Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models

Mathieu Barthet, György Fazekas, and Mark Sandler

Centre for Digital Music
Queen Mary University of London
{mathieu.barthet,gyorgy.fazekas,mark.sandler}@eecs.qmul.ac.uk

Abstract. The prominent status of music in human culture and every day life is due in large part to its striking ability to elicit emotions, which may manifest from slight variation in mood to changes in our physical condition and actions. In this paper, we first review state of the art studies on music and emotions from different disciplines including psychology, musicology and music information retrieval. Based on these studies, we then propose new insights to enhance automated music emotion recognition models.

Keywords: music emotion recognition, mood, metadata, appraisal model

1 Introduction

Since the first empirical works on the relationships between music and emotions [20] [37], a large body of research studies has given strong evidence towards the fact that music can either (i) elicit/induce/evoke emotions in listeners (*felt* emotions), or (ii) express/suggest emotions to listeners (*perceived* emotions), depending on the context [56]. As pointed out by Krumhansl [26], the distinction between *felt* and *perceived* emotions is important both from the theoretical and methodological point of views since the underlying models of representations may differ [71]. One may argue about the fact that music can communicate and trigger emotions in listeners and this has been the subject of numerous debates [37]. However a straightforward demonstration of the latter doe not require a controlled laboratory setting and may be conducted in a common situation, at least in certain cultures, that of watching/listening movies with accompanying soundtracks. In the documentary on film score composer Bernard Hermann [61], the motion picture editor Paul Hirsch (e.g. Star Wars, Carrie) discusses the effect of music in a scene from Alfred Hitchcock's well-known thriller/horror movie *Psycho*, whose soundtrack was composed by Hermann: *"The scene consisted of three very simple shots, there was a close up of her [Janet Lee] driving, there was a point of view of the road in front of her and there was a point of view of the police car behind her that was reflected in the rear mirror. The material was so*

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

simple and yet the scene was absolutely gripping. And I reached over and I turned off the sound to the television set and I realised that the extreme emotional duress I was experiencing was due almost entirely to the music.". With regard to music retrieval, several studies on music information needs and user behaviors have stimulated interest in developing models for the automatic classification of music pieces according to the emotions or mood they suggest. In [28], the responses of 427 participants to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"* showed that emotional/mood states would be used in every third song query, should they be possible. The importance of musical mood metadata was further confirmed in the investigations by Lesaffre et al. [30] which give high importance to affective/emotive descriptors, and indicate that users enjoy discovering new music by entering mood-based queries, as well as those by Bischoff et al. [5] which showed that 15% of the song queries on the web music service Last.fm were made using mood tags. As part of our project Making Musical Mood Metadata (M4) in partnership with the BBC and I Like Music, the present study aims to (i) review the current trends in music emotion recognition (MER), and (ii) provide insights to improve MER models. The remainder of this article is organised as follows. In Section 2, we present the three main types of (music) emotion representations (categorical, dimensional and appraisal). In Section 3, we review MER studies by focusing on those published between 2009 and 2011, and discuss the current trends in terms of features and feature selection frameworks. Section 4 presents state-of-the-art's machine learning techniques for MER. In Section 5, we discuss some of the findings in MER and conclude by highlighting the main implications to improve content and context-based MER models.

2 Representation of Emotions

2.1 Categorical Model

Table 1 presents the main categorical and dimensional emotion models used in the MER studies reviewed in this article. According to the categorial approach, emotions can be represented as a set of categories that are distinct from each others. Ekman's categorical emotion theory [13] introduced *basic* or universal emotions that are expected to have prototypical facial expressions and emotion-specific physiological signatures. The seminal work from Hevner [21] highlighted (i) the bipolar nature of music emotions (e.g. happy/sad), (ii) a possible way of representing them spatially across a circle, as well as (iii) the multi-class and multi-label nature of music emotion classification. Schubert proposed a new taxonomy, the updated Hevner model (UHM) [54], which refined the set of adjectives proposed by Hevner, based on a survey conducted by 133 musically experienced participants. Based on Hevner's list, Russell's circumplex of emotion [44], and Whissell's dictionary of affect [65], the UHM consists in 46 words grouped into nine clusters.

Bischoff et al. [6] and Wang et al. [63] proposed categorical emotion models by dividing the Thayer-Russell Arousal/Valence space (see Section 2.2) into into

Table 1. Categorical and dimensional models of music emotions used in MER. Cat.: Categorical; Dim.: Dimensional; Ref.: References.

Notation	Description	Approach	Ref.
UHM9	Update of Hevner's adjective Model (UHM) including nine categories	Cat.	[54]
AMC5C	5 MIREX audio mood classification (AMC) clusters ("Passionate","Rollicking", "Literate", "Humorous", "Aggressive")	Cat.	[22] [9] [6] [58] [62]
5BE	5 basic emotions ("Happy", "Sad", "Tender", "Scary", "Angry")	Cat.	[12] [45]
AV4Q	4 quadrants of the Thayer-Russell AV space ("Exuberance", "Anxious/Frantic", "Depression", "Contentment")	Cat.	[6] [63]
AV11C	11 subdivisions of the Thayer-Russell AV space ("Pleased", "Happy", "Excited", "Angry", "Nervous", "Bored", "Sad", "Sleepy", "Peaceful", "Relaxed", and "Calm")	Cat.	[19]
AMG12C	12 clusters based on AMG tags	Cat.	[33]
72TCAL500	72 tags from the CAL-500 dataset (genres, instruments, emotions, etc.)	Cat.	[4]
AV4Q-UHM9	Categorisation of UHM9 in Thayer-Russell's quadrants (AV4Q)	Cat.	[40]
AV8C	8 subdivisions of the Thayer-Russell AV space	Cat.	[24]
4BE	4 basic emotions ("Happy", "Sad", "Angry", "Fearful")	Cat.	[59]
4BE-AV	4 basic emotions based on the AV space ("Happy", "Sad", "Angry", "Relaxing")	Cat.	[63]
9AD	Nine affective dimensions from Asmus ("Evil", "Sensual", "Potency", "Humor", "Pastoral", "Longing", "Depression", "Sedative", and "Activity")	Dim.	[2]
AV	Arousal/Valence (Thayer-Russell model)	Dim.	[19]
EPA	Evaluation, potency, and activity (Osgood model)	Dim.	
6D-EPA	6 dim. correlated with the EPA model	Dim.	[35]
AVT	Arousal, valence, and tension	Dim.	[12]

four quadrants (AV4Q). [19] proposed subdivisions of the four AV space quadrants into a larger set, composed of 11 categories (AV11C). Their model, assessed on a prototypical database, led to high MER performance (see Section 3). [22] and [33] proposed mood taxonomies based on the (semi-)automatic analysis of mood tags with clustering techniques. [22] applied an agglomerative hierarchical clustering procedure (Ward's criterion) on similarity data between mood labels mined from the `AllMusicGuide.com` (AMG) website presenting annotations made by professional editors. The procedure generated a set of five clusters which further served as a mood representation model (denoted AMC5C, here) in the MIREX audio mood classification task and has been widely used since (e.g. in [22], [9], [6], and [62]). In this model, the similarity between emotion labels is computed from the frequency of their co-occurence in the dataset. Consequently some of the mood tag clusters may comprise tags which suggest different emotions. Training MER models on these clusters may be misleading for inference systems, as shown in [6] where prominent confusion patterns between clusters are reported (between Clusters 1 and 2, as well as between Clusters 4 and 3). [24] proposed a new categorical model by collecting 4460 mood tags and AV values from 10 music clip annotators and by further grouping them relying on unsupervised classification techniques. The collected mood tags were processed to get rid of synonymous and ambiguous terms. Based on the frequency distribution of the 115 remaining mood tags, the 32 most frequently used tags were retained. The AV values associated with the tags were processed using K-means clustering which led to a configuration of eight clusters (AV8C). The results show that some regions can be identified by the same representative mood tags

as in previous models, but that some of the mood tags present overlap between regions. Categorical approaches have been criticized for their restrictions due to the discretization of the problem into a set of "families" or "landmarks" [39] [8], which prevent to consider emotions which differ from these landmarks. However, as highlighted in the introduction, for music retrieval applications based on language queries, such landmarks (keywords/tags) have shown to be useful.

2.2 Dimensional Model

In contrast to categorical emotion models, dimensional models characterise emotions based on a small number of dimensions intended to correspond to the internal human representation of emotions. The psychologist Osgood [41] devised a technique for measuring the connotative meaning of concepts, called the *semantic differential technique* (SDT). Experiments were conducted with 200 undergraduate students who were asked to rate 20 concepts using 50 descriptive scales (7-point Likert scales whose poles were bipolar adjectives) [41]. Factor analyses accounted for almost 70% of the common variance in a three-dimensional configuration (50% of the total variance remained unexplained). The first factor was clearly identifiable as *evaluative*, for instance representing adjective pairs such as *good/bad*, *beautiful/ugly* (dimension also called *valence*), the second factor identified fairly well as *potency*, for instance related to bipolar adjectives *large/small*, *strong/weak*, *heavy/light* (dimension also called *dominance*), and the third factor appeared to be mainly an *activity* variable, related to adjectives such as *fast/slow*, *active/passive*, *hot/cold* (dimension also called *arousal*). Osgood's EPA model was used for instance in the study [10] investigating how well music (theme tune) can aid automatic classification of TV programmes from BBC Information & Archive. A slight variation of the EPA model was used in [11] with the *potency* dimension being replaced by one related to *tension*. Although Osgood's model has been shown to be relevant to classify affective concepts, its adaptability to music emotions is notwithstanding not straightforward. Asmus [2] replicated Osgood's SDT in the context of music emotions classification. Measures were developed from 2057 participants on 99 affect terms in response to musical excerpts and then factor analysed. Nine affective dimensions (9AD) were found to best represent the measures, two of which were found to be common to the EPA model. Probably because it is harder to visually represent nine dimensions and because it complicates the classification problem, this model has not been used yet in the MIR domain, to our knowledge.

The works that have had the most influence on the choice of emotion representations in MER so far are those from Russell [44] and Thayer [57]. Russell devised a *circumplex model of affect* which consists of a two-dimensional, circular structure involving the dimensions of *arousal* and *valence* (denoted AV and called the *core affect dimensions* following Russell's terminology). Within the AV model, emotions that are across a circle from one another correlate inversely, aspect which is also in line with the semantic differential approach and the bipolar adjectives proposed by Osgood. Schubert [53] developed a measurement interface called the "two-dimensional emotional space" (2DES) using Russell's core affect dimensions and proved the validity of the methodology, experimentally. While

the AV space stood out amongst other models for its simplicity and robustness, higher dimensionality have shown to be needed when seeking for completeness. The potency or dominance dimension related to power and control proposed by Osgood is necessary to make important distinctions between fear and anger, for instance, which are both active and negative states. Fontaine et al. [16] advocated the use of a fourth dimension related to the expectedness or unexpectedness of events, which to our knowledge has not been applied in the MIR domain so far.

A comparison between the categorical, or discrete, and dimensional models has been conducted in [11]. Linear mapping techniques revealed a high correspondence along the core affect dimensions (arousal and valence), and the three obtained dimensions could be reduced to two without significantly reducing the goodness of fit. The major difference between the discrete and categorical models concerned the poorer resolution of the discrete model in characterizing emotionally ambiguous examples. [60] compared the applicability of music-specific and general emotion models, the Geneva Emotional Music Scale (GEMS) [71], the discrete and dimensional AV emotion models, in the assessment of music-induced emotions. The AV model outperformed the other two models in the discrimination of music excerpts, and principal component analysis revealed that 89.9% of the variance in the mean ratings of all the scales (in all three models) was accounted for by two principal components that could be labelled as valence and arousal. The results also revealed that personality-related differences were the most pronounced in the case of the discrete emotion model, aspect which seems to contradict that obtained in [11].

2.3 Appraisal Model

The appraisal approach was first advocated by Arnold [1] who defined appraisal as a cognitive evaluation able to distinguish qualitatively among different emotions. The theory of appraisal therefore accounts for individual differences and variations to responses across time [43], as well as cultural differences [47]. The component process appraisal model (CPM) [48] describes an emotion as a process involving five functional components: cognitive, peripheral efference, motivational, motor expression, and subjective feeling. Banse and Scherer [3] proved the relevance of CPM predictions based on acoustical features of vocal expressions of emotions. Significant correlations between appraisals and acoustic features were also reported in [27] showing that inferred appraisals were in line with the theoretical predictions. Mortillaro et al. [39] advocate that the appraisal framework would help to address the following concerns in automatic emotion recognition: (i) how to establish a link between models of emotion recognition and emotion production? (ii) how to add contextual information to systems of emotion recognition? (iii) how to increase the sensitivity with which weak, subtle, or complex emotion states can be detected? All these points are highly significant for MER with a MIR perspective whereas appraisal models such as the CPM have not yet been applied in the MIR field, to our knowledge. The appraisal framework is especially promising for the development of context-sensitive automatic emotion recognition systems taking into account the environment (e.g. work, or home), the situation (relaxing, performing a task), or the subject (personnality traits),

for instance [39]. This comes from the fact that appraisals themselves represent abstractions of contextual information. By inferring appraisals (e.g. obstruction) from behaviors (e.g. frowning), information about causes of emotions (e.g. anger) can be inferred [7].

3 Acoustical and Contextual Analysis of Emotions

Studies in music psychology [56], musicology [18] and music information retrieval [25] have shown that music emotions were related to different musical variables. Table 2 lists the content and context-based features used in the studies reviewed hereby. Various acoustical correlates of articulation, dynamics, harmony, instrumentation, key, mode, pitch, melody, register, rhythm, tempo, musical structure, and timbre have been used in MER models. Timbre features have shown to provide the best performance in MER systems when used as individual features [52] [73]. Schmidt et al. investigated the use of multiple audio content-based features (timbre and chroma domains) both individually and in combination in a feature fusion system [52] [49]. The best individual features were octave-based spectral contrast and MFCCs. However, the best overall results were achieved using a combination of features, as in [73] (combination of rhythm, timbre and pitch features). Eerola et al. [12] extracted features representing six different musical variables (dynamics, timbre, harmony, register, rhythm, and articulation) to further apply statistical feature selection (FS) methods: multiple linear regression (MLR) with a stepwise FS principle, principle component analysis (PCA) followed by the selection of an optimal number of components, and partial least square regression (PLSR) with a Bayesian information criterion (BIC) to select the optimal number of features. PLSR simultaneously allowed to reduce the data while maximising the covariance between the features and the predicted data, providing the highest prediction rate (R^2=.7) with only two components. However, feature selection frameworks operating by considering all the emotion categories or dimensions at the same time may not be optimal; for instance, features explaining why a song expresses "anger" or why another sounds "innocent" may not be the same. Pairwise classification strategies have been successfully applied to musical instrument recognition [14] showing the interest of adapting the feature sets to discriminate two specific instruments. It would be worth investigating if music emotion recognition could benefit from pairwise feature selection strategies as well.

In addition to audio content features, lyrics have also been used in MER, either individually, or in combination with features belonging to different domains (see multi-modal approaches in Section 4.4). Access to lyrics has been facilitated by the emergence of lyrics databases on the web (e.g. lyricwiki.org, musixmatch.com), some of them providing APIs to retrieve the data. Lyrics can be analysed using standard natural language processing (NLP) techniques. To characterise the importance of a given word in a song given the corpus it belongs to, authors used the term frequency - inverse document frequency (TF-IDF) measure [9] [36]. Methods to analyse emotions in lyrics have been developed using lexical resources for opinion and sentiment mining such as SentiWordNet

Table 2. Content (audio and lyrics) and context-based features used in MER (studies between 2009 and 2011)

Type	Notation	Description	References
		Content-based features	
Articulation	EVENTD	Event density	[12]
Articulation/Timbre	ATTACS	Attack slope	[12]
Articulation/Timbre	ATTACT	Attack time	[12]
Dynamics	AVGENER	Average energy	[19]
Dynamics	INT	Intensity	[40]
Dynamics	INTR	Intensity ratio	[40]
Dynamics	DYN	Dynamics features	[45]
Dynamics	RMS	Root mean square energy	[12] [35] [45]
Dynamics	LOWENER	Low energy	[35]
Dynamics	ENER	Energy features	[36]
Harmony	OSPECENT	Octave spectrum entropy	[12]
Harmony	HARMC	Harmonic change	[12]
Harmony	CHROM	Chroma features	[52]
Harmony	HARMF	Harmony features	[45]
Harmony	RCHORDF	Relative chord frequency	[55]
Harmony	WCHORDD	Weighted chord differential	[35]
Instrum./Rhythm	PERCTO	Percussion template occurrence	[58]
Instrumentation	BASSTD	Bass-line template distance	[58]
Key/Mode	KEY	Key	[19]
Key/Mode	KEYC	Key clarity	[12]
Key/Mode	MAJ	Majorness	[12]
Key/Mode	SPITCH	Salient pitch	[12]
Key/Mode	WTON	Weighted tonality	[35]
Key/Mode	WTOND	Weighted tonality differential	[35]
Pitch/Melody	PITCHMIDI	Pitch MIDI features	[73]
Pitch/Melody	MELOMIDI	Melody MIDI features	[73]
Pitch/Melody	PITCH	Pitch features	[45]
Pitch/Timbre	ZCR	Zero-crossing rate	[73] [72]
Register	CHROMD	Chromagram deviation	[12]
Register	CHROMC	Chromagram centroid	[12]
Rhythm/Tempo	BEATINT	Beat interval	[19]
Rhythm/Tempo	SPECFLUCT	Spectrum fluctuation	[12]
Rhythm/Tempo	TEMP	Tempo	[12]
Rhythm/Tempo	PULSC	Pulse clarity	[12]
Rhythm/Tempo	RHYCONT	Rhythm content features	[73]
Rhythm/Tempo	RHYSTR	Rhythm strength	[40]
Rhythm/Tempo	CORRPEA	Correlation peak	[40]
Rhythm/Tempo	ONSF	Onset frequency	[40]
Rhythm/Tempo	RHYT	Rhythm features	[45]
Rhythm/Tempo	SCHERHYT	Scheirer rhythm features	[55]
Rhythm/Tempo	PERCF	Percussive features	[36]
Structure	MSTRUCT	Multidimensional structure features	[12]
Structure	STRUCT	Structure features	[45]
Timbre	SPECC	Spectral centroid	[6] [35] [73]
Timbre	HARMSTR	Harmonic strength	[19]
Timbre	MFCC	Mel frequency cepstral coefficient	[6] [4] [58] [73] [62] [45] [52] [49] [72] [59] [51] [45]
Timbre	SPECC	Spectral centroid	[12] [73] [72] [50] [52] [40] [55]
Timbre	SPECS	Spectral spread	[12]
Timbre	SPECENT	Spectral entropy	[12]
Timbre	SPECR	Spectral rolloff	[12] [73] [72] [50] [52] [40] [55]
Timbre	SF	Spectral flux	[73] [72] [50] [52] [40] [55]
Timbre	OBSC	Octave-based spectral contrast	[50] [52] [49] [51] [40] [29]
Timbre	RPEAKVAL	Ratio between average peak and valley strength	[40]
Timbre	ROUG	Roughness	[12]
Timbre	TIM	Timbre features	[45]
Timbre	SPEC	Spectral features	[36]
Timbre	ECNTT	Echo Nest timbre feature	[51] [36]
Lyrics	SENTIWORD	Occurence of sentiment word	[9]
Lyrics	NEG-SENTIW	Occurrence of sentiment word with negation	[9]
Lyrics	MOD-SENTIW	Occurrence of sentiment word with modifier	[9]
Lyrics	WORDW	Word weight	[9]
Lyrics	LYRIC	Lyrics feature	[73]
Lyrics	RSTEMFR	Relative stem frequency	[55]
Lyrics	TF-IDF	Term frequency - Inverse document frequency	[9] [36]
Lyrics	RHYME	Rhyme feature	[63]
		Context-based features	
Social tags	TAGS	Tag relevance score	[4]
Web-mined tags	DOCRS	Document relevance score	[4]
Metadata	ARTISTW	Artist weight	[9]
Metadata	META	Metadata features (e.g. artist's name, title)	[55]

(measures of positivity, negativity, objectivity) [9], and the affective norm for English words (measures of arousal, valence, and dominance) [36]. Since meaning emerges from subtle word combinations and sentence structure, research is still needed to develop new features characterising emotional meanings in lyrics. [63] proposed a feature to characterise rhymes whose patterns are relevant to emotion expression, as poems can attest. To attempt to improve the performance of MER systems only relying on content-based features, and in order to bridge the semantic gap between the raw data (signals) and high-level semantics (meanings), several studies introduced context-based features. [9], [6], [4], and [62] used music tags mined from websites known to have good quality information about songs, albums or artists (e.g. bbc.co.uk, rollingstone.com), social music platform (e.g. last.fm), or web blogs (e.g. livejournal.com). Social tags are generally fused with audio features to improve overall performance of the classication task [6] [4] [62].

4 Machine Learning for Music Emotion Recognition

4.1 Early Categorical Approaches and Multi-Label Classification

Associating music with discrete emotion categories was demonstrated by the first works that used an audio-based approach. Li et al. [31] used a song database hand-labelled with adjectives belonging to one of 13 categories and trained Support Vector Machines (SVM) on timbral, rhythmic and pitch features. The authors report large variation in the accuracy of estimating the different mood categories, with the overall accuracy (F score) remaining below 50%. Feng et al. [15] used a Back Propagation Neural Network (BPNN) to recognise to which extent music pieces belong to four emotion categories ("happiness", "sadness", "anger", and "fear"). They used features related to tempo (fast-slow) and articulation (staccato-legato), and report 66% and 67% precision and recall, respectively. However, the actual accuracy of detecting each emotion fluctuated considerably. The modest results obtained with early categorical approaches can be attributed to the difficulty in assigning music pieces to any single category, and the ambiguity of mood adjectives themselves. For these reasons subsequent research have moved on to use multi-label, fuzzy or continuous (dimensional) emotion models.

In multi-label classification, training examples are assigned multiple labels from a set of disjoint categories. MER was first formulated as a multi-label classification problem by Wieczorkowska et al. [66] applying a classifier specifically adopted to this task. In a recent study, Sanden and Zhang [46] examined multi-label classification in the general music tagging context (emotion labelling is seen as a subset of this task). Two datasets, the CAL500 and approximately 21,000 clips from Magnatune (each associated with one or more of 188 different tags) were used in the experiments. The clips were modeled using statistical distributions of spectral, timbral and beat features. The authors tested Multi-Label k-Nearest Neighbours (MLkNN), Calibrated Label Ranking (CLR), Backpropagation for Multi-Label Learning (BPMLL), Hierarchy of Multi-Label Classifiers (HOMER), Instance Based Logistic Regression (IBLR) and Binary Relevance

Table 3. Content-based music emotion recognition (MER) models (studies between 2009 and 2011). [a]: *F-measure;* [b]: *Accuracy;* [c]: *Coefficient of determination* R^2*;*[d]*: Average Kullback-Leibler divergence;* [e]*: Average distance;* [f]*: Mean* l^2 *error.* SSD: statistical spectrum descriptors. BAYN: Bayesian network. ACORR: Autocorrelation. Best reported configurations are indicated in bold.

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Lin et al. (2009) [33]	Audio	AMG (1535)	Cat. (AMG12C)	track	MARSYAS (436)	SVM	56.00%[a]
Han et al. (2009) [19]	Audio	AMG (165)	Cat. (AV11C)	track	KEY, AVGENER, TEMP, σ(BEATINT), σ(HARMSTR)	**SVR**, SVM, GMM	94.55%[b]
Eerola et al. (2009) [12]	Audio	Soundtrack110 (110)	Cat. (5BE) & Dim. (AV & AVT)	15.3 s (avg)	RMS, SPECC, SPECS, SPECENT, ROUG, OSPECENT, HARMC, KEYC, MAJ, CHROMC, CHROMD, SPITCH, SPECFLUCT, TEMP, PULSC, EVENTD, ATTACS, ATTACT, MSTRUCT (29)	MLR + STEPS, PCA + FS, **PLSR + DT**	70%[c] (avg)
Tsunoo et al. (2010) [58]	Audio	CAL500 (240)	Cat. (AMC5C)	track	PERCTO (4), BASSTD (80), 26 M,σ MFCCs, 12 M,σ corr(Chroma)	**TEML + SVM**	56.4%[d]
Zhao et al. (2010) [73]	Audio	Chin. & West. (24)	Cat. (AV4Q)	30s	**PITCH (5), RHYT (6), MFCCs (10), SSDs (9)**	**BAYN**	74.9%[b]
Schmidt et al. (2010) [50]	Audio	MoodSwings Lite (240)	Dim. (AV)	1s	OBSC	MLR, LDS Kalman, LDS KALF, **LDS KALFM**	2.88[d]
Schmidt et al. (2010) [52]	Audio	MoodSwings Lite (240)	Cat. (AV4Q) & Dim. (AV)	1s	**MFCCs**, CHROM (12), SSDs, OBSC	SVM / PLSR, **SVR**	0.137[e]
Schmidt & Kim (2010) [49]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs**, ACORR(CHROM), SSDs, **OBSC**	MLR, PLSR, **SVR**	3.186 / 13.61[d]
Myint & Pwint (2010) [40]	Audio	Western pop (100)	Cat. (AV4Q-UHM9)	segment	INT, INTR, SSD, OBSC, RHYSTR, CORRPEA, RPEAKVAL, M(TEMP), M(ONSF)	OAO FSVM	37%[b]
Lee et al. (2011) [29]	Audio	Clips (1000)	Dim. 2 (AV)	20s	OBSC	**SVM**	67.5%[b]
Mann et al. (2011) [35]	Audio	TV theme tunes (144)	Dim. (6D-EPA)	track	RMS, LOWENER, SPECC, WTON, WTOND, WCHORDD, TEMP	**SVM**	80-94%[b]
Vaizman et al. (2011) [59]	Audio	Piano, Vocal (76)	Cat. (4BE)	track	34 MFCCs	DTM	60%[a]
Schmidt & Kim (2011) [51]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs (20)**, OBSC, ECNTTs (12)	MLR, **CRF**	0.122[f]
Saari et al. (2011) [45]	Audio	Film soundtrack (104)	Cat. (5BE)	track	52 (DYN, RHY, PITCH, HARM, TIM, STRUCT) + MFCCs (14)	**NB**, k-NN, SVM, SMO	59.4%[b]
Wang et al. (2011) [63]	Lyrics	Chinese songs (500)	Cat. (4BE-AV)	track	TF-IDF, RHYME	MLR, NB, SVM-SMO, DECT (J48)	61.5%[a]

Table 4. Multi-modal music emotion recognition (MER) models (studies between 2009 and 2011). [a]*: F-measure;* [b]*: Accuracy;* [c]*: Mean average precision;* [d]*: Coefficient of determination* R^2. FSS: Feature subset selection. Best reported configurations are indicated in bold.

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Dang & Shirai (2009) [9]	Lyrics, Web-mined Tags	LiveJournal, LyricWiki (6000)	Cat. (AMC5C)	track	TF/IDF, SENTIWORD, NEG-SENTIW, MOD-SENTIW, WORDW, ARTISTW	SVM, **NB**, Graph-based	57.44%[b]
Bischoff et al. (2009) [6]	Audio, Social tags	Last.fm, AMG (1192)	Cat. (AMC5C) & AV4Q	30s	MFCCs, TEMP, CHROM (12), SPECC, ... / log(TF)	**SVM (RBF)**, LOGR, RANF, GMM, K-NN, DECT, **NB**	57.2%[a]
Barrington et al. (2009) [4]	Audio, Social tags, Web-mined tags	Last.fm, CAL500 (500)	Cat. (72TCAL500)	30s	MFCCs (39), Δ MFCCs, $\Delta\Delta$ MFCCs, CHROM (12) / + 8-GMM, TAGRS, DOCRS	**CSA**, RANB, KC-SVM	53.8%[c]
Wang et al. (2010) [62]	Audio, Social tags	Last.fm, WordNet, AMG (1804)	Cat. (AMC5C)	track	MARSYAS (138) & PSYSOUND3 + FSS / MFCCs + GMM	SVM PPK-RBF / NRQL	60.6%[b]
Zhao et al. (2010) [73]	Audio, Lyrics, MIDI	Chinese songs (500)	Cat. (AV4Q)	track	MFCCs, LPC, SPECC, SPECR, SPECF, ZCR, ... (113) / N-GRAM LYRIC (2000) / PITCH-MIDI, MELOMIDI (101)	**SVM**, NB, DECT	61.6%[b]
Schuller et al. (2011) [55]	Audio, Lyrics, Metadata	NTWICM, lyricsDB, LyricWiki (2648)	Dim. (AV)	track	RCHORDF (22), SCHERHYT (87), SPECC,... (24) / RSTEMFR (393), META (152)	**ConceptNet, Porter stemming, UREPT**	.60 (A) & .74 (V)[d]
McVicar et al. (2011) [36]	Audio, Lyrics	EchoNest API, lyric-smode.com, ANEW (119 664)	Dim.(AV)	track	TF-IDF, ECNT (65)	**CCA**	N/A

kNN (BRkNN) models, and two separate evaluations were performed using the two datasets. In both cases, the CLR classifier using a Support Vector Machine (CLR_{SVM}) outperformed all other approaches (peak F_1 score of 0.497 and precision of 0.642 on CAL500). However, CLR with Decision Trees, BPMLL, and MLkNN also performed competitively.

4.2 Fuzzy Classification and Emotion Regression

A possible approach to account for subjectivity in emotional responses is the use of fuzzy classification incorporating fuzzy logic into conventional classification strategies. The work of Yang et al. [70] was the first to take this route. As opposed to associating pieces with a single or a discrete set of emotions, fuzzy classification uses fuzzy vectors whose elements represent the likelihood of a piece belonging to each respective emotion categories in a particular model. In [70], two classifiers, Fuzzy k-NN (FkNN) and Fuzzy Nearest Mean (FNM), were tested using a database of 243 popular songs and 15 acoustic features. The authors performed 10-fold cross validation and reported 68.22% and 70.88% mean accuracy for the two classifiers respectively. After applying stepwise backward feature selection, the results improved to 70.88% and 78.33%.

The techniques mentioned so far rely on the idea that emotions may be organised in a simple taxonomy consisting of a small set of universal emotions (e.g. happy or sad) and more subtle differences within these categories. Limitations of this model include *i)* the fixed set of classes considered, *ii)* the ambiguity in the meaning of adjectives associated with emotion categories, and *iii)* the potential heterogeneity in the taxonomical organisation. The use of a continuous emotion space such as Thayer-Russell's Arousal-Valence (AV) space and corresponding dimensional models is a solution to these problems. In the first study that addresses these issues [69], MER was formulated as a regression problem to map high-dimensional features extracted from audio to the two-dimensional AV space directly. AV values for *induced* emotion were collected from 253 subjects for 195 popular recordings. After basic dimensionality reduction of the feature space, three regressors were trained and tested: Multiple Linear Regression (MLR) as baseline, Support Vector Regression (SVR) and Adaboost.RT, a regression tree ensemble. The authors reported coefficient of determination statistics (R^2) with peak performance of 58.3% for arousal, and 28.1% for valence using SVR. Han et al. [19] used SVR for training distinct regressors to predict arousal and valence both in terms of Cartesian and polar coordinates of the AV space. A policy for partitioning the AV space (AV11C) and mapping coordinates to discrete emotions was used, and an increase in accuracy from 63.03% to 94.55% was obtained when polar coordinates were used in this process. Notably Gaussian Mixture Model (GMM) classifiers performed competitively in this study. Schmidt et al. [52] showed that Multi-Level Least-Squares Regression (MLSR) performs comparably to SVR at a lower computational cost. An interesting observation is that combining multiple feature sets does not necessarily improve regressor performance, probably due to the curse of dimensionality. The solution was seen in the use of different fusion topologies, i.e. using separate regressors for each

feature set. Huq et al. [23] performed a systematic evaluation of content-based emotion recognition to identify a potential *glass ceiling* in the use of regression. 160 audio features were tested in four categories, timbral, loudness, harmonic, and rhythmic (with or without feature selection), as well as different regressors in three categories, Linear Regression, variants of regression trees and SVRs with Radial Basis Function (RBF) kernel (with or without parameter optimisation). Ground truth data were collected to indicate *induced* emotion, as in [69], by averaging arousal and valence scores from 50 subjects for 288 music pieces. Confirming earlier findings that arousal is easier to predict than valence, peak R^2 of 69.7% (arousal) and 25.8% (valence) were obtained using SVR-RBF. The authors concluded that small database size presents a major problem, while the wide distribution of individual responses to a song spreading in the AV space was seen as another limitation. In order to overcome the subjectivity and potential nonlinearity of AV coordinates collected from users, and to ease the cognitive load during data collection, Yang et al. proposed a method to automatically determine the AV coordinates of songs using pair-wise comparison of relative emotion differences between songs using a ranking algorithm [67]. They demonstrated that the increased reliability of ground truth pays off when different learning algorithms are compared. In [68], the authors modeled emotions as probability distributions in the AV space as opposed to discrete coordinates. They developed a method to predict these distributions using *regression fusion*, and reported a weighted R^2 score of 54.39%.

4.3 Methods for Music Emotion Variation Detection

It can easily be argued however that emotions are not necessarily constant during the course of a piece of music, especially in classical recordings. The problem of Music Emotion Variation Detection (MEVD) can be approached from two perspectives: the detection of time-varying emotion as a continuous trajectory in the AV space, or finding music segments that are correlated with well defined emotions. The task of dividing the music into several segments which contain homogeneous emotion expression was first proposed by Lu et al. [34]. In [70], the authors also proposed MEVD but by classifying features resulting from 10s segments with 33.3% overlap using a fuzzy approach, and then computing arousal and valence values from the fuzzy output vectors. Building on earlier studies, Schmidt et al. [50] demonstrated that emotion distributions may be modeled as 2D Gaussian distributions in the AV space, and then approached the problem of time-varying emotion tracking. In [50], they employed Kalman filtering in a linear dynamical system to capture the dynamics of emotions across time. While this method provided smoothed estimates over time, the authors concluded that the wide variance in emotion space dynamics could not be accommodated by the initial model, and subsequently moved on to use Conditional Random Fields (CRF), a probabilistic graphical model to approach the same problem [51]. In modeling complex emotion-space distributions as AV *heatmaps*, CRF outperformed the prediction of 2D Gaussians using MLR. However, the CRF model has higher computational cost.

4.4 Multi-Modal Approaches and Fusion Policies

The combination of multiple feature domains have become dominant in recent MER systems and a comprehensive overview of combining acoustic features with lyrics, social tags and images (e.g. album covers) is presented in [25]. In most works, the previously discussed machine learning techniques still prevail, however, different feature fusion policies may be applied ranging from concatenating normalised feature vectors (early fusion) to boosting, or ensemble methods combining the outputs of classifiers or regressors trained on different feature sets independently (late fusion). Late fusion is becoming dominant since it solves the issues related to tractability, and the curse of dimensionality affecting early fusion. Bischoff et al. [6] showed that classification performance can be improved by exploiting both audio features and collaborative user annotations. In this study, SVMs with RBF kernel outperformed logistic regression, random forest, GMM, K-NN, and decision trees in case of audio features, while the Naïve Bayes Multinomial classifier produced the best results in case of tag features. An experimentally-defined linear combination of the results then outperformed classifiers using individual feature domains. In a more recent study, Lin et al. [32] demonstrated that genre-based grouping complements the use of tags in a two-stage multi-label emotion classification system reporting an improvement of 55% when genre information was used. Finally, Schuller [55] et al. combined audio features with metadata and Web-mined lyrics. They used a stemmed bag of words approach to represent lyrics and editorial metadata, and also extracted mood concepts from lyrics using natural language processing. Ensembles of REPTrees (a variant of Decision Trees) are used in a set of regression experiments. When the domains were considered in isolation, the best performance was achieved using audio features (chords, rhythm, timbre), but taking into account all the modalities improved the results.

5 Discussion and Conclusions

The results from the audio mood classification (AMC) task ran at MIREX from 2007 to 2009, and that of studies published between 2009 and 2011 reviewed in this article, suggest the existence of a "glass ceiling" for MER at F-measure about 65%. In a recent study [45], high-level features (mode "majorness" and key "clarity") have shown to enhance emotion recognition in a more robust way than low-level features. In line with these results, we claim that in order to improve MER models, there is a need for new mid or high-level descriptors characterising musical clues, more adapted to *explain* our conditioning to musical emotions than low-level descriptors. Some of the findings in music perception and cognition [56], psycho-musicology [17] [18], and affective computing [39] have not yet been exploited or adapted to their full potential for music information retrieval. Most of the current approaches to emotion recognition articulate on black-box models which do not take into account the interpretability of the relationships between features and emotion components; this is a disadvantage when trying to understand the underlying mechanisms [64]. Other emotion representation models, the appraisal models [39], attempt to predict the association between

appraisal and emotion components making possible to interpret the relationships. Despite the promising applications of semantic web ontologies in the field of MIR, the ontology approach has only be scarcely used in MER. [62] proposed a music-mood specific ontology grounded in the Music Ontology [42], in order to develop a multi-modal MER model relying on audio content extraction and semantic association reasoning. Such approach is promising since the system from [62] achieved a performance increase of approximately 20% points (60.6%) in comparison with the system by Feng, Cheng and Yang (FCY1), proposed at MIREX 2009 [38]. Recent research focuses on the use of regression and attempt to estimate continuous-valued coordinates in emotion spaces, which may then be mapped to an emotion label or a broader category. The choice between regression and classification is however not straightforward, as both categorical and dimensional emotion models have strengths and weaknesses for specific applications. Retrieving labels or categories given the estimated coordinates is often necessary, which requires a mapping between the dimensional and categorical models. This may not be available for a given model, may not be valid from a psychological perspective, and may also be dependent on extra-musical circumstances. With regard to the use of multiple modalities, most studies to date confirm that the strongest factors enabling emotion recognition are indeed related to the audio content. However a glass ceiling seems to exist which may only be vanquished if both contextual features and features from different musical modalities are considered.

Acknowledgments. This work was partly funded by the TSB project 12033-76187 "Making Musical Mood Metadata" (TS/J002283/1).

References

1. Arnold, M.B.: Emotion and personality. Columbia University Press, New York (1960)
2. Asmus, E.P.: Nine affective dimensions (Test manual). Tech. rep., University of Miami (1986)
3. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. J. of Pers. and Social Psy. 70, 614–636 (1996)
4. Barrington, L., Turnbull, D., Yazdani, M., Lanckriet, G.: Combining audio content and social context for semantic music discovery. In: Proc. ACM SIGIR (2009)
5. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proc. ACM CIKM. pp. 193–202 (2008)
6. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification - a hybrid approach. In: Proc. ISMIR. pp. 657–662 (2011)
7. Castellano, G., Caridakis, G., Camurri, A., Karpouzis, K., Volpe, G., Kollias, S.: Body gesture and facial expression analysis for automatic affect recognition, pp. 245–255. Oxford University Press, New York (2010)
8. Cowie, R., McKeown, G., Douglas-Cowie, E.: Tracing emotion: an overview. Int. J. of Synt. Emotions (2012)
9. Dang, T.T., Shirai, K.: Machine learning approaches for mood classification of songs toward music search engine. In: Proc. ICKSE (2009)
10. Davies, S., Allen, P., Mann, M., Cox, T.: Musical moods: a mass participation experiment for affective classification of music. In: Proc. ISMIR. pp. 741–746 (2011)
11. Eerola, T.: A comparison of the discrete and dimensional models of emotion in music. Psychol. of Mus. 39(1), 18–49 (2010)
12. Eerola, T., Lartillot, O., Toivianien, P.: Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: Proc. ISMIR (2009)
13. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA (1978)
14. Essid, S., Richard, G., David, B.: Musical instrument recognition by pairwise classification strategies. IEEE Trans. on Audio, Speech, and Langu. Proc. 14(4), 1401–1412 (2006)

15. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. Proc. ACM SIGIR pp. 375–376 (2003)
16. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotions is not two-dimensional. Psychol. Sc. 18(2), 1050–1057 (2007)
17. Gabrielsson, A.: Emotional expression in synthesizer and sentograph performance. Psychomus. 14, 94–116 (1995)
18. Gabrielsson, A.: The influence of musical structure on emotional expression, pp. 223–248. Oxford University Press (2001)
19. Han, B.J., Dannenberg, R.B., Hwang, E.: SMERS: music emotion recognition using support vector regression. In: Proc. ISMIR. pp. 651–656 (2009)
20. Hevner, K.: Expression in music: a discussion of experimental studies and theories. Psychol. Rev. 42(2), 186–204 (1935)
21. Hevner, K.: Experimental studies of the elements of expression in music. Am. J. of Psychol. 48(2), 246–268 (1936)
22. Hu, X., Downie, J.S.: Exploring mood metadata: relationships with genre, artist and usage metadata. In: Proc. ISMIR (2007)
23. Huq, A., Bello, J.P., Rowe, R.: Automated music emotion recognition: A systematic evaluation. J. of New Mus. Res. 39(3), 227–244 (2010)
24. Kim, J.H., Lee, S., Kim, S.M., Yoo, W.Y.: Music mood classification model based on Arousal-Valence values. In: Proc. ICACT. pp. 292–295 (2011)
25. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G.: Music emotion recognition: a state of the art review. Proc. ISMIR pp. 255–266 (2010)
26. Krumhansl, C.L.: An exploratory study of musical emotions and psychophysiology. Can. J. of Exp. Psychol. 51(4), 336–353 (1997)
27. Laukka, P., Elfenbein, H.A., Chui, W., Thingujam, N.S., Iraki, F.K., Rockstuhl, T., Althoff, J.: Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotation emotion appraisals. In: Devillers, L., Schuller, B., Cowie, R., Douglas-Cowie, E., Batliner, A. (eds.) Proc. of LREC work. on Corp. for Res. on Emotion and Affect. pp. 53–57. European Language Resources Association, Paris (2010)
28. Lee, J.A., Downie, J.S.: Survey of music information needs, uses, and seeking behaviors: preliminary findings. In: Proc. ISMIR (2004)
29. Lee, S., Kim, J.H., Kim, S.M., Yoo, W.Y.: Smoodi: Mood-based music recommendation player. In: Proc. IEEE ICME. pp. 1–4 (2011)
30. Lesaffre, M., Leman, M., Martens, J.P.: A user oriented approach to music information retrieval. In: Proc. Content-Based Retriev. Conf. Daghstul Seminar Proceedings, Wadern Germany (2006)
31. Li, T., Ogihara, M.: Detecting emotion in music. Proc. ISMIR pp. 239–240 (2003)
32. Lin, Y.C., Yang, Y.H., Chen, H.H.: Exploiting online music tags for music emotion classification. ACM Trans. on Mult. Comp. Com. and App. 7S(1), 26:1–15 (2011)
33. Lin, Y.C., Yang, Y.H., Chen, H.H., Liao, I.B., Ho, Y.C.: Exploiting genre for music emotion classification. In: Proc. IEEE ICME. pp. 618–621 (2009)
34. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. on Audio, Speech, and Langu. Proc. 14(1), 5–18 (2006)
35. Mann, M., Cox, T.J., Li, F.F.: Music mood classificationof television theme tunes. In: Proc. ISMIR. pp. 735–740 (2011)
36. McVicar, M., Freeman, T., De Bie, T.: Mining the correlation between lyrical and audio features and the emergence of mood. In: Proc. ISMIR. pp. 783–788 (2011)
37. Meyer, L.B.: Emotion and meaning in music. The University of Chicago press (1956)
38. MIREX: Audio mood classification (AMC) results. `http://www.music-ir.org/mirex/wiki/2009:Audio_Music_Mood_Classification_Results` (2009)
39. Mortillaro, M., Meuleman, B., Scherer, R.: Advocating a componential appraisal model to guide emotion recognition. Int. J. of Synt. Emotions (2012 (in press))
40. Myint, E.E.P., Pwint, M.: An approach for multi-label music mood classification. In: Proc. ICSPS. vol. VI, pp. 290–294 (2010)
41. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning. University of Illinois Press, Urbana (1957)
42. Raimond, Y., Abdallah, S., Sandler, M., Frederick, G.: The music ontology. In: Proc. ISMIR. Vienna, Austria (2007)
43. Roseman, I.J., Smith, C.A.: Appraisal theory: Overview, assumptions, varieties, controversies, pp. 3–19. Oxford University Press, New York (2001)
44. Russell, J.A.: A circumplex model of affect. J. of Pers. and Social Psy. 39(6), 1161–1178 (1980)
45. Saari, P., Eerola, T., Lartillot, O.: Generalizability and simplicity as criteria in feature selection: application to mood classification in music. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(6), 1802–1812 (2011)
46. Sanden, C., Zhang, J.: An empirical study of multi-label classifiers for music tag annotation. Proc. ISMIR pp. 717–722 (2011)
47. Scherer, K.R., Brosch, T.: Culture-specific appraial biases contribute to emotion disposition. Europ. J. of Person. 288, 265–288 (2009)
48. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal processes in emotion: Theory, methods, research. Oxford University Press, New York (2001)

49. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions from audio. In: Proc. ISMIR. pp. 465–470 (2010)
50. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions using Kalman filtering. In: Proc. ICMLA. pp. 655–660 (2010)
51. Schmidt, E.M., Kim, Y.E.: Modeling musical emotion dynamics with conditional random fields. In: Proc. ISMIR. pp. 777–782 (2011)
52. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proc. ACM SIGMM MIR. pp. 267–273 (2010)
53. Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. Austral. J. of Psychol. 51(3), 154–165 (1999)
54. Schubert, E.: Update of the Hevner adjective checklist. Percept. and Mot. Skil. pp. 117–1122 (2003)
55. Schuller, B., Weninger, F., Dorfner, J.: Multi-modal non-prototypical music mood analysis in continous space: reliability and performances. In: Proc. ISMIR. pp. 759–764 (2011)
56. Sloboda, J.A., Juslin, P.N.: Psychological perspectives on music and emotion, pp. 71–104. Series in Affective Science, Oxford University Press (2001)
57. Thayer, J.F.: Multiple indicators of affective responses to music. Dissert. Abst. Int. 47(12) (1986)
58. Tsunoo, E., Akase, T., Ono, N., Sagayama, S.: Music mood classification by rhythm and bass-line unit pattern analysis. In: Proc. ICASSP. pp. 265–268 (2010)
59. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling dynamic patterns for emotional content in music. In: Proc. ISMIR. pp. 747–752 (2011)
60. Vuoskoski, J.K.: Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. Music. Sc. 15(2), 159–173 (2011)
61. Waletzky, J.: Bernard Hermann Music For the Movies. DVD Les Films d'Ici / Alternative Current (1992)
62. Wang, J., Anguerra, X., Chen, X., Yang, D.: Enriching music mood annotation by semantic association reasoning. In: Proc. Int. Conf. on Mult. (2010)
63. Wang, X., Chen, X., Yang, D., Wu, Y.: Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In: Proc. ISMIR. pp. 765–770 (2011)
64. Wehrle, T., Scherer, K.R.: Toward computational modelling of appraisal theories, pp. 92–120. Oxford University Press, New York (2001)
65. Whissell, C.M.: The dictionary of affect in language, vol. 4, pp. 113–131. Academic Press, New York (1989)
66. Wieczorkowska, A., Synak, P., Ras, Z.W.: Multi-label classification of emotions in music. Proc. Intel. Info. Proc. and Web Min. pp. 307–315 (2006)
67. Yang, Y.H., Chen, H.H.: Ranking-based emotion recognition for music organisation and retrieval. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(4), 762–774 (2010)
68. Yang, Y.H., Chen, H.H.: Prediction of the distribution of perceived music emotions using discrete samples. IEEE Trans. on Audio, Speech, and Langu. Proc. 19(7), 2184–2195 (2011)
69. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: A regression approach to music emotion recognition. IEEE Trans. on Audio, Speech, and Langu. Proc. 16(2), 448–457 (2008)
70. Yang, Y.H., Liu, C.C., Chen, H.H.: Music emotion classification: A fuzzy approach. Proc. ACM Int. Conf. on Mult. pp. 81–84 (2006)
71. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: Differentiation, classification, and measurement. Emotion 8(4), 494–521 (2008)
72. Zhao, Y., Yang, D., Chen, X.: Multi-modal music mood classification using co-training. In: Proc. Int. Conf. on Comp. Intel. and Soft. Eng. (CiSE). pp. 1–4 (2010)
73. Zhao, Z., Xie, L., Liu, J., Wu, W.: The analysis of mood taxonomy comparison between Chinese and Western music. In: Proc. ICSPS. vol. VI, pp. 606–610 (2010)