies showed that different emotional states are usually associated with different music features [27, 53, 82]. For example, "contentment" is related to a slow tempo, low intensity, and a soft timbre, whereas "exuberance" is related to a fast tempo, high intensity, and a cheerful timbre. At the same time, Gabrielsson et al. [27] and Thayer [82] noted that there are corresponding relations between the dimensional models and music features.

Among these features, intensity is a basic feature, which is highly correlated with arousal and is used to classify the arousal dimension [114]. Its level is obtained by simply measuring the amplitude, and because of its short-time dynamic, the intensity of a song, which consists of some descriptive features, such as the centroid of loudness and low energy rate, in general is extracted in a frame [53]. Meanwhile, as a basic element of music, the timbre describes the sound quality [25] and is represented by features such as mel-frequency cepstrum coefficients (MFCCs), spectrum shape, and spectral contrast (SC). Another frequently used descriptive feature, which is associated with the timbre, is Daubechies wavelet coefficient histograms (DWCHs) [107]. The local and global information of the spectrum can be better expressed using histograms to estimate the probability distribution of it. In addition, rhythm, a long-term feature, is one of the most widely used features in MEC [25]. It describes some of the patterns that appear and repeat in a music segment. There are

Table 1 Illustration of the commonly used descriptive features in MER

No.	Feature	Abbreviation	Dimension	Description
I	Mel-frequency Cepstrum Coefficients	MFCCs	20-D	Low-dimensional representation of the spectrum obtained from the Mel-scale distortion [91]; reflects the nonlinear frequency sensitivity of the human auditory system
II	Octave-based Spectral Contrast	OSC	14-D	Rough representation of the harmonic content of the frequency domain based on the identification of the peak and Valley values of the frequency spectrum [79, 91]
III	Statistical Spectrum Descriptors	SSDs	4-D	Four-dimensional feature composed of the spectral centroid, flux, rolloff and flatness, and usually related to the structure of timbre [79]
IV	Chromagram	Chroma	12-D	An effective method for the estimation of the western pitch components in a short time time interval; essentially a circular version of the log twisted spectrogram [75]
V	Daubechies Wavelet Coefficient Histograms	DWCH	21-D	The mean (7-D), the standard deviation (7-D) and the subband energy (7-D) of spectrum [107]

three types of rhythm feature that are closely related to the human emotion response: rhythm strength, rhythm regularity, and tempo [51, 68].

The above analysis indicates that the abstraction level characteristics of music, such as intensity, timbre and rhythm, are reflected by descriptive features that can be directly extracted from audio files (Fig. 2). As shown in Table 1, some of the most commonly used descriptive features are introduced.

In addition to the music features described above, the genre information, lyric features, etc. of music can also be used to achieve emotion recognition, and were widely used in the studies reported in [7, 12, 50, 100, 102]. For example, the study of Li and Ogihara [50] showed that the genre information of music can improve the accuracy of emotion detection. There are two means of obtaining the genre of music. First, the genre information may be manually annotated and stored in the database can be queried. Second, considerable research has been conducted on automatic genre identification in the field of music, and the methods used are similar to MER methods[13, 29]. One shortcoming of using the genre information is that when it is incorrect it has a negative effect on the results of emotion recognition [87].

Furthermore, the features of song lyrics are very important, since only "saying things" can affect a human's emotion. This type of features was seldom used in earlier studies, because it is quite difficult to analyze the features of lyrics from music segments [87]. Wu et al. [95] conducted a study of the lyrics which related to music emotion and gave the corresponding first three words of these five music emotions. As shown in Table 2, there is no contradiction between the corresponding words of each emotion, and some combinations of these words can frequently be linked to the corresponding emotion in this table. For example the combination of "snow" and "river" can be linked to the "delicate" emotion, the combination of "dream" and "need" can be linked to the "sweet" emotion, etc. In addition, Yang et al. [101] proposed an emotion recognition method based on the latent Dirichlet distribution to recognize the emotion of lyrics. By extracting the uni-gram features of lyrics and using the VA model, they captured the corresponding relations between the features of lyrics and the emotion values. Their experimental results showed that this scheme can effectively improve the accuracy of the regression prediction of the emotion, and the lyrics can be used to predict the emotion of the song. At the same time, combined with music features or ground-truth data [30, 95], analyzing the lyrics can achieve better prediction results.

2.1.2 Feature selection method

The selection of the appropriate features is a difficult task, because researchers are frequently unsure which features are useful, and it is difficult to extract the perfect features.

Table 2 Top three words corresponding to five music emotions [95]

Emotion	Major words (sort by weight)
Delicate	River (0.016), snow (0.013), night (0.012)
Cheerful	Forever (0.019), remember (0.018), happy (0.016)
Restrained	Little (0.016), oceam (0.015), love (0.012)
Rollicking	Saturday (0.023), night (0.021), baby (0.018)
Sweet	Dream (0.022), need (0.021), call (0.017)

Although researchers can formulate a limited hypothesis about which type of features may be suitable based on their experience and knowledge, this may lead to the accuracy of machine learning techniques being limited because of the limitation of their hypothesis [58]. Intuitively, we may think that one means of improving the performance of machine learning is to increase the number of features. However, experiments have shown that, although the performance can thus be improved to a certain extent, using too many features leads to performance degradation [114]. For example, a study of music emotion conducted by Wang et al. [91] showed that the results of using MFCC or contrast features alone in many cases is superior to using them in combination. That is, the automatic emotion annotation results of the acoustic emotion Gaussians (AEG) model did not benefit from using a combination of multiple features. Yang et al. compared the regression results when using the entire feature set and when using the 15 features extracted by the PsySound tool (called the Psy15 features) in regression prediction experiments of music emotion [107, 108], and found that using the Psy15 features could not only greatly reduce the computational complexity, but also avoid serious overfitting problems. It is obvious that an increase in the feature dimension simply cannot improve the performance effectively, and thus, we need a more effective means of selecting the most suitable features.

The Karhunen–Loeve (KL) transform method used by Lu et al. [53] was an effective method to remove the cross-correlative features from original features. They applied it to the features of the training sets in the feature selection step to map the extracted features to an orthogonal space and obtain a new feature space, thereby extracting the unrelated features and reducing the computational complexity. The stepwise backward selection method was used to greedily delete the sequence of the worst features until the accuracy could no longer be improved [109] and in contrast the stepwise forward selection method was used to increase the number of features to find the best combination of the feature sets [33]. Nevertheless, in all the above methods, it is necessary to evaluate the system performance by performing the actual course of testing. It is known that some classifiers are computation-intensive [58] and, therefore, the classification process takes a considerable amount of time to complete. Therefore, there remains a class of dimensionality reduction techniques that select features through the statistical analysis of feature space, such as principal component analysis (PCA) [108, 111] and factor analysis [57]. By forming a linear combination set of the new features that retain the variation of the original features in some fashion, the multi-dimensional data can be mapped into a low-dimensional subspace. In addition, as for the lyrics, by comparing the classification performance of two corpora before and after the application of the syntactic rules, Kim et al. [41] found that the application of the partial syntactic analysis can to a certain extent improve the classification performance of the models.

2.2 Ground-truth data

In early studies, researchers began to explore the relations between music features and ground-truth data in MER. Hevner, who proposed the Hevner circumplex model by dividing the 67 basic emotion adjectives into 8 emotion clusters according to their similarity, detected the emotion values of six features, namely, tone, speed, pitch, rhythm, harmony, and melody, and studied the relations between them and emotion. He used his model to prove the effect and expression of emotions that music can convey in 1936 [31]. In his experiments, the variation in human emotion perception induced by music emotion variables, such as the modality, rhythm, harmony, and melody, were studied. The results showed that there was a clear consistency in the subjects' perception of music, and thus the reliability of the ground-truth data for identifying the emotion was proved.

As shown in Fig. 2, there are three means of collecting ground-truth data: select the list of adjectives, mark the level of music features, or directly label dimensional models, such as the VA model. In this paper, we call the adjectives annotation method (AA), the music features annotation method (MFA), and the dimensional models annotation method (DMA), respectively.

2.2.1 Adjectives annotation method

The AA method is in general performed by means of a questionnaire survey. It requires the subjects to choose the adjectives that are suitable for describing the emotion of the song which they are listening from the list of adjectives in the categorical models (Fig. 2). The results of this annotation method are adjectives, that is, label-type ground-truth data. This method simply takes the emotions as the discrete emotion labels with clear emotional significance, is easy to apply, and is consistent with humans' subjective feeling, which is more conducive to the design of personalized systems. It also has a lower requirement for the professionalism of the subjects in terms of music perception, requiring only that they be able to select the emotion adjectives accurately. The computational complexity can be significantly reduced when the emotional granularity is large and, therefore, it is usually used in MEC, as shown in "Appendix". Nonetheless, because of the ambiguity of the definition of adjectives in the categorical models, it can lead to strong subjectivity in the label-type ground-truth data and low accuracy of the results. That is to say, the same person may have different feelings about the same music segments. Another disadvantage of this method is that the accuracy of classification and the number of emotion categories restrict each other [50], and the definition of emotion categories is an unsolved problem.

At present, there are many public music datasets that were established using the AA method. For example, the audio mood classification (AMC) task of MIREX was the only benchmark in the field of MER for many years [20]. Its aim is to promote the research of MER and provide a baseline for comparison. The creators of the AMC task provided 600 songs for non-commercial use. However, the problem is that the dataset uses five types of discrete emotion for classification, some of them without sources in the psychology literature and others with the semantic and acoustic repetition [46]. This dataset includes only static annotation results, which does not conform to the nature of the time variation of music emotion. In addition, Turnbull et al. [85] established the CAL500 dataset, which consists of 500 popular music songs, each of which has been labeled with adjectives by at least 3 subjects.

2.2.2 *Music features annotation method*

The reason why the dimensional models, such as the VA emotion model [69, 71], are widely used in MER is that this type of model can be used to set up the corresponding relations between music features and emotional states simply and intuitively through the coordinates. Psychologists state that for the VA model, while the arousal dimension is related to tempo (fast or slow), pitch (high or low), loudness level (high or low), and timbre (bright or soft), the valence dimension is related to mode (major or minor) and harmony (consonant or dissonant) [27]. Based on this principle, the MFA method requires the subjects to rate the music features while listening to music, and then obtains the label- or numerical-type ground-truth data according to the correlations between music features and dimensions.

For example, Juslin [37] found two music dimensions, tempo and articulation, which can explain the transfer of emotional content between the performers and audiences, as well as their relations with the four basic human emotions. Based on these, Feng et al. [23] established the emotion annotation rules (Fig. 2) to annotate the emotion of songs by means of tempos (fast or slow) and articulations (staccato or legato). Goyal et al. [28] collected the annotations of the level of music features from the user feedback based on the VA model and used the correlations between the acoustic features (i.e., tempo, pitch, and intensity) and the valence and arousal dimensions, respectively, in their experiments. Thus, they obtained the fuzzy relational feedback values between the perceived emotion and music features of each music segment, to fine-tune the emotion recognition systems further.

Although intuitively it can be seen that the MFA method does not require subjects to annotate the emotions directly, and thus avoids a certain degree of subjectivity, a large error remains in the annotations of the features. On the other hand, this method refers only to the music features that are associated with the dimensions of dimensional models and neglects the other features, which will lead to the impersonal prediction results and an unacceptable emotion output. Consequently, this method is seldom used in practice and a corresponding public dataset has thus far not been provided.

2.2.3 *Dimensional models annotation method*

The DMA method requires that the subjects annotate dimension values in the dimensional models directly, such as arousal and valence values in the VA model, while listening to music [36]. It expresses music as a point in the emotion space, and obtains numerical-type ground-truth data. Thus, this method demands that the professionalism of the subjects in terms of music perception be extremely high, that is, the selection of the subjects is quite strict, and is a time-consuming and labor-intensive process. Nevertheless, it can establish the corresponding relations between the music emotion that human perceived and the emotion models directly, and effectively avoid the ambiguity of the AA method and the limitation of the MFA method. To a great extent, it improves the accuracy of MER and is widely used in MER accordingly. However, numerical-type ground-truth data are focused mainly on the values of music emotion in the emotion space and to a certain degree, the method ignores the correspondence of the human subjective perception. Therefore, we need to further explore how to combine numerical data with label data to ensure both the accuracy of the identification and the human subjective perception.

Soleymani et at. set up a public free MediaEval Database for Emotional Analysis in Music (DEAM)[5] consisting of 1802 songs from 2013 to 2015, which are annotated with emotions using the DMA method, to help researchers in their MER studies. The emotion annotation tool applied in Soleymani's experiment uses an online annotation interface. The dynamics of the valence and arousal values are obtained by sliding the mouse around into interface to indicate the current emotion [79], as shown in Fig. 2. At the same time, to remove some unnecessary subjective factors and obtain more real ground-truth data, Soleymani considered other factors, such as the mood of the subjects at the time of listening

[5] http://cvml.unige.ch/databases/DEAM/.

to the music and the different time-points of the day. The MoodSwings Turk dataset (MTurk) [80] contains 240 pop songs selected from the "uspop2002 database". On the basis of MoodSwings [42], a more original approach was used in it for recruiting subjects to collect emotion labels through Mechanical Turk, an online outsourcing software of Amazon. Each subject was asked to randomly select 11 segments, and then, annotate the second-by-second VA values of these music segments in the VA graphical interface. Next, an automatic verification procedure was used to remove implausible annotations. Finally, each segment was marked by 7–23 subjects. The disadvantage of this dataset is that only the features which extracted from the audio files are public, and if we need other features associated with the music emotion, then this dataset is not applicable. The DEAM and MTurk datasets give the second-by-second emotion values of segments, thus are two currently well known publicly available datasets for music emotion variation detection (MEVD) and have been widely used. In addition, other numerical datasets exist, such as the AMG1608 dataset [10], in which each song has an average of 15–32 numerical-type annotations, the DEAP120 dataset [44], which includes video segments comprising the physiological signals of the subjects, Yang's MER60 dataset [110], and the LJ2M dataset [52].

The statistics in "Appendix" show that the DMA and AA method are two frequently used ground-truth data acquisition approaches. The DMA method increases the annotation burden of the subjects, but it can directly establish the corresponding relations between the music emotion and emotion models and effectively avoid the ambiguity of the AA method. At the same time, the numerical-type ground-truth data can also be easily transformed into the adjectives in the corresponding region of the emotion space (i.e., the label-type ground-truth data) by the dimensional models. In addition, Kim et al. [40] proposed an MEC model that uses emotion labels and VA values based on the VA space. The experimental results showed the feasibility of the combination of label- and numerical-type ground-truth data and contributed to a further improvement in the accuracy of MER.

2.2.4 Selection methods for ground-truth data

Ground-truth data need to be collected carefully because of the subjectivity of emotions, and three problems arise:

- First, different subjects give different emotion annotations and, therefore, each segment needs to be annotated by many subjects [70]. This takes a considerable amount of time and manpower.
- Second, because of the vagueness and inexactness of human annotation, there may be many reasons why undesirable annotations are produced; for example, careless labeling, tired subjects, the failure of some subjects to understand segments, etc. There may also be software reasons, such as that some browsers failed to record the movement of the slider.
- Finally, the biggest problem in content-based music information retrieval is the sharing of music content. Audio files are frequently protected by copyright, although the ground-truth data may be published. The MER research process is being hindered by this problem [103].

To solve the difficulties of emotion annotation, on the one hand many researchers are committed to collecting public datasets, such as MER60 [110], MTurk [80], and AMC [20], to provide reliable and free datasets to researchers in this field. On the other hand, we need to filter the poor data to solve the problem of poor annotations. The specific filtering methods can be divided into two classes:

- The methods in one class select the subjects, that is, the subjects have to pass a test to prove their thorough understanding of the annotation tasks, so that only those with a higher acuity are selected as the formal subjects for emotion annotation [79].
- The methods in the second class find the noise data through a certain filtering method. For example, Speck et al. [80] provided the complete process of using an outlier detection algorithm to remove invalid numerical-type ground-truth data. First, the experts' annotations are used as the baseline, and a classifier (i.e., a one-class SVM) is applied to learn the experts' second-by-second annotations of each music segment, to set up the decision boundary and filter out the outliers. Because the machine cannot achieve a complete and accurate classification, the second step is to compare the trajectory of the subjects' annotations with that of the experts over time. If the similarity of these two trajectories is high, the annotations of subjects are marked "accepted"; otherwise, they are marked "rejected" or "unknown".

In addition, in the process of designing the annotation experiments, it is necessary to focus on reducing the problem of human fatigue, because emotion annotation is a time-consuming and labor-intensive process. Common practices include reducing the length of the music segments [108], using sample songs (their emotions need to be recognized by the majority) to better express the specific meaning of each emotion, allowing the subject to skip a segment when he/she feels that there is no emotion category that can describe its emotional content, designing a user-friendly interface (UI), and using online games, the so-called "Games with a Purpose" [88], to facilitate the collection of the ground-truth data. At present, the typical games are MoodSwings (a numerical-type annotation game) [42, 61, 75], Herd It (a game that combines multiple types of annotations) [3], TagATune (a label-type annotation game) [47], etc. Moreover, it is necessary to place certain restrictions on the time that the subjects spend annotating and the number of annotations that they perform. In general, the time is not more than one hour and the number should not be excessive; as shown in "Appendix", 10–20 segments per subject is optimal.

3 Music emotion recognition methods

An exciting but also the most challenging task in the study of music is to create a computation model that can recognize the emotional content of music signals and organize music databases based on emotions [43]. The essence of such a model is that it uses the music features that are associated with emotion to establish the mapping from the low-level music features to the high-level affective semantic features and realizes the automatic annotation of music emotion. Katayose et al. [39] were the first to achieve the emotion extraction of music based on symbol files in 1988 and Feng et al. [23] were the first to set up an emotion-based music retrieval system for audio files in 2003. Also in 2003, Li et al. [50] treated the emotion detection problem as a multi-label classification problem and divided music into several classes for the first time. Since then, there has been an increasing number of research results in the field of MER, as shown in "Appendix". Based on the different modeling methods, these MER methods can be classified into two categories: classification and regression. In accordance with the different combinations of the data features during the modeling phase, they can be classified into three categories: those that use the different music features only, those that use the ground-truth data only, and those that use the combination of these two types of data features, as shown in Fig. 3:

- The analysis of music features in Sect. 2.1 showed that correlations exit between music features and emotions. For example, while a high note can convey the "excitement" emotion, a low note can convey the "sadness" emotion. When these notes are connected, more powerful and nuanced emotional states can be felt [28]. Therefore, an MER method based on the correlations between music features and emotions is feasible. However, the results of Bartoszewski et al. [5] and Patra et al. [68] showed that it was difficult to accurately recognize music emotion by means of an MER method that used only the music features during the modeling phase (Fig. 3a). The reason for this is that music segments conveying different emotions may have similar music features; for example, music segments conveying the "excited" or "happy" emotion have similar root-mean-square (RMS) and tempo values and, therefore, are likely to be confused in the clustering process and affect the accuracy of classification.
- Intuitively, the results of MER methods that use ground-truth data only without music features must be highly subjective because of the randomness, vagueness, and inexactness of human annotation. In Sect. 2.2, it was mentioned that there are many problems in the acquisition process of ground-truth data and that the main problem is the impact of individual differences [110]. As the ground-truth data are not controllable, there are few MER methods that use only the ground-truth data. Nonetheless, the experimental results of Kim et al. [40] showed the feasibility of combining the label- with the numerical-type ground-truth data, and contributed to the further improvement of the accuracy of MER (Fig. 3b).
- MER is a process that involves using computation models (Fig. 3c, supervised learning) to extract and analyze music features, form the mapping relations between music features and the emotion space, and recognize the emotions that music expresses [43]. Therefore, methods that use the combination of music features and ground-truth data during the modeling phase are most frequently used in MER, and thus, are specially discussed in the following sections.

As mentioned in the introduction, the ground-truth data are classified into the label- and numerical-type, and the choice of the different type of ground-truth data plays an important role in the selection of machine learning methods. We can choose only the classification method when selecting the label-type ground-truth data. This is because the label-type ground-truth data are obtained by selecting the corresponding adjectives in categorical models, and the emotions are considered discrete labels. Instead, we can choose the classification or regression methods when selecting the numerical-type ground-truth data. This is because, on the one hand, we can simply translate the numerical values into labels (e.g. by means of the four quadrants of the VA space) for classification, and on the other hand, we can also realize the mapping between music features and numerical-type ground-truth data to predict the music emotion in the form of numerical-value or continuous probability distribution. Focusing on the case where the ground-truth data and music features are

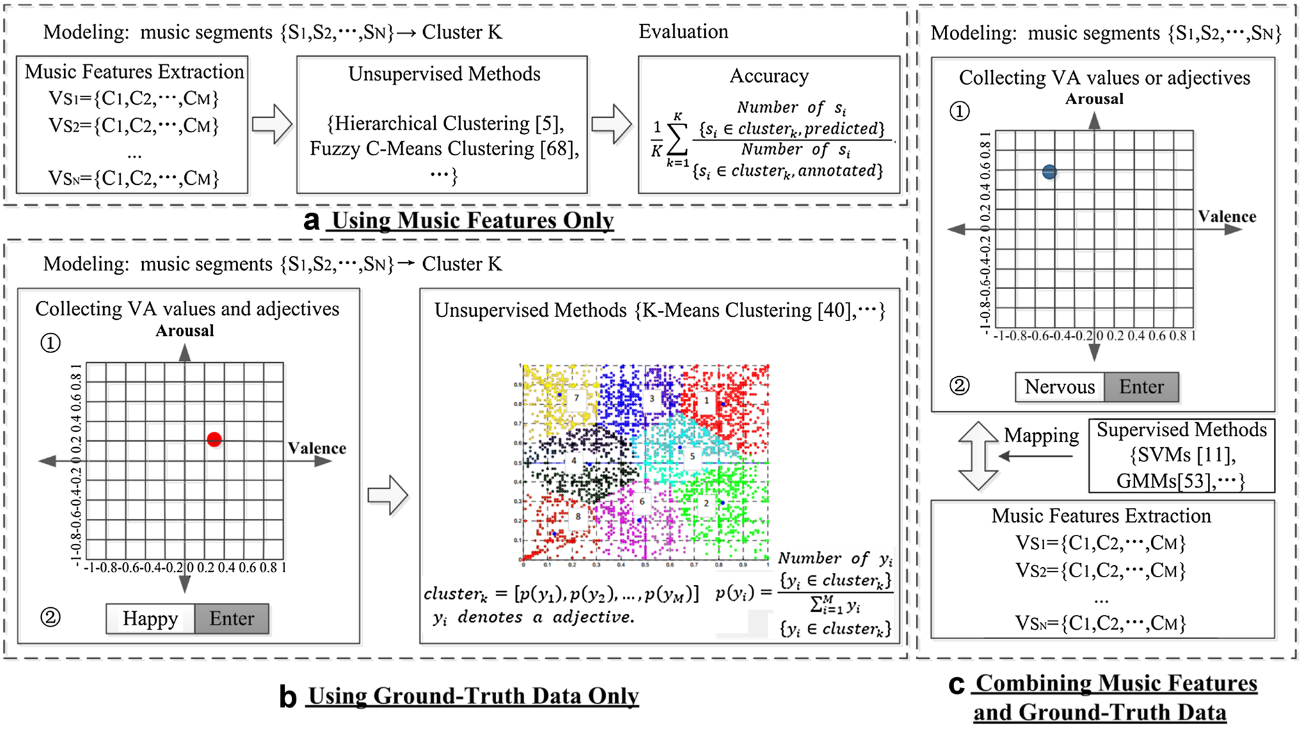

Modeling: music segments {S1,S2,⋯,SN}→ Cluster K
Evaluation
Music Features Extraction
VS1={C1,C2,⋯,CM}
VS2={C1,C2,⋯,CM}
...
VSN={C1,C2,⋯,CM}
Unsupervised Methods
{Hierarchical Clustering [5], Fuzzy C-Means Clustering [68], ⋯}
Accuracy
$\frac{1}{K}\sum_{k=1}^{K}\frac{Number\ of\ s_i\ \{s_i \in cluster_k, predicted\}}{Number\ of\ s_i\ \{s_i \in cluster_k, annotated\}}$
a Using Music Features Only
Modeling: music segments {S1,S2,⋯,SN}→ Cluster K
Collecting VA values and adjectives
①
Arousal
Valence
②
Happy
Enter
Unsupervised Methods {K-Means Clustering [40],⋯}
$cluster_k = [p(y_1), p(y_2), \ldots, p(y_M)]$
y_i denotes a adjective.
$p(y_i) = \frac{Number\ of\ y_i\ \{y_i \in cluster_k\}}{\sum_{i=1}^{M} y_i\ \{y_i \in cluster_k\}}$
b Using Ground-Truth Data Only
Modeling: music segments {S1,S2,⋯,SN}
Collecting VA values or adjectives
①
Arousal
Valence
②
Nervous
Enter
Mapping
Supervised Methods {SVMs [11], GMMs[53],⋯}
Music Features Extraction
VS1={C1,C2,⋯,CM}
VS2={C1,C2,⋯,CM}
...
VSN={C1,C2,⋯,CM}
c Combining Music Features and Ground-Truth Data

Fig. 3 Three different combinations of the data features that MER methods use during the modeling phase. **a** In modeling phase, Bartoszewski et al. [5] and Patra et al. [68] developed the unsupervised classifier for MEC based on the simple audio features. Then, the accuracy was calculated by comparing the predictions with the ground-truth data that subjects annotated in the evaluation phase. **b** In modeling phase, Kim et al. [40] divided the VA values of music segments into eight regions using the k-means clustering algorithm and expressed each region with a possibility distribution of adjectives using statistical methods. Then, the emotion labels of a song can be predicted according to its VA values. **c** Methods that use the combination of music features and ground-truth data during the modeling phase are most frequently used in MER

combined, in this section we further classify and compare MER methods according to the different types of ground-truth data, as shown in Fig. 4.

3.1 Combining label-type ground-truth data and music features

Because of the limitation of the early ground-truth data acquisition methods, which made more use of the AA method, and the universality of classification methods, the method most widely used in studies in the extant literature of MER is the classification method called MEC. The goal of this method is to obtain one or more emotion labels corresponding to a music segment. According to the different number of annotations for each music segment, we can divide the MER classification methods into single-label classification, multi-label classification, and a special case of multi-label classification called fuzzy classification, as shown in Fig. 4.

- Single-label classification expresses the music emotion as a certain single emotion label, that is, the emotion of music segment s_i: $\text{emotion}_{s_i} = y_i$, where y_i is an adjective. This method is simple and intuitive and its computational complexity is low, but the subjectivity and dynamics of human emotion perception are ignored and the classification accuracy is inversely proportional to the number of emotion categories. Therefore, it is gradually being eliminated.
- Multi-label classification is a type of classification method that can be compared with regression methods. While it takes into account the vagueness and inexactness of human annotation and classifies the emotion of music segment into a number of emotion categories: $\text{emotion}_{s_i} = [y_1, y_2, \ldots, y_i]$; fuzzy classification expresses the emotion of music segment as the discrete possibility distribution of a number of emotion categories: $\text{emotion}_{s_i} = [p(y_1), p(y_2), \ldots, p(y_i)]$. As compared with single-label classification methods, the complexity of multi-label classification methods is large; however, the accuracy is improved and the influence of subjectivity on the experimental results can be reduced, as shown in "Appendix".

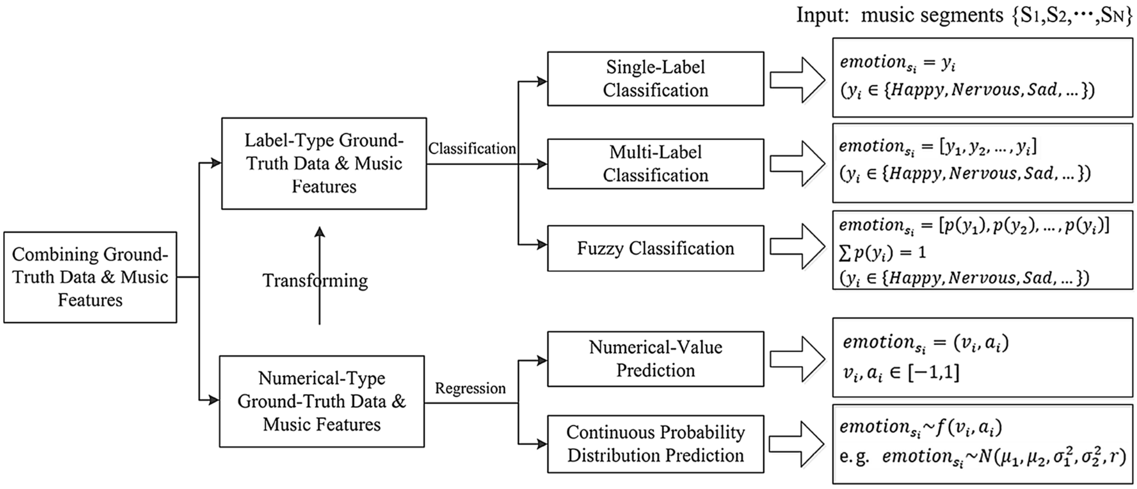

Input: music segments {S1,S2,⋯,SN}
Combining Ground-Truth Data & Music Features
Label-Type Ground-Truth Data & Music Features
Transforming
Numerical-Type Ground-Truth Data & Music Features
Classification
Regression
Single-Label Classification
Multi-Label Classification
Fuzzy Classification
Numerical-Value Prediction
Continuous Probability Distribution Prediction
$emotion_{s_i} = y_i$
$(y_i \in \{Happy, Nervous, Sad, \ldots\})$
$emotion_{s_i} = [y_1, y_2, \ldots, y_i]$
$(y_i \in \{Happy, Nervous, Sad, \ldots\})$
$emotion_{s_i} = [p(y_1), p(y_2), \ldots, p(y_i)]$
$\sum p(y_i) = 1$
$(y_i \in \{Happy, Nervous, Sad, \ldots\})$
$emotion_{s_i} = (v_i, a_i)$
$v_i, a_i \in [-1,1]$
$emotion_{s_i} \sim f(v_i, a_i)$
e.g. $emotion_{s_i} \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, r)$

Fig. 4 MER methods classification according to the different types of ground-truth data

During the past decade, researchers have conducted a considerable amount of research on MEC. Support vector machines (SVMs) [11, 66], *k*-nearest neighbor (KNN) [15, 66], neural networks (ANNs) [23, 83], Bayesian network [97], Gaussian mixture models (GMMs) [53, 89] decision trees (DTs) [48, 54, 63], etc. are the classification methods that are frequently used in MEC. Among these, the SVM is one of the most frequently used methods. It is a supervised learning method that can analyze data and identify patterns, and has advantages in solving small-sample, nonlinear, and high-dimensional pattern recognition problems. The KNN classifier is a type of instance-based classification method and is more frequently used in practice, although its results are not ideal. It is more suitable for automatic classification of larger sample sizes of classes, whereas for smaller sample sizes it is prone to error.

At present, many studies exist in which the performance of classification methods was compared. For example, Pao and Cheng [66] conducted a performance comparison of weighted discrete KNN (W-D-KNN), KNN, and SVM using the same music dataset, MER60 [110]. The results showed that the accuracy of the W-D-KNN classifier was the highest. Kim et al. [41] executed MEC based on the lyrics using three machine learning methods, NaiveBayes (NB), hidden Markov model (HMM), and SVM. The results showed that the classification performance of the SVM method is optimal, reaching 53.6%. In a paper reporting their experiments on the emotion classification of popular Indian music, Ujlambkar et al. [86] presented the order of efficacy of 11 types of classification algorithms sequentially according to the comparison results: NB, SVM, J48, RandomTree and so on.

At the same time, in experiments on MEC methods [48, 63, 86] in which SVM was compared with other classifiers, such as KNN, DT, and GMM, it was found that SVM is usually more efficient than the other classifiers. Therefore, to achieve better classification results, these methods are used in combination with SVM. For example, for each node in the decision tree, Ma et al. [54] used SVM to create a decision hierarchy, which can achieve multi-label classification of music emotion.

In recent years, with the development of deep learning, researchers have applied deep learning methods in MEC and achieved good results. Li et al. [49] as well as Schmidt and Kim [74], used deep belief networks (DBN) to learn the sparse features of music and thus classified music emotions. Chen et al. [8] used the deep Gaussian process (deep GP) to classify music emotions into nine classes in VA space.

3.1.1 Single-label classification

Single-label classification expresses the music emotion as a certain single emotion label. By means of analyzing the music primitives and making rules, Katayose et al. [39] were the first to achieve music emotion extraction based on symbol files. Feng et al. [23] used a two-layer BP neural network to establish an emotion-based music retrieval system based on audio files. As an example of using music features and label-type ground-truth data to establish MER models, Lu et al. proposed [51] and further discussed in detail [53] an MEVD method of the audio signals of classical music. In the detection of music emotion, first the experimental dataset was annotated and filtered by 3 experts. Then, three types of music features, i.e., strength, tone, and rhythm, and a GMM were used to establish a hierarchical framework to focus on different features in different emotion detection tasks. Finally, the detection of four types of music emotion was realized. The results showed that the precision and recall rates of the system reached 67 and 66%, respectively. Inspired by the research of Siegwart and Scherer who used psychology experiments to investigate the emotion that is expressed by singing voices in drama music, Xu et al. [99] found that using singing voices alone may be effective for separating the "calm" from the "sad" emotion, but this effectiveness is lost when the voices are mixed with accompanying music. Therefore, the idea of using source separation to improve the accuracy of MEC and choosing SVM, a baseline classifier that has been proved to have a good classification performance in MIREX, as the classifier of the system were considered. The results showed that source separation can effectively improve the performance of MER from 37.1 to 49.1% and the classification performance for accompanying music (49.1%) is slightly better than that of singing voices (47.6%).

3.1.2 Multi-label classification

The experimental results of Feng [23] and the single-label classification instances, such as those of Xu [99] and others, indicate that the accuracy performance of single-label classification is about 50%. Even when there are sufficient labels, the recognition accuracy is still far from satisfactory, and this has constituted the bottleneck in the study of MER in recent years [105]. As reported in MIREX, in 2011 the best classification accuracy of the five categories of emotion classification tasks had reached 69.5%, and this performance had still not been improved until now. One of the main reasons for this is that single-label classification methods represent the emotion of each song as a deterministic, discrete emotion label, without considering the influences of the ambiguity of the emotion categories and the subjectivity of human annotation. Therefore, multi-label classification was introduced into other theories to support the subjectivity of human evaluation and classify the emotion of music segment into a number of emotion categories, or fuzzy-set and rough-set theories were applied to alleviate the influence of subjectivity. The methods can be classified into two forms. In one form, multi-label classification, several emotion labels are assigned to a music segment at the same time. In the second form, fuzzy classification, the emotion of music segment is expressed as the discrete possibility distribution of a number of emotion categories.

Li et al. [50] was the first to treat the emotion detection problem as a multi-label classification problem and they divided the music into several classes. The selected 499 30-s music segments used in the experiments were labeled by a 39 year-old woman, and they were classified into 13 adjective groups and 6 super-groups by using SVMs. The results showed that, as compared with those of the single-label classification of Feng [23], the precision and recall rates were reduced. This was attributed to the poor acquisition of the label-type ground-truth data; that is, the annotation results of a single subject contain strong subjectivity, and the subject faced many difficult decisions. Therefore, in subsequent studies more annotations were obtained for each segment. Furthermore, the experimental results showed that when more emotion categories were used, the classification accuracy was low, and when the number of emotion categories was reduced, the overall classification accuracy was improved. At the same time, using the genre information can improve the performance of emotion detection.

Later, based on the two types of emotion classification models defined in [50], Wieczorkowska et al. [94] used the KNN algorithm to verify the effects of multi-label classification on recognition accuracy. The accuracy of the 13-category classification was 27.1% ($k = 13$), which was better than the previous experimental results (single-label classification, 20%) and the accuracy of the 6- category classification was 38.62% ($k = 15$). Although the overall classification accuracy was low, the study showed that the results of multi-label classification are more accurate than those of single-label classification.

In addition, Trohidis [84] was the first to present a systematic comparison of multi-label classification algorithms, i.e., Relevance Binary (BR), Powerset Label (LP), Random k-label sets (RAkEL). The results showed that among these three algorithms, the performance of BR is the worst, because it simply transforms a multi-label classification task into multiple single-label classification tasks and disregards the correlations between labels. The LP algorithm takes the correlations between labels into account, and therefore, its performance is better than that of BR. The performance of RAkEL is best, because it is a combination of multiple LPs; its average accuracy reached 79%. Wu et al. [95] proposed a novel hierarchical Bayesian model in which MER is modeled as a multi-label, multi-layer, multi-instance, and multi-view learning problem. In this model, the music is represented as a hierarchical multi-instance structure. From the results of comparing the three single-instance models proposed by Trohidis [84] and the multi-instance M3LDA model proposed by Nguyen et al. [64], it was concluded that the multi-instance model is suitable for MER and the accuracy of multi-label classification can be improved by considering the correlations between multiple labels.

3.1.3 Fuzzy classification

Fuzzy classification is a special case of multi-label classification, which expresses the emotion of a music segment as the discrete possibility distribution of a number of emotion categories. Yang et al. [109] presented one of the first attempts to take the subjectivity of human perception into consideration in MEC, using the fuzzy classification method to classify music emotions. The proposed MEC system combined the music features with label-type ground-truth data, used the fuzzy nearest neighbor classifier (FKNN) and fuzzy nearest mean classifier (FNM) to estimate the weight of each emotion category expressed by fuzzy vectors, and compared them. A 10-fold cross-validation technique was used to evaluate the classification results. It was concluded that the performance of the FNM classifier (78.33%) was better than that of the FKNN classifier (70.88%), and the accuracy of FKNN classification is lower for a sparsely distributed sample. In addition, Myint et al. [62] used a fuzzy support vector machine (FSVM) for MER and Goyal et al. [28] obtained the fuzzy relational feedback values between the perceived emotion and the music features of each music segment by setting up the annotation of the level of music features in the user feedback of the established MER system.

3.2 Combining numerical-type ground-truth data and music features

As well as label-type ground-truth data, numerical-type ground-truth data transformed by means of the coordinates can be used for classification training. Schmidt et al. [75] classified 240 15-s music segments containing multiple VA values, which were collected from the MoodSwings game, into four classes according to the four quadrants of the VA space the SVM classifier. The results showed that the music segments with a quite similar distribution of VA values were divided into completely different emotion categories in reality, and the authors deduced that the main reason for the poor performance of this classification was the strict quantification used for the essentially continuous numerical-type ground-truth data. Thus, they further considered identifying the emotions using a regression method.

As mentioned previously, single-label classification methods suffer problems, such as the ambiguity of emotion categories and the subjectivity of human annotation. Although multi-label classification methods alleviate the subjectivity of human annotation to some extent, they still disregard the ambiguity, individual differences, and time-varying dynamics of music emotion and, therefore, they cannot effectively achieve music retrieval. Yang [109] considered using the geometric relations of the four emotion categories in VA emotion space to calculate the specific VA values of music segments and presented a method for calculating VA values using fuzzy vectors:

$$\text{Valence of } x_u = \{\mu_{1u} + \mu_{4u} - \mu_{2u} - \mu_{3u}\}, \tag{1}$$

$$\text{Arousal of } x_u = \{\mu_{1u} + \mu_{2u} - \mu_{3u} - \mu_{4u}\}, \tag{2}$$

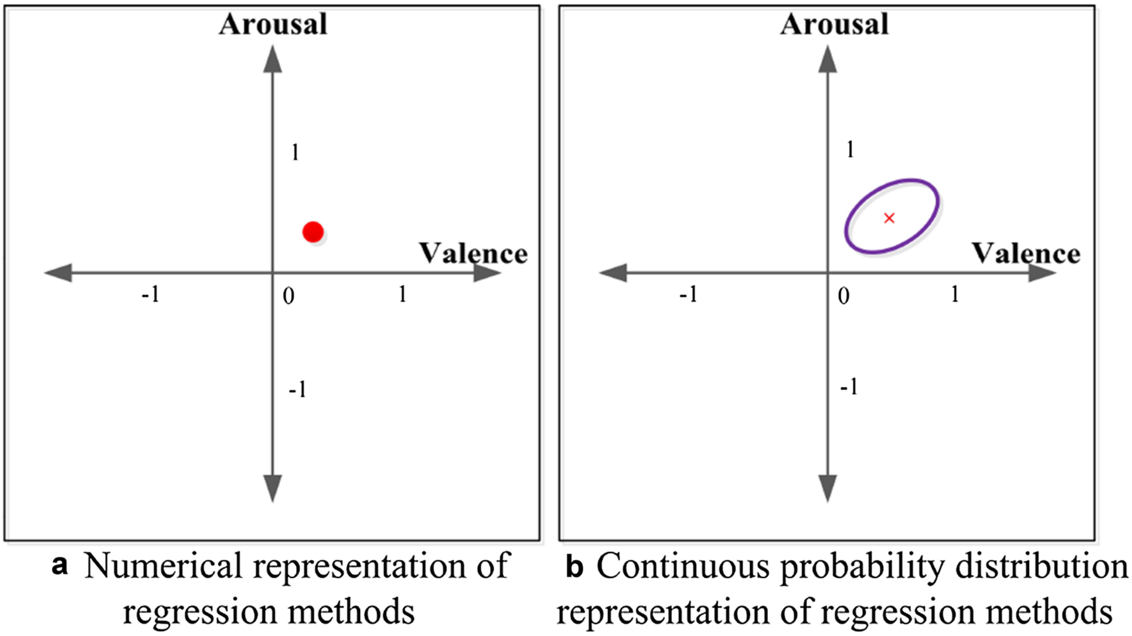
where μ_{yu}, $y = \{1, 2, 3, 4\}$ is defined as the weight of the input sample x_u belonging to each emotion category. However, the theoretical basis of this calculation method is not mature, the transformations between emotion categories are not independent of each other, and the geometric relations between the valence and arousal dimensions are uncertain. Therefore, it is not suitable to use the geometric operations to obtain the VA values.

Dimensional models, such as the Russell's VA model and the Thayer model, in which the regression method is used, can effectively solve these problems. At the same time, research on the prediction of music emotion shows that, in the case of using the same acoustic features, the performance of the parametric regression method is better than that of the label-based classification method [43]. This is why regression methods are used.

The regression method uses the values of the dimensional space to define the emotions, and music features and numerical-type ground-truth data are combined to form the dataset. As shown in Fig. 4, the prediction methods can be divided into two categories according to the different representations.

- One category is numerical-value prediction, which expresses the music emotion as a point that reflects the specific position of emotion in the dimensional space [107, 110]. For example, in the VA space, the emotion of music segment s_i: $\text{emotion}_{s_i} = (v_i, a_i)$, $v_i, a_i \in [-1, 1]$, as shown in Fig. 5a. This avoids the ambiguity of emotion categories, but the prediction results are greatly affected by the subjectivity of human annotation.
- The second category is continuous probability distribution prediction, which expresses the music emotion as a continuous probability distribution in the dimensional space [72, 73, 91]: $\text{emotion}_{s_i} = f(v_i, a_i)$, where $f(\cdot)$ is a bivariate Gaussian probability density function, as shown in Fig. 5b. It can reflect the different emotional experiences of different subjects when listening to the same song, alleviate the influence of the subjectivity of human annotation, and help researchers establish a more personalized music retrieval and recommendation system.

There are many methods to express music emotion as a single, discrete point in the emotion space. Among these, multiple linear regression (MLR) [79, 107] is a standard regression method that is used as a baseline algorithm for MER owing to its relatively low computational complexity and its effectiveness. MLR assumes that there is a



a Numerical representation of regression methods **b** Continuous probability distribution representation of regression methods

Fig. 5 Two representations of the results of regression methods

linear relation between variables and uses the least square estimation method to estimate it. Least squares regression (LSR) [75] is one of the simplest regression techniques in MER. Partial least squares regression (PLSR) [14] is a linear regression method, which has the advantages of PCA, canonical correlation analysis, and multiple linear regression analysis. It is very effective when highly linear correlations exist within variables, and can better solve the problem that the number of samples is less than the number of variables. In contrast, support vector regression (SVR) [14, 34, 107] is a basic nonlinear regression algorithm, which maps the input vector into a high dimensional feature space nonlinearly using the kernel trick and constructs the linear decision function in high-dimensional space to realize the nonlinear regression. SVMs have been found to be very competitive in comparison experiments on existing machine learning methods. Other regression methods, such as the Gaussian process regression [91] and maximum a posteriori linear regression (MAPLR) [9] algorithms, have also been applied in MER. It can be noted that regression methods generate more reliable results and allow quantitative analysis of their performance. In addition, they establish the regressors for the dimensions of the dimensional models such as valence and arousal separately, rather than assuming that there is any geometric relation between them [107, 108], and learn the prediction rules according to the corresponding numerical-type ground-truth data. To display adjectives that inform users how emotion is expressed by a song, the regression results can easily be converted to the emotion labels of the corresponding region in the emotion space.

Thus far, two methods have been developed to express the results of music emotion in the form of continuous probability distribution. One uses the Gaussian RBF kernel of SVM [80], and the second uses the Gaussian probability density function of GMM [72, 73, 91]. For example, Wang et al. [91] proposed a novel acoustic emotion Gaussians model (AEG) based on GMM, which realized a generative process of music emotion perception in a probabilistic and parametric framework. In addition, in many studies, methods combining the numerical values and continuous probability distribution to predict emotions were used. These methods use a regression method (such as SVR or MLR) of the numerical-value prediction to calculate the weight of each component in GMM [80] or the parameters of the Gaussian probability density function [72, 73] to get the corresponding continuous probability distribution further.

3.2.1 Numerical-value prediction

The MER regression methods require that the data features be a combination of the music features and the numerical-type ground-truth data, to allow emotion recognition using the mapping relations between the music features and the numerical values. The move from categorical models of emotion to dimensional models shows the general development trend of MER. Yang et al. was the first to propose [107] a regression method that regards MER as a regression problem and they improved on it [108]. They used three regression methods, MLR, SVR, and AdaBoost.RT, to model the music emotion as a point represented by VA values in the emotion space. The results showed that the best R^2 are obtained using the SVR method. (R^2 statistics is a criterion to measure the performance of the regression algorithms, which is frequently interpreted as the proportion of the corresponding changes described by regressors in the model; a negative R^2 statistics value means that the prediction model is worse than calculating the mean of samples.) The classification accuracy A_c of the SVR reached 84% (arousal dimension) and 68% (valence dimension), and is competitive with that of previous classification methods. In addition, they concluded that arousal values were more easily predicted than valence values [107].

The successful use of regression methods attracted a large number of researchers. Most of the original regression methods expressed the music emotion as a single, deterministic point in the emotion space, such as those of Hu and Yang [32], Yang et al. [101], Neocleous et al. [63], Deng et al. [14], and Schmidt et al. [75]. However, the methods encounter problems similar to those that single-label classification encounter, that is, the music emotion is expressed as a single, deterministic point without considering the influence of the subjectivity, ambiguity, and other factors of human annotation.

The relevant research and the application of music retrieval models presented by Juslin et al. [38] fully showed that the standard of music selection is usually related to the individual's emotional state at the time of the selection. In practice, it has been realized that individual differences or personalization has an important impact on the success of MER systems. However, the impact of personalization was not considered in previous studies and there has been no quantitative assessment of this impact on the performance of MER systems. Thus, Yang et al. [110] proposed two methods to study the impact of individual differences on the performance of MER. They established a group-wise MER system (GWMER) to assess the importance of each individual factor (such as gender, cultural background, and music experience) and a personalized MER system (PMER) to verify whether the elimination of individual differences could in fact improve the prediction accuracy of a system. The comparative results indicate that:

- There was no significant improvement in the performance of the GWMER system as compared to the general MER system, and the influence of individual differences on valence values is greater than that on arousal values. Next, one individual factor was too subtle to reflect the experimental results and, therefore, it is necessary to use a combination of individual factors or other methods to describe the personalization.
- The PMER system compares three types of situation: completely generalized, partially personalized, and completely personalized. It was observed that the valence value prediction performances were significantly improved. Overall, the performance of the partially personalized type is the best, so that the argument that, in reality, personalization plays an important role in MER is supported.

Since then, Yang et al. [92] have proposed a personalized MER system via model adaptation based on the research of personalization, and Janssen et al. [35] have investigated a robust personalized affective music player, which has been widely used.

3.2.2 *Continuous probability distribution prediction*

The fuzzy classification method, which was mentioned previously, expresses the music emotion as a discrete possibility distribution rather than a deterministic label, and can effectively alleviate the impact of the subjectivity of human annotation on the music emotion perception; however, it still cannot avoid the ambiguity of emotion categories. Moreover, whereas using regression methods to predict the emotion of music segment can effectively eliminate the error caused by the ambiguity of the emotional adjectives in categorical models, it produces excessive error when using a point in the emotion space to express the emotion of the corresponding music segment because of the subjectivity of human annotation. Thus, Schmidt et al. [72] proposed a continuous probability distribution representation of music emotion in the time-varying prediction of music emotions. This representation can fit the VA values of the majority of samples in the emotion space. The continuous probability distribution of the music emotion is combined with the advantages of the fuzzy classification and numerical-value prediction, can explain the subjectivity of human emotion perception better and avoid the ambiguity of emotion categories. Therefore, continuous probability distribution representation has become a MER research hotspot in recent years.

Schmidt et al. [72] first used a single two-dimensional Gaussian distribution to fit the second-by-second VA values of a music segment that multiple subjects annotated, and obtained the Gaussian distribution of the ground-truth data in 2010. Furthermore, to establish an automatic filtering system for invalid data, Speck et al. [80] proposed a method that used the Gaussian RBF kernel of SVM to fit the VA values of a music segment that experts had annotated as a continuous probability distribution, to obtain the decision boundary. On the basis of the study above, Wang et al. [91] proposed a novel AEG model based on GMM, which considered the emotion of a music segment as a form of continuous probability distribution, and realized a generative process of music emotion perception in a probabilistic and parametric framework. The experimental datasets were two ready-made music corpora, MER60 [110] and MTurk [80], which included annotated VA values. Two sets of GMMs with the same number of Gaussian distributions were used to fit the data: one, called acoustic GMM, was used to fit the music features, and the second, called VA GMM, was used to fit the VA values that subjects annotated. Finally, an automatic MER system was established to realize the automatic labeling and retrieval of the music emotion.

To improve the proposed AEG model for the concrete application scenes, Wang et al. [90] proposed an acoustic–visual emotion Gaussian model (AVEG) and learned the relations between music, video, and emotion according to a music video corpus, DEAP [44]. The AVEG model is used to predict the continuous probability distributions in the emotion space of a music segment and a video sequence based on its low-level acoustic features respectively, and to match the music and video by measuring the similarity between the two corresponding distributions. In addition, Wang et al. modeled a personalized MER system for the target user in a probabilistic framework based on the maximum a posteriori (MAP) algorithm [92] and the linear regression (LR) adaptive algorithm [9], respectively, on the basis of the AEG model.

4 Discussion

As an emerging research field, MER has received an increasing amount of attention in academic research and for industrial applications. It is aimed to classify music into different kinds of emotion categories, such as pleasant, sad, and angry, so that individual users can select the appropriate song according to their emotional requirements. A comparison of the existing method types of automatic MER, where a combination of music features and ground-truth data is applied, according to the different categories of machine learning methods (i.e., the classification and regression methods) is shown in Table 3. The classification methods can be classified as single-label and multi-label classification according to the number of labels, and the regression method can be classified as numerical-value prediction and continuous probability distribution prediction according to the different representations of the prediction results.

In addition, MEVD or emotion tracking, an extended model based on the classification and regression methods, which can reflect the time-varying dynamics of the music emotion, has received increasing attention in recent years

Table 3 Comparison of the existing method types of automatic MER

Predicted results	Method types	Type of ground-truth data	Emotion models	Description	References[a]
An adjective label	Single-label classification	Label-type	Categorical model	Predicted a single,discrete emotion label of a music segment	[5, 8, 23, 53, 66, 75, 99, 106]
Adjective labels	Multi-label classification	Label-type	Categorical model	Predicted the discrete emotion labels of music segments	[1, 28, 40, 50, 68, 94–96, 109]
Numerical values	Numerical-value prediction	Numerical-type	Dimensional model	Predicted the emotion values of a music segment	[32–34, 45, 56, 65, 75, 77, 79, 107, 108, 110]
Continuous probability distribution	Continuous probability distribution prediction	Numerical-type	Dimensional model	Predicted the continuous probability distributions of music segments	[72, 73, 89–93]
MEVD	Classification and regression prediction	Label-type and numerical-type	Categorical and dimensional models	Detected the variety of music emotion	[34, 45, 53, 56, 65, 72, 73, 75, 77, 108, 109]

[a] References in this table are detailed introduced in "Appendix", including the format of the music records, the length of the music segments, music features that they use, etc

[109]. It was first proposed by Lu et al. [53] based on the single-label classification method. This method searches for emotion boundaries based on the waveforms of three extracted features and divides the entire song into several independent segments, which contain a uniform expression of emotions. Thus, Lu et al. could classify the emotion of each subsegment and further express the emotion variation trend of the song. On the basis of this study, Yang et al. [108] achieved an MEVD system based on regression methods, and represented the music emotion in the VA emotion space in the form of an emotion variation curve. In addition to methods that connect the points in the emotion space to form an emotion variation curve, there is a type of MEVD method, which was investigated in the experiments on the time-varying emotion distribution prediction of audio files conducted by Schmidt and Kim [72], that connects the two-dimensional Gaussian distribution of each subsegment in the VA emotion space in accordance with the time information to obtain the entire emotion variation trend of a segment. As shown in Fig. 6, while red ellipses indicate the Gaussian distribution fitted by the ground-truth data, the blue ellipses indicate the Gaussian distribution which is predicted using the regression method. The shade of both ellipses becomes darker as time progresses. We can see that there is little variation in the arousal dimension, while the emotion of the valence dimension varies from "positive" to "negative". Gabrielsson found that for capturing the continuous variation of emotion, the dimensional representation was better than categorical representation and was more effective [26]. Therefore, thus far MER methods have been based on dimensional models.

Therefore, the three main research directions in the field of MER are MEC, emotion regression prediction, and MEVD. Although considerable effort has been invested, there are still many unsolved issues in MER, such as the relatively low accuracy of the emotion classification, the selection of the emotional granularity [50, 87], the vagueness and inexactness of human annotation, and the low annotation accuracy of valence values [110]. This section summarizes these issues and accordingly discusses future research directions.

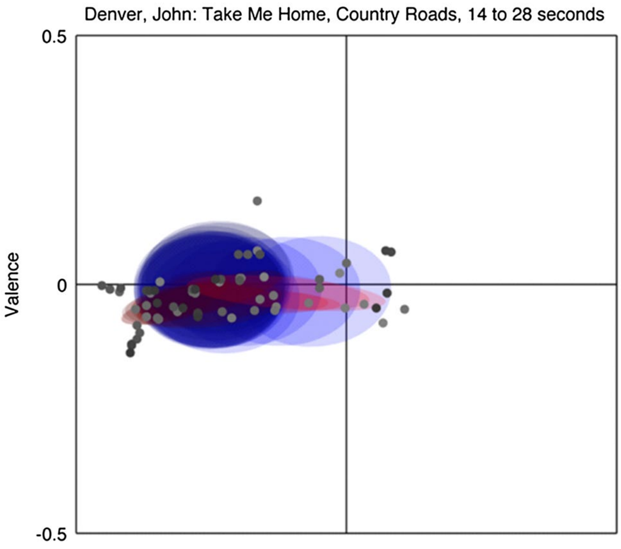


Fig. 6 Emotion variation trend that the continuous probability distribution prediction formed [72]

4.1 Deep learning methods for MEC

MEC is an important branch of emotion recognition. However, as mentioned previously, since 2011 the best classification accuracy of the five categories of emotion classification reached 69.5% in MIREX and this performance had still not been improved until now. And when more categories are used in emotion classification tasks, the classification accuracy is lower. This may be because, on the one hand, the granularity of the emotion classification directly affects to the accuracy of the classification results [50, 87] for the categorical models; that is, although in general the greater the number of categories, the more satisfactory is a method, in this case methods with fewer categories perform better. On the other hand, the classification accuracy is limited by the existing machine learning methods. In recent years, deep learning has contributed significantly to improving the accuracy of multi-class classification using the multi-layer representation and abstract learning and it has been successfully applied in the field of speech recognition, audio processing, and so on. Therefore, according to the excellent characteristics of deep learning in multi-class classification, we can consider using it to improve the classification accuracy of music emotions. Recently, Schmidt and Kim [74] used deep belief networks (DBN) to learn the sparse features of music and Chen et al. [8] employed the deep Gaussian process (deep GP) to achieve the classification of nine music emotions in the VA emotion space. In comparison with the classification results of SVM (67.4% accuracy) and standard Gaussian process (63% accuracy), the deep GP results were better (71.3% accuracy). At the same time, it was shown that there is still plenty of room to improve the accuracy of MEC.

4.2 Combining the label- and numerical-type ground-truth data

The existing MEC methods mostly divide the emotion model simply into four basic groups, or into more groups, for example, six and eight. These methods lack a theoretical foundation and do not consider the size of the datasets, the practicability of music information retrieval, and so on. The previous analysis showed that MER is used mainly as technical support for the emotion-based organization and management of music information retrieval systems. When there are a few emotion categories, although MER can achieve a higher classification accuracy, it cannot easily reflect the subtleties of music emotions and cannot meet the retrieval demands of users. Thus, its practicability is low. However, when more emotion categories are used, it leads to semantic repetition among adjectives and the phenomenon that users cannot distinguish them. Although the introduction of the regression method avoids the problem of selecting the emotional granularity and the music emotion is expressed as a point or continuous probability distribution in the emotion space using dimensional models, it focuses on the values of the music emotion in the emotion space, and ignores the correspondence with human perception. Thus, it leads to the retrieval of a song that humans find unacceptable. Therefore, the selection of the emotional granularity is a problem that urgently needs to be solved.

The selection of the emotion granularity must be based on the corresponding dataset and meet the requirements of practical applications. The method that Kim et al. proposed to establish the corresponding relations between emotion labels and VA values [40], and the relations between the categorical and dimensional emotion semantics of music that Wang et al. explored by combining the label- and numerical-types ground-truth data [89], gives us inspiration concerning the selection of the emotion granularity. Based on the combination of label- and numerical-type ground-truth data, we can obtain the corresponding region of each adjective in the emotion space for different datasets, and clearly define the emotion of the song while locating the position of it in the emotion space, so as to improve the performance of MER in terms of accuracy and practicality.

4.3 Considering the emotional differences between songs for ground-truth data

The collection of ground-truth data is a difficult problem in MER, because on the one hand it takes a considerable amount of time and manpower, and on the other hand, it is difficult to guarantee the accuracy of data because of the subjectivity of human emotion perception. Speck et al. [80] proposed a method in which experts' annotations are used as the baseline to set up the decision boundary and filter out outliers, and they obtained good annotation results. However, this approach requires that each song be annotated by the experts and then marked by general subjects. The establishment of an MER system requires a dataset consisting of a large number of songs and, therefore the cost is high. At present, many music emotion annotation games have been designed and are widely used, such as MoodSwings [42, 61, 75], Herd It [3], and TagATune [47], which can ease the pressure on the subjects and reduce the cost. We can obtain many ground-truth data using them, but much more noise is present in the data and no benchmark exists for data filtering. To solve this problem, Yang and Chen [104] proposed a ranking-based emotion annotation method. By requiring subjects to compare the emotions of pairs of songs and decide which song has a higher arousal or valence value, rather than to give specific emotion values directly, this method can ease the pressure on the subjects and enhance the accuracy of ground-truth data.

According to this solution, we consider the emotional differences between songs as the basis for achieving a combination of these two annotation methods.

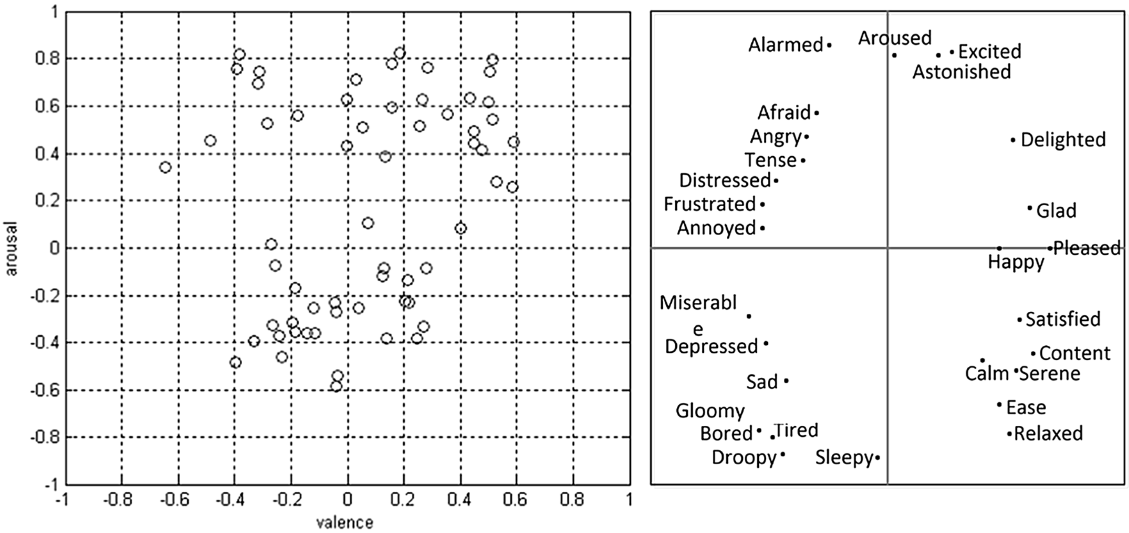
- First, a small number of songs is selected for annotation by experts as the baseline of the system.
- Next, the relations between the features of an unlabeled and a labeled song are compared to represent the relative position of the two songs in the emotion space based on the correspondence between the dimensions and music features.
- Finally, based on the emotional decision boundary of the labeled song, the emotion decision boundary of the unlabeled-song is estimated to rectify the ground-truth data that subjects annotated. This not only reduces costs but also improves the accuracy of the ground-truth data.

4.4 Source separation for the annotation of valence value

The low accuracy of valence value annotation is an unsolved problem in MER. There are two reasons for this phenomenon.

- First, the conclusion has been reached that the arousal dimension gives the measure of the degree of "exciting" or "calming", and there are many features associated with it, such as loudness, tempo, and pitch. While the valence dimension indicates the degree of positive or negative emotion, there are few salient features relevant to it [27].
- Second, the impact of individual differences on the evaluation of valence is greater than that on arousal, and it is a fair possibility that different people perceiving the same song will give an opposite value [108].

In Fig. 7a, which shows the distribution of the ground-truth data, it can be seen that the subjects can easily distinguish the high-arousal and low-arousal songs, because there are two distinct clusters along the arousal dimension. However, most of neutral values are assigned to the valence dimension, that is, the point along the valence dimension is centered around 0. The majority of the neutral annotations are a result of averaging the values, that is, half of the people think the song is positive and half thinks it is negative. The results show that the arousal value is more likely to be predicted accurately than the valence value [110].



a Distribution of the numerical-type ground-truth data[110]

b Russell circumplex emotion model[71]

Fig. 7 Source separation for the annotation of valence value

The research of Siegwart and Scherer, which included psychology experiments on the emotions that are expressed by singing voices in drama music, found that using only singing voices may be effective for separating the "calm" from the "sad" emotion, but the effectiveness is lost when the voice is mixed with the accompanying music [99]. As shown in Fig. 7b, "sad" and "calm" are located on either side of the valence dimension, representing the negative and positive emotions, respectively. Therefore, Xu et al. [99] used the idea of source separation to improve the accuracy of MEC and achieved remarkable results. The experimental results showed that the classification accuracy of the "calm" emotion is increased from 26.1 to 69.6% after source separation, and the error probability of classifying "calm" as "sad" is reduced from 47.8 to 26.1%. Thus, we should consider the idea of separating the accompanying music from the singing voices (or lyrics). Through the emotion recognition of the singing voices and the accompanying music respectively and the more explicit division of positive or negative emotions, the annotation accuracy of the valence dimension can be improved.

5 Conclusion

The organization of music according to emotions is a natural process for human beings, but, because of the ambiguity of the definition of emotion categories, it is a very difficult task for a machine. Therefore, researchers are attempting to establish an automatic emotion recognition system using various methods, and thus, MER is receiving an increasing amount of attention. This paper mainly reviewed the current MER methods based on the different combinations of data features that they use during the modeling phase, that is, from the three aspects of music features only, ground-truth data only, and their combination. We classified and compared the machine learning methods based on different types of ground-truth data in MER. In addition, to provide researchers with a clear understanding of the data features and modeling methods of MER, we summarized and classified the extraction and selection methods of music features and different types of ground-truth data, as well as the existing public datasets, to establish the corresponding relations between data features and the MER models. Finally, we discussed three research directions and the existing problems of MER and put forward some suggestions for future research directions.

Appendix

See Table 4.

Table 4 Summary of the literature in MER (studies between 2003 and 2017)

References	Formats	Genre types	Emotion models	Songs (no.)	Length	Annotation methods	Types of ground-truth	Object types	Annotations (no.)	Music features	Methods
Feng et al. [23]	Audio	Popular	4 classes	223	–	MFA	Label-type	–	1	Tempo and articulation	BP neural network classifier
Li et al. [50]	Audio	Ambient, Jazz, etc	13/6 classes	499	30 s	AA	Label-type	Expert	1	Timbre, rhythmic and pitch	SVM
Yang et al. [109]	Audio	Popular	4 classes (2D-VA)[a]	195	25 s	AA	Label-type	Subjects	Everyone	Psy15	FKNN and FNM
Lu et al. [53]	Audio	Classical	4 classes (2D-VA)[a]	800	20 s	AA	Label-type	Experts	3	Intensity, timbre and rhythm	Hierarchical GMM
Wiec-zorkowska [94]	Audio	Classical	13/6 classes	875	30 s	AA	Label-type	Subjects	–	Timbre	KNN
Yang et al. [107, 108]	Audio	Popular	2D-VA	195	25 s	DMA	Numerical-type	Subjects	10	PsySound[b], Marsyas[b], spectral contrast and DWCH	MLR, SVR and AdaBoost.RT
Yang et al. [110]	Audio	Popular	2D-VA	MER60	25 s	DMA	Numerical-type	Subjects	40	PsySound[b] and Marsyas[b]	SVR
Bartoszewski et al. [5]	MIDI	–	2D-QA	104/70	–	–	–	Subjects	3	Pitch, duration and velocity	Agglomerative clustering and SOM neural network
Pao et al. [66]	Audio	Popular	4 classes (2D-VA)[a]	60	25 s	DMA	Numerical-type	Subjects	40	PsySound[b] and Marsyas[b]	W-D-KNN, K-NN and SVM
Wu et al. [96]	Audio	Sound-tracks	8 classes (UHM)[a]	200	30 s	AA	Label-type	Subjects	28.2	PsySound[b], Marsyas[b], spectral contrast and DWCH	SVM
Schmidt et al. [75]	Audio	Popular	2D-VA	MTurk240	15 s	DMA	Numerical-type	Subjects	7–23	MFCCs, chroma, SSDs and OSC	Classification: SVM Regression: LSR and SVR
Schmidt et al. [72]	Audio	Popular	2D-VA	MTurk240	15 s	DMA	Numerical-type	Subjects	7–23	MFCCs, chroma, SSDs and OSC	MLR, PLR and SVR

Table 4 continued

References	Formats	Genre types	Emotion models	Songs (no.)	Length	Annotation methods	Types of ground-truth	Object types	Annotations (no.)	Music features	Methods
Schmidt et al. [73]	Audio	Popular	2D-VA	MTurk240	15 s	DMA	Numerical-type	Subjects	7–23	OSC	MLR (Kalman filtering and *k*-means)
Kim et al. [40]	Audio	Jazz, R&B, Rock, etc	8 classes (2D-VA)[a]	446	20 s	AA and DMA	Label- and numerical-type	Subjects	10	–	*k*-means clustering algorithm
Wang et al. [91]	Audio	Popular	2D-VA	300	15 s/30 s	DMA	Numerical-type	Subjects	40/(7–23)	MFCCs, chroma, SSDs and OSC	GMM → AEG
Soleymani et al. [79]	Audio	Classical, Rock, etc	2D-VA	1000	45 s	DMA	Numerical-type	Subjects	10	Rastamat toolbox, Ellis and Echonest API	MLR
Patra et al. [68]	Audio	Hindi music	5 classes	250	30 s	–	–	Subjects	5	Rhythm, timbre and intensity	Fuzzy C-means clustering algorithm
Imbrasaite et al. [34]	Audio	Popular	2D-VA	MTurk240	15 s	DMA	Numerical-type	Subjects	7–23	MFCCs, chroma, SSDs and OSC	SVR and CCRF
Goyal et al. [28]	Audio	–	2D-VA	–	–	MFA	Numerical-type	Subjects	–	Tempo, intensity and pitch	BK fuzzy relational interval computing
Xu et al. [99]	Audio	Popular	4 classes (2D-VA)[a]	267	15 s	AA	Label-type	Subjects	–	Timbre, melody, rhythm, etc.	SVM
Wu et al. [95]	Audio	Popular	122 classes	1493	–	AA	Label-type	Experts	15–32	Timbre, pitch, loudness, rhythm and lyrics	Hierarchical bayesian model
Hu et al. [32]	Audio/music video	Popular –	2D-VA	CH496 MER60 DEAP120	30 s 30 s 1 min	DMA	Numerical-type	Experts Subjects Subjects	3 40 14–16	Energy, rhythm, pitch, tonal and timbre	RBF-based SVR
Chen et al. [8]	Audio		9 classes (2D-VA)[a]	1080	1min	AA	Label-type	Experts	15–32	Rhythm, dynamics, timbre, pitch and tonality	Deep GP
Madsen et al. [56]	Audio	Popular	2D-VA	20	15 s	DMA	Numerical-type	Subjects	13	MFCC, chroma and loudness	GP

Table 4 continued

References	Formats	Genre types	Emotion models	Songs (no.)	Length	Annotation methods	Types of ground-truth	Object types	Annotations (no.)	Music features	Methods
Ahsan et al. [1]	Audio	Classical, Rock, etc	6 classes	100	30 s	AA	Label-type	Subjects	–	Rhythmic and timbre	KNN, SVM, ML-KNN and M3L
Hu et al. [33]	Audio	Popular	2D-VA	MER60 CH818 AMG1608	30 s 30 s 1 min	DMA	Numerical-type	Subjects Experts Subjects	40 3 15–32	Loudness, pitch, rhythm, timbre and harmony	RBF-based SVR

[a] Emotion categories that are obtained according to 2D-VA or UHM emotion models

[b] Music features that are extracted by PsySound or Marsyas tools

References

1. Ahsan, H., Kumar, V., Jawahar, C.: Multi-label annotation of music. 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–5. IEEE, New York (2015)
2. Balkwill, L.L., Thompson, W.F.: A cross-cultural investigation of the perception of emotion in music: psychophysical and cultural cues. Music Percept. Interdiscip. J. **17**(1), 43–64 (1999)
3. Barrington, L., O'Malley, D., Turnbull, D., Lanckriet, G.: User-centered design of a social game to tag music. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 7–10. ACM, New York (2009)
4. Barthet, M., Fazekas, G., Sandler, M.: Music emotion recognition: from content-to context-based models. In: International Symposium on Computer Music Modeling and Retrieval, pp. 228–252. Springer, Berlin (2012)
5. Bartoszewski, M., Kwasnicka, H., Markowska-Kaczmar, U., Myszkowski, P.B.: Extraction of emotional content from music data. In: Computer Information Systems and Industrial Management Applications, 2008. CISIM'08. 7th, pp. 293–299. IEEE, New York (2008)
6. Bernatzky, G., Presch, M., Anderson, M., Panksepp, J.: Emotional foundations of music as a non-pharmacological pain management tool in modern medicine. Neurosci. Biobehav. Rev. **35**(9), 1989–1999 (2011)
7. Chapaneri, S., Lopes, R., Jayaswal, D.: Evaluation of music features for PUK kernel based genre classification. Procedia Comput. Sci. **45**, 186–196 (2015)
8. Chen, S.H., Lee, Y.S., Hsieh, W.C., Wang, J.C.: Music emotion recognition using deep gaussian process. In: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 495–498. IEEE, New York (2015)
9. Chen, Y.A., Wang, J.C., Yang, Y.H., Chen, H.: Linear regression-based adaptation of music emotion recognition models for personalization. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2149–2153. IEEE, New York (2014)
10. Chen, Y.A., Yang, Y.H., Wang, J.C., Chen, H.: The amg1608 dataset for music emotion recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 693–697. IEEE, New York (2015)
11. Chin, Y.H., Lin, C.H., Siahaan, E., Wang, I.C., Wang, J.C.: Music emotion classification using double-layer support vector machines. In: 2013 International Conference on Orange Technologies (ICOT), pp. 193–196. IEEE, New York (2013)
12. Chin, Y.H., Lin, P.C., Tai, T.C., Wang, J.C.: Genre based emotion annotation for music in noisy environment. In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, pp. 863–866. IEEE, New York (2015)
13. Corrêa, D.C., Rodrigues, F.A.: A survey on symbolic data-based music genre classification. Expert Syst. Appl. **60**, 190–210 (2016)
14. Deng, J.J., Leung, C.H., Milani, A., Chen, L.: Emotional states associated with music: classification, prediction of changes, and consideration in recommendation. ACM Trans. Interact. Intell. Syst. (TiiS) **5**(1), 1–36 (2015)
15. Dewi, K.C., Harjoko, A.: Kid's song classification based on mood parameters using k-nearest neighbor classification method and self organizing map. In: 2010 International Conference on Distributed Framework and Applications (DFmA), pp. 1–5. IEEE, New York (2010)
16. Dingle, G.A., Kelly, P.J., Flynn, L.M., Baker, F.A.: The influence of music on emotions and cravings in clients in addiction

treatment: a study of two clinical samples. Arts Psychother. **45**, 18–25 (2015)

17. Dobashi, A., Ikemiya, Y., Itoyama, K., Yoshii, K.: A music performance assistance system based on vocal, harmonic, and percussive source separation and content visualization for music audio signals. In: Proceedings of SMC, pp. 99–104 (2015)
18. Dornbush, S., Fisher, K., McKay, K., Prikhodko, A., Segall, Z.: Xpod-a human activity and emotion aware mobile music player. In: 2005 2nd Asia Pacific Conference on Mobile Technology, Applications and Systems, pp. 1–6. IEEE, New York (2005)
19. Downie, J.S.: The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. Acoust. Sci. Technol. **29**(4), 247–255 (2008)
20. Downie, X., Laurier, C., Ehmann, M.: The 2007 mirex audio mood classification task: Lessons learned. In: Proc. 9th Int. Conf. Music Inf. Retrieval, pp. 462–467 (2008)
21. Eerola, T., Vuoskoski, J.K.: A comparison of the discrete and dimensional models of emotion in music. Psychol. Music **39**(1), 18–49 (2011)
22. Fan, S., Tan, C., Fan, X., Su, H., Zhang, J.: Heartplayer: a smart music player involving emotion recognition, expression and recommendation. In: International Conference on Multimedia Modeling, pp. 483–485. Springer, Berlin (2011)
23. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 375–376. ACM, New York (2003)
24. Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., Koelsch, S.: Universal recognition of three basic emotions in music. Curr. Biol. **19**(7), 573–576 (2009)
25. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimed. **13**(2), 303–319 (2011)
26. Gabrielsson, A.: Emotion perceived and emotion felt: same or different? Musicae Scientiae **5**(1 suppl), 123–147 (2002)
27. Gabrielsson, A., Lindström, E.: The influence of musical structure on emotional expression. Oxford University Press (2001)
28. Goyal, S., Kim, E.: Application of fuzzy relational interval computing for emotional classification of music. In: 2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW), pp. 1–8. IEEE, New York (2014)
29. Grimaldi, M., Cunningham, P.D., Kokaram, A.: Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification. Multimed. Syst. **11**(5), 422–437 (2006)
30. He, H., Chen, B., Guo, J.: Emotion recognition of pop music based on maximum entropy with priors. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 788–795. Springer, Berlin (2009)
31. Hevner, K.: Experimental studies of the elements of expression in music. Am. J. Psychol. **48**(2), 246–268 (1936)
32. Hu, X., Yang, Y.H.: Cross-cultural mood regression for music digital libraries. In: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 471–472. IEEE Press, New York (2014)
33. Hu, X., Yang, Y.H.: Cross-dataset and cross-cultural music mood prediction: a case on western and Chinese pop songs. IEEE Trans. Affect. Comput. **8**(2), 228–240 (2017)
34. Imbrasaitė, V., Baltrušaitis, T., Robinson, P.: Emotion tracking in music using continuous conditional random fields and relative feature representation. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6. IEEE, New York (2013)
35. Janssen, J.H., van den Broek, E.L., Westerink, J.H.: Tune in to your emotions: a robust personalized affective music player. User Model. User-Adapt. Interact. **22**(3), 255–279 (2012)
36. Jun, S., Rho, S., Han, B.J., Hwang, E.: A fuzzy inference-based music emotion recognition system. In: 5th International Conference on Visual Information Engineering, 2008. VIE 2008, pp. 673–677. IET, Stevenage (2008)
37. Juslin, P.N.: Cue utilization in communication of emotion in music performance: relating performance to perception. J. Exp. Psychol. Hum. Percept. Perform. **26**(6), 1797–1812 (2000)
38. Juslin, P.N., Sloboda, J.A.: Music and emotion: theory and research. Oxford University Press (2001)
39. Katayose, H., Imai, M., Inokuchi, S.: Sentiment extraction in music. In: 9th International Conference on Pattern Recognition, 1988, pp. 1083–1087. IEEE, New York (1988)
40. Kim, J., Lee, S., Kim, S., Yoo, W.Y.: Music mood classification model based on arousal–valence values. In: 2011 13th International Conference on Advanced Communication Technology (ICACT), pp. 292–295. IEEE, New York (2011)
41. Kim, M., Kwon, H.C.: Lyrics-based emotion classification using feature selection by partial syntactic analysis. In: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, pp. 960–964. IEEE, New York (2011)
42. Kim, Y.E., Schmidt, E.M., Emelle, L.: Moodswings: a collaborative game for music mood label collection. ISMIR **8**, 231–236 (2008)
43. 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: Proc. ISMIR, pp. 255–266. Citeseer (2010)
44. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
45. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. Eighteenth Int. Conf. Mach. Learn. ICML **1**, 282–289 (2001)
46. Laurier, C., Herrera, P., Mandel, M., Ellis, D.: Audio music mood classification using support vector machine. MIREX Task Audio Mood Classif. 2–4 (2007)
47. Law, E.L., Von Ahn, L., Dannenberg, R.B., Crawford, M.: Tagatune: a game for music and sound annotation. In: ISMIR, vol. 3, p. 2 (2007)
48. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9), 1162–1171 (2011)
49. Li, J., Gao, S., Han, N., Fang, Z., Liao, J.: Music mood classification via deep belief network. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1241–1245. IEEE, New York (2015)
50. Li, T., Ogihara, M.: Detecting emotion in music. ISMIR **3**, 239–240 (2003)
51. Liu, D., Lu, L., Zhang, H.J.: Automatic mood detection from acoustic music data. In: Proceedings of Ismir, pp. 81–87 (2003)
52. Liu, J.Y., Liu, S.Y., Yang, Y.H.: Lj2m dataset: toward better understanding of music listening behavior and user mood. In: 2014 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, New York (2014)
53. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. Audio Speech Lang. Process. **14**(1), 5–18 (2006)

54. Ma, A., Sethi, I., Patel, N.: Multimedia content tagging using multilabel decision tree. In: International Symposium on Multimedia, pp. 606–611 (2009)
55. MacDorman, K.F., Ho, C.C.: Automatic emotion prediction of song excerpts: index construction, algorithm design, and empirical comparison. J. N. Music Res. **36**(4), 281–299 (2007)
56. Madsen, J., Jensen, B.S., Larsen, J.: Learning combinations of multiple feature representations for music emotion prediction. In: Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, pp. 3–8. ACM, New York (2015)
57. Mcadams, S., Giordano, B.L., Mcadams, S., Giordano, B.L.: The oxford handbook of music psychology. Jew. Q. Rev. **11**, 72–80 (2009)
58. McKay, C.: Automatic genre classification of midi recordings. Ph.D. thesis, McGill University (2004)
59. McKay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets. In: ISMIR, vol. 2004, pp. 525–530. Citeseer (2004)
60. Mehrabian, A.: Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr. Psychol. **14**(4), 261–292 (1996)
61. Morton, B.G., Speck, J.A., Schmidt, E.M., Kim, Y.E.: Improving music emotion labeling using human computation. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 45–48. ACM, New York (2010)
62. Myint, E.E.P., Pwint, M.: An approach for multi-label music mood classification. In: 2010 2nd International Conference on Signal Processing Systems (ICSPS), vol. 1, pp. V1–290. IEEE, New York (2010)
63. Neocleous, A., Ramirez, R., Perez, A., Maestre, E.: Modeling emotions in violin audio recordings. In: Proceedings of 3rd International Workshop on Machine Learning and Music, pp. 17–20. ACM, New York (2010)
64. Nguyen, C.T., Zhan, D.C., Zhou, Z.H.: Multi-modal image annotation with multi-instance multi-label lda. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1558–1564. AAAI Press, Cambridge (2013)
65. Panda, R., Paiva, R.P.: Using support vector machines for automatic mood tracking in audio music. Audio Engineering Society Convention, pp. 1–8. Audio Engineering Society, New York (2011)
66. Pao, T.L., Cheng, Y.M., Yeh, J.H., Chen, Y.T., Pai, C.Y., Tsai, Y.W.: Comparison between weighted d-knn and other classifiers for music emotion recognition. In: 3rd International Conference on Innovative Computing Information and Control, 2008. ICICIC'08, pp. 530–530. IEEE, New York (2008)
67. Park, S.H., Ihm, S.Y., Jang, W.I., Nasridinov, A., Park, Y.H.: A music recommendation method with emotion recognition using ranked attributes. Computer Science and its Applications, pp. 1065–1070. Springer, Berlin (2015)
68. Patra, B.G., Das, D., Bandyopadhyay, S.: Unsupervised approach to hindi music mood classification. Mining Intelligence and Knowledge Exploration, pp. 62–69. Springer, Berlin (2013)
69. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**(03), 715–734 (2005)
70. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. J. Mach. Learn. Res. **11**, 1297–1322 (2010)
71. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161–1178 (1980)
72. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions from audio. In: International Society for Music Information Retrieval Conference, Ismir 2010, Utrecht, The Netherlands, August, pp. 465–470 (2010)
73. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions using kalman filtering. In: 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 655–660. IEEE, New York (2010)
74. Schmidt, E.M., Kim, Y.E.: Learning emotion-based acoustic features with deep belief networks. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 65–68. IEEE, New York (2011)
75. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 267–274. ACM, New York (2010)
76. Schubert, E.: Update of the Hevner adjective checklist. Percept. Motor Skills **96**(3 suppl), 1117–1122 (2003)
77. Schubert, E.: Modeling perceived emotion with continuous musical features. Music Percept. Interdiscipl. J. **21**(4), 561–585 (2004)
78. Sloboda, J.A., Juslin, P.N.: Psychological perspectives on music and emotion. Oxford University Press (2001)
79. 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, pp. 1–6. ACM, New York (2013)
80. Speck, J.A., Schmidt, E.M., Morton, B.G., Kim, Y.E.: A comparative study of collaborative vs. traditional musical mood annotation. In: ISMIR, pp. 549–554. Citeseer (2011)
81. Thayer, R.E.: Toward a psychological theory of multidimensional activation (arousal). Motiv. Emot. **2**(1), 1–34 (1978)
82. Thayer, R.E., McNally, R.J.: The biopsychology of mood and arousal. Cognit. Behav. Neurol. **5**(1), 65 (1992)
83. Tomo, T.P., Enriquez, G., Hashimoto, S.: Indonesian puppet theater robot with gamelan music emotion recognition. In: 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1177–1182. IEEE, New York (2015)
84. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. ISMIR **8**, 325–330 (2008)
85. 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, pp. 439–446. ACM, New York (2007)
86. Ujlambkar, A.M., Attar, V.Z.: Mood classification of Indian popular music. In: Proceedings of the CUBE International Information Technology Conference, pp. 278–283. ACM, New York (2012)
87. Van De Laar, B.: Emotion detection in music, a survey. Twente Stud. Conf. IT **1**, 1–7 (2006)
88. Von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006)
89. Wang, J.C., Yang, Y.H., Chang, K., Wang, H.M., Jeng, S.K.: Exploring the relationship between categorical and dimensional emotion semantics of music. In: Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies, pp. 63–68. ACM, New York (2012)
90. Wang, J.C., Yang, Y.H., Jhuo, I.H., Lin, Y.Y., Wang, H.M., et al.: The acousticvisual emotion guassians model for automatic generation of music video. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 1379–1380. ACM, New York (2012)
91. 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 International Conference on Multimedia, pp. 89–98. ACM, New York (2012)

92. Wang, J.C., Yang, Y.H., Wang, H.M., Jeng, S.K.: Personalized music emotion recognition via model adaptation. In: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pp. 1–7. IEEE, New York (2012)
93. Wang, J.C., Yang, Y.H., Wang, H.M., Jeng, S.K.: Modeling the affective content of music with a gaussian mixture model. IEEE Trans. Affect. Comput. **6**(1), 56–68 (2015)
94. Wieczorkowska, A., Synak, P., Raś, Z.W.: Multi-label classification of emotions in music. In: Intelligent Information Processing and Web Mining, pp. 307–315. Springer, Berlin (2006)
95. Wu, B., Zhong, E., Horner, A., Yang, Q.: Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 117–126. ACM, New York (2014)
96. Wu, T.L., Jeng, S.K.: Probabilistic estimation of a novel music emotion model. In: International Conference on Multimedia Modeling, pp. 487–497. Springer, Berlin (2008)
97. Wu, W., Xie, L.: Discriminating mood taxonomy of Chinese traditional music and western classical music with content feature sets. In: Congress on Image and Signal Processing, 2008. CISP'08, vol. 5, pp. 148–152. IEEE, New York (2008)
98. Xiao, Z., Dellandrea, E., Dou, W., Chen, L.: What is the best segment duration for music mood analysis? In: 2008 International Workshop on Content-Based Multimedia Indexing, pp. 17–24. IEEE, New York (2008)
99. Xu, J., Li, X., Hao, Y., Yang, G.: Source separation improves music emotion recognition. In: Proceedings of International Conference on Multimedia Retrieval, pp. 423–426. ACM, New York (2014)
100. Xue, H., Xue, L., Su, F.: Multimodal music mood classification by fusion of audio and lyrics. In: International Conference on Multimedia Modeling, pp. 26–37. Springer, Berlin (2015)
101. Yang, D., Chen, X., Zhao, Y.: A lda-based approach to lyric emotion regression. Knowledge Engineering and Management, pp. 331–340. Springer, Berlin (2011)
102. Yang, D., Lee, W.S.: Disambiguating music emotion using software agents. ISMIR **4**, 218–223 (2004)
103. Yang, Y.H., Chen, H.H.: Music Emotion Recognition. CRC Press, Boca Raton (2011)
104. Yang, Y.H., Chen, H.H.: Ranking-based emotion recognition for music organization and retrieval. IEEE Trans. Audio Speech Lang. Process. **19**(4), 762–774 (2011)
105. Yang, Y.H., Chen, H.H.: Machine recognition of music emotion: a review. ACM Trans. Intell. Syst. Technol. (TIST) **3**(3), 40 (2012)
106. Yang, Y.H., Hu, X.: Cross-cultural music mood classification: a comparison on English and Chinese songs. Int. Soc. Music Inf. Retr. Conf. Ismir **2012**, 19–24 (2012)
107. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: Music emotion classification: a regression approach. In: IEEE International Conference on Multimedia and Expo, pp. 208–211. IEEE, New York (2007)
108. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: A regression approach to music emotion recognition. IEEE Trans. Audio Speech Lang. Process. **16**(2), 448–457 (2008)
109. Yang, Y.H., Liu, C.C., Chen, H.H.: Music emotion classification: a fuzzy approach. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 81–84. ACM, New York (2006)
110. Yang, Y.H., Su, Y.F., Lin, Y.C., Chen, H.H.: Music emotion recognition: the role of individuality. In: Proceedings of the International Workshop on Human-Centered Multimedia, pp. 13–22. ACM, New York (2007)
111. Yazdani, A., Kappeler, K., Ebrahimi, T.: Affective content analysis of music video clips. In: Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies, pp. 7–12. ACM, New York (2011)
112. Yoon, J.H., Cho, J.H., Lee, J.K., Lee, H.E., et al.: Method and apparatus for visualizing music information. US Patent 20,160,035,323 (2016)
113. Yu, Y.C., You, S.D., Tsai, D.R.: Magic mirror table for social-emotion alleviation in the smart home. IEEE Trans. Consum. Electron. **58**(1), 126–131 (2012)
114. Zhang, J.L., Huang, X.L., Yang, L.F., Xu, Y., Sun, S.T.: Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods. Multimedia Syst. **23**(2), 251–264 (2017)