Music Emotion Recognition: From Content- to 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 striking ability of music to elicit emotions assures its prominent status in human culture and every day life. Music is often enjoyed and sought for its ability to induce or convey emotions, which may manifest in anything from a slight variation in mood, to changes in our physical condition and actions. Consequently, research on how we might associate musical pieces with emotions and, more generally, how music brings about an emotional response is attracting ever increasing attention. First, this paper provides a thorough review of studies on the relation of music and emotions from different disciplines. We then propose new insights to enhance automated music emotion recognition models using recent results from psychology, musicology, affective computing, semantic technologies and music information retrieval.

Keywords: music emotion, mood, recognition, retrieval, metadata, model, arousal, valence, multi-modal, ontology, appraisal, review, state of the art.

1 Introduction

Since the first empirical works on the relationships between music and emotions [25,46] a large body of research studies has given strong evidence towards the fact that - depending on contextual information - music can either (i) elicit/induce/evoke emotions in listeners (*felt* emotions), or (ii) express/suggest emotions to listeners (*perceived* emotions) [71]. As pointed out by Krumhansl [33], the distinction between *felt* and *perceived* emotions is important both from the theoretical and methodological points of views since the models of representations may differ. Felt emotions relate to the observation that listeners may experience an emotional response to music, whereas perceived emotions relate to the fact that music can communicate qualities associated with emotions [73]. [91] devised a scale to analyse music-induced emotions - the Geneva emotional music scale (GEMS) - and showed that the underlying taxonomic model of emotions differed from the models which were devised in studies investigating the representations of perceived music emotions (e.g. [25]). One may argue whether music can communicate and trigger emotions in listeners and this has been the subject of numerous debates [46]. However a demonstration of the latter does not

M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 228–252, 2013.
© Springer-Verlag Berlin Heidelberg 2013

require a controlled laboratory setting and can be undertaken while watching films. In the documentary about film score composer Bernard Hermann [79], 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 horror movie *Psycho*, the soundtrack of which was composed by Hermann: *"I was home one night and Psycho was on and I saw a scene in which Janet Lee had stolen some money. [...] The scene consisted of three very simple shots, there was a close up of her 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 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."*. Such effect is in line with Chion's theory that music, by "adding value" to the image, causes the filmgoer to construe the image differently [11][1].

With regard to music retrieval, several studies on music information needs and user behaviors have highlighted an interest in developing models for the automatic classification of music pieces according to the emotions or mood they suggest[2]. In [37], 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, where possible, emotional/mood states would be used in every third song query. The importance of musical mood metadata was further confirmed in the investigations by [39] 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 [8] 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 analysis of human annotations of music emotions on editorial resources such as `AllMusicGuide.com` (AMG) showed that emotion recognition can be viewed as a multi-class (different classes of emotions) and multi-label (different mood tags for each track) classification or regression problem in which a music piece is associated with a set of emotions [31] (e.g. a track can be described as being "soft", "tender" and "peaceful"). In a generic way, music emotion recognition models can be described as the combination of two components: a detection component (feature extraction and feature selection), and an inference component (machine learning, fusion of results). If MER studies were still sparse in 2006 [43], MER has since become a burgeoning

[1] The analysis of the effects of music on emotion perceived in film goes beyond the scope of this article, and we refer the reader to [12] and [53] for thoughtful discussions and investigations on this subject.

[2] We will employ the words music emotion and mood interchangeably since their distinctions are out of the scope of this article. If not specified otherwise, they will refer to emotions suggested by music, rather than felt emotions. We refer the reader to the work of Meyer [46] for a discussion on the differences between emotion ("temporary and evanescent") and mood ("relatively permanent and stable").

field, as highlighted by the growing number of publications on this topic within the music information retrieval community (see Sect. 3). In parallel to MIR research, psychologists improved emotion/mood representation models, as well as measurement techniques (see Sect. 2). The two main types of computational models in MIR (content- and context-based) are closely linked with the distinctions of music meaning formulated by Meyer [46]. On one hand, content-based approaches may be associated with the "absolutist" point of view which sees "the meaning of music as being essentially intramusical (non-referential)" [46], a facet coined as *intrinsic sources* of emotions by Sloboda and Juslin [71]. On the other hand, context-based approaches may be associated with the "referentialist" point of view which contends that "music also communicates meanings which in some way refer to the extramusical world of concepts, actions, emotional states, and character", facet later coined as *extrinsic sources* of emotions in [71]. Meyer also put forward that absolute and referential meanings are not mutually exclusive and "can and do coexist in one and the same piece of music". This point of view corresponds well with the paradigm underlying hybrid approaches to the MER problem which combine content- and context-based models and are by essence multi-modal (mixing together audio, symbolic notations, lyrics, social tags, etc.). The annual evaluation campaign Music Information Retrieval Evaluation eXchange (MIREX) collocated with the International Society for Music Information Retrieval (ISMIR) conference launched a task on audio mood classification (AMC) in 2007. The reported F-measures of MIREX state-of-the arts' MER models rose from 62% in 2007 to 66% in 2009. Although great improvements have been made in pattern recognition systems, the analysis of the 2007-2009 MIREX results and that of studies published between 2009 and 2011, reviewed in Sect. 3, suggest the existence of a "glass ceiling" for MER at F-measure around 65%. Such bottleneck for MIR machine learning models was highlighted by the systematic evaluations performed in the experiments from [4], in the context of music similarity measures. In order to overcome these limitations, hybrid and multi-modal approaches have been proposed, by taking advantage of social metadata, web-mined tags, semantic reasoning [6,80], music symbolic notations [92], and/or lyrics [35]. The recent developments of such multi-modal MER approaches are not unrelated to the ever growing amount of music resources on the web, data management infrastructures and application programming interfaces (APIs), as well as the advances in the closely related field of social media retrieval. As argued in [80] "combining information from sources like web-based, text and other sorts of multi-modal information with content-based features in an efficient way could be one of the solutions to break the bottleneck of pure content-based method".

The remainder of this article is organised as follows. In Sect. 2, we present the three main types of (music) emotion representations (categorical, dimensional and appraisal), before discussing aspects related to taxonomy and ontology. In Sect. 3, we review MER studies by focusing on those published between 2009 and 2011, and discuss aspects linked with databases, features, feature selection frameworks, and emotion variation across time. Sect. 4 presents state of the art machine learning techniques for MER. Throughout the article and in Sect. 5,

Table 1. Categorical and dimensional models of music emotions used in MER

Notation	Description	Approach	Ref.
UHM9	Update of Hevner's adjective Model (UHM) in nine categories	Cat.	[68]
AMC5C	5 MIREX audio mood classification (AMC) clusters ("Passionate","Rollicking", "Literate", "Humorous", "Aggressive")	Cat.	[27] [14] [9] [75] [80]
5BE	5 basic emotions ("Happy", "Sad", "Tender", "Scary", "Angry")	Cat.	[18] [58]
AV4Q	4 quadrants of the Thayer-Russell AV space ("Exuberance", "Anxious/Frantic", "Depression", "Contentment")	Cat.	[9] [81]
AV11C	11 subdivision categories of the Thayer-Russell AV space	Cat.	[24]
AMG12C	12 clusters based on AMG tags	Cat.	[42]
72TCAL500	72 tags from the CAL-500 dataset (genres, instruments, emotions, etc.)	Cat.	[6]
AV4Q-UHM9	Categorisation of UHM9 in Thayer-Russell's quadrants (AV4Q)	Cat.	[49]
AV8C	8 subdivision categories of the Thayer-Russell AV space	Cat.	[29]
4BE	4 basic emotions ("Happy", "Sad", "Angry", "Fearful")	Cat.	[77]
4BE-AV	4 basic emotions based on the AV space ("Happy", "Sad", "Angry", "Relaxing")	Cat.	[81]
9AD	Nine affective dimensions from Asmus	Dim.	[3]
AV	Arousal/Valence (Thayer-Russell model)	Dim.	[24]
EPA	Evaluation, potency, and activity (Osgood model)	Dim.	
6D-EPA	6 dim. correlates with the EPA model	Dim.	[44]
AVT	Arousal, valence, and tension	Dim.	[18]

we discuss some of the findings in MER and highlight the main implications to improve content- and context-based MER models.

2 Representation of Emotions

Table 1 presents the main categorical and dimensional emotion models employed in the MER studies reviewed in Sect. 3 and 4, and gives the associated notations used throughout the article.

2.1 Categorical Model

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 [19] was formulated a century after that proposed by Darwin [15], centred on *basic* or universal emotions that are expected to have prototypical facial expressions and emotion-specific physiological signatures. Ekman developed the facial action coding system (FACS), a system to taxonomize human facial expressions. The facial action coding system affect interpretation dictionary (FACSAID) relates an emotion category (e.g. happy) to action units (AU), coding the contraction or relaxation of one or more muscles (e.g. 6+12).

The scientific study of emotions in music has often been conducted in conjunction with the analysis of musical expression [25]. In order to secure the responses of individual listeners to music in a simple and objective way while leaving them enough freedom not to force their judgements, Hevner devised a list of 66 emotion-related adjectives, arranged in 14 groups. Listeners were asked to check all the adjectives they found appropriate to describe the music [26]. The meanings or affective characteristics of music pieces were further ascertained by comparing the numbers of votes for different adjectives. Hevner proposed an arrangement of eight adjective groups organised around a circle in order to simplify the selection task, so that "any two adjacent groups should have some characteristics in common, and that the groups at the extremities of any diameter of the circle should be as unlike each other as possible". This study was seminal since it highlighted (i) the bipolar nature of music emotions (e.g. happy/sad), (ii) a possible way of representing them spatially across a circle (on which Thayer-Russell's model is based [57]; see Sect. 2.2), 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) [68], 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 [57], and Whissell's dictionary of affect [83], the UHM consists in 46 words grouped into nine clusters.

Some categorical approaches have emerged from dimensional approaches based on the organisation of the Thayer-Russell Arousal/Valence (AV) space (see Sect. 2.2) into a set of "landmark" or "family" areas. This procedure has been followed for instance in [9] and [81] where the Thayer/Russel space was divided into four quadrants (AV4Q). [9] considered the following four quadrants (Q): Q1 - high energy/high stress ("anxious, frantic"), Q2 - high energy/low stress ("exuberance"), Q3 - low energy/low stress ("contentment"), and Q4 - low energy/high stress ("depression"). Similarly, [81] proposed: Q1 - high energy/low stress ("happy, exciting"), Q2 - high energy/high stress ("angry, anxious"), Q3 - low energy/high stress ("sad, bored"), and Q4 - low energy/low stress ("relaxing, serene"). The results from [9] report classification confusions between the quadrants 1 and 4 which, according to the authors, come from the fact that both quadrants are associated with emotional states involving high stress (negative valence), and that the arousal dimension did not ease the differentiation between them. [24] proposed subdivisions of the four AV space quadrants into a larger set composed of 11 categories (AV11C: "pleased", "happy", "excited", "angry", "nervous", "bored", "sad", "sleepy", "peaceful", "relaxed", and "calm") associated with the middle of the space. Their model, assessed on a prototypical database, led to high MER performance (see Table 3).

[27] and [42] proposed mood taxonomies based on the (semi-)automatic analysis of mood tags with clustering techniques. In a study exploring the relationships between mood, genre, artist, and usage metadata, [27] 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 led to 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 [27,14,9], and [80]). In this model, the similarity between emotion labels is computed based on the frequency of their co-occurence in the dataset. Consequently some of the mood tag clusters may comprise tags which suggest

different emotions: e.g. “literate” and “bittersweet” in cluster 3, “witty”, “humorous”, and “whimsical” in cluster 4. In contrast, some of the terms belonging to different clusters present close semantic associations: e.g. “literate” and “witty” (cluster 3 and 4, respectively). Training MER models on these clusters may be misleading for inference systems, as shown in [9] where prominent confusion patterns between clusters were reported (between clusters 1 and 2, as well as between clusters 4 and 3).

By combining findings from categorical and dimensional approaches, [49] proposed a mood taxonomy model by grouping the eight clusters associated with Schubert’s UHM across the four quadrants (Q) of the AV space: Q1 (UHM groups I, II, IX: “exuberance”), Q2 (UHM group VIII: “anxious”), Q3 (UHM groups V, VII: “depression”), and Q4 (UHM groups III, IV, and VI: “contentment”). However, the resulting classification accuracies have shown to be good only for the first quadrant (68% of correct classifications). [29] 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” [48], or “landmarks” [13], which prevent consideration of emotions which differ from these landmarks. However, for music retrieval applications based on language queries, such landmarks (keywords/tags) have shown to be useful.

2.2 Dimensional Model

In contrast with the categorical approach, the dimensional approach to emotion representation consists in characterising emotions based on a small number of dimensions intended to correspond to the internal human representation of emotions.

The psychologist Osgood [52] devised a technique for measuring the connotative meaning of concepts, called the *semantic differential technique (SDT)*. It involves the rating of words on a set of bipolar adjectives (e.g. happy/sad). 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) [52]. 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 like *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*). The SDT was later applied by Osgood [51] in thirty different cultures for 620 concepts validating the evaluation, potency, activity (EPA) model of representation of emotions, and the results were formulated in an *Atlas of affective meaning.* Osgood's EPA model was used for instance in the study [16] investigating how well music (theme tune) can aid automatic classification of TV programmes from BBC Information & Archives. A slight variation of the EPA model was used in [17] 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 not straightforward. In other words, it is reasonable to make the assumption that music emotions may be represented by a different set of dimensions than that uncovered for affective concepts, in general. Asmus [3] replicated Osgood's semantic differential technique 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 (potency and activity) were found to be common to the EPA model: "evil", "sensual", "potency", "humor", "pastoral", "longing", "depression", "sedative", and "activity". 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 of Russell [57] and Thayer [72]. 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). As in Hevner's circular representation of emotion-related adjectives, and Schlosberg's proposal that emotions are organised in a circular arrangement [62], within the AV model, emotions that are across a circle from one another correlate inversely (e.g. sadness/happiness). This characteristic is also in line with the semantic differential approach and the bipolar adjectives proposed by Osgood. Thayer's findings confirmed the relevance of the AV model in the musical domain where emotion classes can be defined in terms of arousal or energy (how exciting/calming musical pieces are) and valence or stress (how positive/negative musical pieces are). Schubert [67] 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. However, results obtained in [90] suggest that arousal and valence are not fully independent, even though they are two axes in the 2D Thayer-Russell space.

While the AV space stood out amongst other models for its simplicity and robustness, higher dimensionality has shown to be needed when seeking 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. [22] advocated the use of a fourth dimension related to the expectedness or unexpectedness of events, which to our knowledge has not been used in the MIR domain so far. It is worth mentioning that none of these dimensions can represent more complex and subtle emotional states such as *pride/shame* or *shy/extroverted* in a straightforward manner. As to whether such emotional states can be musically-induced requires investigation. Following the dimensional approach to emotion representation, several teams have focused on obtaining continuous representations of emotions from human labelers across time, both in the domains of affective computing for audiovisual recordings (e.g. FEELtrace [13]), and music (e.g. 2DES [67], MoodSwings [63]).

2.3 Comparison between Categorical and Dimensional Models

A comparison between the categorical, or discrete, and dimensional models has been conducted in [17]. 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. [78] compared the applicability of music-specific and general emotion models, the Geneva emotional music scale (GEMS), and 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, an aspect which seems to contradict the findings obtained in [17].

2.4 Appraisal Model

As described in [48], *"appraisal models are a third alternative perspective on emotion: they combine elements of dimensional models - emotions as emergent results of underlying dimensions - with elements of discrete theories - emotions have different subjective qualities - and add a definition of the cognitive mechanisms at the basis of emotions"*. The appraisal approach was first advocated by Arnold [2] 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 [56], as well as for some cultural differences [60]. Appraisal models attempt to explain the differentiation of emotional states with different configurations of the underlying appraisal dimensions which are then mapped to emotion labels. The component process appraisal model (CPM) [61] describes an emotion as a process involving five functional components: cognitive, peripheral efference, motivational, motor expression, and subjective feeling. Banse and Scherer [5] proved the relevance of CPM predictions based on acoustical features of vocal expressions of emotions. The acoustic features characterising 100 vocal affect bursts, representing five emotions, were successfully related to the power and control parts of the appraisal component of coping potential. Significant correlations between appraisals and acoustic features were also reported in [34] showing that inferred appraisals were in line with the theoretical predictions.

Mortillaro et al. [48] 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 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 (personality traits), for instance [48]. 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 the causes of emotions (e.g. anger) can be uncovered [10].

2.5 Ontology

Despite the promising applications of semantic web ontologies in the field of MIR (see e.g. [32]), the ontology approach has been scarcely used in MER. [80] proposed a music-mood specific ontology grounded in the Music Ontology (see [54,55]), in order to develop a multi-modal MER model relying on audio content extraction and semantic association reasoning. Such an approach is promising since the system from [80] 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 [47].

3 Acoustical and Contextual Analysis of Emotions

3.1 Databases

Several music emotion annotation databases produced by the MIR research community have been made publicly available to facilitate the training, assessment and systematic comparison of music emotion recognition models. Developing musical mood annotation databases is a challenge for several reasons: as discussed in the previous section, the choice of emotion representation is not obvious, the task can be very time-consuming, ground truth annotations remain subjective, and often several labelers are required to reach for consistency. The CAL500 dataset comprises emotion labels for 500 songs by 500 unique artists [76]. Each song was annotated by three (non expert) reviewers using a set of 174 music tags, from which 18 were mood tags. The labellers annotated songs as a whole, rather than over time, a choice justified by the fact that mood is believed to be less prone to changes over time in popular music as opposed to classical music. The popular online music streaming service Last.fm has built up a "folksonomy" of 960 000 tags [31] (analytics from 2007) from which between 13% [80] and 20% [9], depending on the set of considered songs/artists, have been estimated to be related to mood. [80] published a dataset of 1804 tracks covering about 21 genres, with labels from the AMC5C mood tag clusters, derived from the AMG classification. [30] devised an online collaborative music mood annotation game, MoodSwings, where players annotate 30s-long music clips from the uspop2002 database [7], across time in the AV space. [70] built the "Now That's What I Call Music!" (NTWICM) database containing 2648 tracks from over five different genres (e.g. pop, rock, rap, R&B, electronic). Arousal and valence emotion annotations were conducted on 5-point Likert scales by four labelers. Eerola et al. [18,17] established a set of stimuli for the study of music-mediated emotions. A large pilot study established a set of 110 film music excerpts, half of which were moderately and highly representative examples of five discrete emotions ("anger", "fear", "sadness", "happiness", and "tenderness"), and the other half were moderate and high examples of the six extremes of three bipolar dimensions (valence, energy arousal and tension arousal).

3.2 Content- and Context-Based Features

Finding the acoustical clues predicting music emotions is one of the most challenging aspect in the development of music emotion recognition systems. Studies in music psychology [71], musicology [23] and music information retrieval [31] 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, while Tables 3 and 4 present the architectures of the associated content-based and multimodal MER models, respectively. Various acoustical correlates of articulation, dynamics, harmony, instrumentation, key, mode, pitch, register, rhythm, tempo, musical structure and timbre have been used in MER models. It can be seen from Table 2 that timbre features are the most commonly used in MER models. This is due to the fact that they have shown to provide the best performance in MER systems when used as individual features [66,93]. Indeed, Schmidt et al. investigated the use of multiple audio content-based features both individually and in combination in a feature fusion system [66,63]. They tested timbre descriptors (mel frequency cepstral coefficients, spectral centroid, spectral rolloff, spectral flux, octave-based spectral contrast, modeling peaks and gaps between harmonics), and chroma descriptors. 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 [93] (combination of rhythm, timbre and pitch features). Eerola et al. [18] 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, principal component analysis (PCA) followed by the selection of an optimal

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	[18]
Articulation/Timbre	ATTACS	Attack slope	[18]
Articulation/Timbre	ATTACT	Attack time	[18]
Dynamics	AVGENER	Average energy	[24]
Dynamics	INT	Intensity	[49]
Dynamics	INTR	Intensity ratio	[49]
Dynamics	DYN	Dynamics features	[58]
Dynamics	RMS	Root mean square energy	[18] [44] [58]
Dynamics	LOWENER	Low energy	[44]
Dynamics	ENER	Energy features	[45]
Harmony	OSPECENT	Octave spectrum entropy	[18]
Harmony	HARMC	Harmonic change	[18]
Harmony	CHROM	Chroma features	[66]
Harmony	HARMF	Harmony features	[58]
Harmony	RCHORDF	Relative chord frequency	[70]
Harmony	WCHORDD	Weighted chord differential	[44]
Instrum./Rhythm	PERCTO	Percussion template occurrence	[75]
Instrumentation	BASSTD	Bass-line template distance	[75]
Key/Mode	KEY	Key	[24]
Key/Mode	KEYC	Key clarity	[18]
Key/Mode	MAJ	Majorness	[18]
Key/Mode	SPITCH	Salient pitch	[18]
Key/Mode	WTON	Weighted tonality	[44]
Key/Mode	WTOND	Weighted tonality differential	[44]
Pitch/Melody	PITCHMIDI	Pitch MIDI features	[93]
Pitch/Melody	MELOMIDI	Melody MIDI features	[93]
Pitch/Melody	PITCH	Pitch features	[58]
Pitch/Timbre	ZCR	Zero-crossing rate	[93] [92]
Register	CHROMD	Chromagram deviation	[18]
Register	CHROMC	Chromagram centroid	[18]
Rhythm/Tempo	BEATINT	Beat interval	[24]
Rhythm/Tempo	SPECFLUCT	Spectrum fluctuation	[18]
Rhythm/Tempo	TEMP	Tempo	[18]
Rhythm/Tempo	PULSC	Pulse clarity	[18]
Rhythm/Tempo	RHYCONT	Rhythm content features	[93]
Rhythm/Tempo	RHYSTR	Rhythm strength	[49]
Rhythm/Tempo	CORRPEA	Correlation peak	[49]
Rhythm/Tempo	ONSF	Onset frequency	[49]
Rhythm/Tempo	RHYT	Rhythm features	[58]
Rhythm/Tempo	SCHERHYT	Scheirer rhythm features	[70]
Rhythm/Tempo	PERCF	Percussive features	[45]
Structure	MSTRUCT	Multidimensional structure features	[18]
Structure	STRUCT	Structure features	[58]
Timbre	HARMSTR	Harmonic strength	[24]
Timbre	MFCC	Mel frequency cepstral coefficient	[9] [6] [75] [93] [80] [58] [66] [63] [92] [77] [65] [58]
Timbre	SPECC	Spectral centroid	[9] [18] [93] [92] [64] [66] [49] [44] [70]
Timbre	SPECS	Spectral spread	[18]
Timbre	SPECENT	Spectral entropy	[18]
Timbre	SPECR	Spectral rolloff	[18] [93] [92] [64] [66] [49] [70]
Timbre	SF	Spectral flux	[93] [92] [64] [66] [49] [70]
Timbre	OBSC	Octave-based spectral contrast	[64] [66] [63] [65] [49] [38]
Timbre	RPEAKVAL	Ratio between average peak and valley strength	[49]
Timbre	ROUG	Roughness	[18]
Timbre	TIM	Timbre features	[58]
Timbre	SPEC	Spectral features	[45]
Timbre	ECNTT	Echo Nest timbre features	[65] [45]
Lyrics	SENTIWORD	Occurence of sentiment word	[14]
Lyrics	NEG-SENTIW	Occurrence of sentiment word with negation	[14]
Lyrics	MOD-SENTIW	Occurrence of sentiment word with modifier	[14]
Lyrics	WORDW	Word weight	[14]
Lyrics	LYRIC	Lyrics feature	[93]
Lyrics	RSTEMFR	Relative stem frequency	[70]
Lyrics	TF-IDF	Term frequency - Inverse document frequency	[14] [45]
Lyrics	RHYME	Rhyme feature	[81]
Context-based features			
Social tags	TAGS	Tag relevance score	[6]
Web-mined tags	DOCRS	Document relevance score	[6]
Metadata	ARTISTW	Artist weight	[14]
Metadata	META	Metadata features (e.g. artist's name, title)	[70]

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

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Lin et al. (2009) [42]	Audio	AMG (1535)	Cat. (AMG12C)	track	MARSYAS (436)	SVM	56.00%^{a}
Han et al. (2009) [24]	Audio	AMG (165)	Cat. (AV11C)	track	KEY, AVGENER, TEMP, σ(BEATINT), σ(HARMSTR)	**SVR**, SVM, GMM	94.55%^{b}
Eerola et al. (2009) [18]	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) [75]	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) [93]	Audio	Chin. & West. (24)	Cat. (AV4Q)	30s	**PITCH (5), RHYT (6), MFCCs (10), SSDs (9)**	**BAYN**	74.9%^{b}
Schmidt et al. (2010) [64]	Audio	MoodSwings Lite (240)	Dim. (AV)	1s	OBSC	MLR, LDS Kalman, LDS KALF, **LDS KALFM**	2.88^d
Schmidt et al. (2010) [66]	Audio	MoodSwings Lite (240)	Cat. (AV4Q) & Dim. (AV)	1s	**MFCCs**, CHROM (12), SSDs, OBSC	SVM / PLSR, **SVR**	0.137^e
Schmidt & Kim (2010) [63]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs**, ACORR(CHROM), SSDs, **OBSC**	MLR, PLSR, **SVR**	3.186 / 13.61^d
Myint & Pwint (2010) [49]	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) [38]	Audio	Clips (1000)	Dim. 2 (AV)	20s	OBSC	**SVM**	67.5%^{b}
Mann et al. (2011) [44]	Audio	TV theme tunes (144)	Dim. (6D-EPA)	track	RMS, LOWENER, SPECC, WTON, WTOND, WCHORDD, TEMP	**SVM**	80-94%^{b}
Vaizman et al. (2011) [77]	Audio	Piano, Vocal (76)	Cat. (4BE)	track	34 MFCCs	DTM	60%^{a}
Schmidt & Kim (2011) [65]	Audio	MoodSwings Lite (240)	Dim. (AV)	15s / 1s	**MFCCs (20)**, OBSC, ECNTTs (12)	MLR, **CRF**	0.122^f
Saari et al. (2011) [58]	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) [81]	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}: r^2. FSS: Feature subset selection. Feature notations are given in Table 2.

Reference	Modalities	Dtb (# songs)	Model (notation)	Decision hor.	Features (no.)	Machine learn.	Perf.
Dang & Shirai (2009) [14]	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) [9]	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) [6]	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) [80]	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) [93]	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) [70]	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) [45]	Audio, Lyrics	EchoNest API, lyricsmode.com, ANEW (119 664)	Dim.(AV)	track	TF-IDF, ECNT (65)	**CCA**	N/A

number of components, and partial least square regression (PLSR) with a Bayesian information criterion (BIC) to select the optimal number of features. PCA showed to be too sensitive to the covariance between the features and the predicted data. In contrast, 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 [20] 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 Sect. 4.6). Lyrics can indeed be semantically rich and expressive and have been shown to impact the way we perceive music [1]. 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 natural language processing (NLP) techniques. A standard way to represent text is to use a bag-of-words approach which characterises documents as vectors of words. To characterise the importance of a given word in a song given the corpus it belongs to, authors have used term frequency - inverse document frequency (TF-IDF) measure [14,45]. Methods to analyse emotions in lyrics have been developed using lexical resources for opinion and sentiment mining such as SentiWordNet (measures of positivity, negativity, objectivity) [14] and the affective norm for English words (measures of arousal, valence, and dominance) [45]. Since meaning emerges from subtle word combinations and sentence structure, research is still needed to develop new features characterising emotional meanings in lyrics. [81] proposed a feature to characterise rhymes whose patterns are relevant to emotion expression, as poems exemplify.

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. [14,9,6,80] 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 [9,6,80].

3.3 Temporal Aspects

MER models are also influenced by the duration of the audio segments chosen to make the classification decisions. Research on music emotions has shown that the fastest emotion-related responses take less than a second [13]. In [69], the author recommends a sampling rate of at least 2 Hz when collecting trace measurements. However, it is not clear yet to what extent such results depend on the material and dimension which are traced since some visual stimuli have been shown to evoke fear-related responses in the amygdala in about 12 ms [36]. However, most MER models rely on long term decision horizon (e.g. whole track [42,24,14,75,80,93,44,77,58,81,70,45], or 30-s long segment [9,6,93]). Algorithms identifying emotions on long term decision horizons are not bound to predict only a single emotion category per song since they may be associated with multi-label classification schemes, i.e. several emotion labels per decision (see Sect. 4). Other MER models use short-term decision horizons (e.g. 1 s [63,64,65,66]), in order to take into account the effects of music across time. Such an approach led to the development of methods for music emotion variation detection (see Sect. 4.5).

4 Machine Learning for Music Emotion Recognition

In most music information systems, emotion is seen as a high-level semantic feature. Thus the first step in utilising emotion-related information is devising a method that associates features from one or more of the above sources with mood categories or alternatively an emotion state in a continuous space. Machine learning techniques have become predominant for bridging this semantic gap. Initial approaches in MER were grounded on emotion recognition techniques developed for speech, or previous work within the MIR community on genre classification. Noting the similarity in architectural requirements, the first methods include the works of Feng et al. [21] and Li et al. [40]. Subsequently developed techniques can be characterised by their training method and expected outputs as follows: *multi-class single-label classification* (training samples are assigned a discrete emotion category, and the best estimate is chosen as output), *multi-label classification* (estimate multiple emotion categories simultaneously), *fuzzy classification* (probability estimates in each possible category), and *regression* (an estimate of emotion state in a continuous space).

From a high-level perspective, the first three approaches rely on a categorical model (Sect. 2.1) while regression relies on a dimensional model (Sect. 2.2). Given articles already covering early approaches to MER in detail (e.g. [31,50,86]), more emphasis is placed on state of the art and recent regression-based techniques in the following review.

4.1 Early Categorical Approaches

Associating music with discrete emotion categories was demonstrated by the first works that used an audio-based approach. Li et al. [40] 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. [21] used a Backpropagation 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.

4.2 Multi-label Classification

Early approaches demonstrate that content-based models of musical emotion are feasible. However, the ambiguity in the results 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. [84] applying a classifier specifically adopted to this task. The first systematic evaluation comparing several multi-label classification algorithms including Binary Relevance (BR), Label Powerset (LP), Random k-label sets (RAkEL) and Multi-Label k-Nearest Neighbours (MLkNN) was performed by Trohidis et al. [74], with RAkEL reaching 79% average precision using a dataset of 593 songs and simple rhythm and timbre features. In a recent study, Sanden and Zhang [59] 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. Besides the above algorithms, the authors tested Calibrated Label Ranking (CLR), Backpropagation for Multi-Label Learning (BPMLL), Hierarchy of Multi-Label Classifiers (HOMER), Instance-Based Logistic Regression (IBLR) and Binary Relevance 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 0.642 precision on CAL500). However, CLR with Decision Trees, BPMLL, and MLkNN also performed competitively.

4.3 Fuzzy Classification

Irrespective of considering induced or attributed emotion, people do not generally feel or perceive the same emotions. Several studies conclude that accommodating subjectivity is among the primary challenges in categorical emotion recognition models, while this was also demonstrated in a systematic evaluation using a non-categorical model [28]. A possible approach to account for subjectivity is the use of fuzzy classification incorporating fuzzy logic into conventional classification strategies. The work of Yang et al. [89] 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 category in a particular model. In [89], 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%. In some sense fuzzy classification may be seen as a special case of multi-label classification, but it is also a step towards continuous non-categorical models of emotion discussed in the next section.

4.4 Emotion Regression

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 [88], 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. A 114-dimensional feature space was constructed including spectral contrast features, wavelet coefficient histograms, as well as spectral (e.g. spectral centroid) and musicological (e.g. chords) features. After basic dimensionality reduction, 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. These results were then improved using feature selection.

Han et al. [24] 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 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. [66] show 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. [28] 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 was collected to indicate *induced* emotion, as in [88], 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. However, none of the variations in the experimental setup led to substantial improvement. 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 [85]. They demonstrated that the increased reliability of ground truth pays off when different learning algorithms are compared. In [87], 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%.

4.5 Methods for Music Emotion Variation Detection

The techniques discussed so far focus on detecting emotions from songs or short clips in a static manner. 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. [43]. In [89], the authors also proposed MEVD but by classifying features resulting from 10-s 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. [64] demonstrated that emotion distributions may be modeled as two-dimensional Gaussian distributions in the AV space, and then approached the problem of time-varying emotion tracking in two successive publications. In [64], 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 [65]. 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.

4.6 Multi-modal Approaches and Fusion Policies

When trying to account for the subjectivity of music related emotions, several factors other than audio may also be taken into account. Some of these factors, such as the acculturation of the listener, are extra-musical, or present in other modalities like lyrics. The combination of multiple feature domains has 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 [31]. 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.

Despite the need for a complex architecture, combining multiple modalities pays off well since different feature domains are often complementary. Bischoff et al. [9] 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 the case of audio features, while the Naive Bayes Multinomial classifier produced the best results in the 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. [41] 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 is used. Finally, Schuller et al. [70] 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 all modalities into account improved the results. However, they were not equally reliable, which promoted late fusion with a weighted combination of unimodal predictions. The decision between late and early fusion was not always clear cut however, since finding fusion weights was subject to overfitting.

5 Discussion and Conclusions

Although approaches relying on web social data and web documents are promising, they target commercial popular music repertoires for which web resources (e.g. blogs) are available and can be mined. Such approaches can't be applied straightforwardly to production music (music used in film, television, radio and other media, and often referred to as “mood music”) which don't benefit from the same media exposure as commercial music. The semantic analysis of lyrics offers promising perspectives, however it can't be applied to instrumental music, which represents a large corpus of classical and jazz music, alternative and progressive rock, and the most part of production music catalogues, for instance.

For such reasons, there is still a need to refine purely content-based methods, in addition to continuing development of hybrid approaches. [75] put forward the dominance of timbral features in music emotion recognition over pitch, and rhythmic features, for instance. As showed in Sect. 3, a large part of MER models rely on spectral timbre descriptors, such as the mel frequency cepstral coefficients, MFCCs, used in more than half of the studies reviewed hereby, as well as the octave-based spectral contrast (OBSC) and spectral descriptors, used in a third of the reviewed studies. This is related to the fact that spectral timbre descriptors have shown to provide the best correct classification rates when they were coupled with state of the art machine learning techniques in MER (see Table 3), audio music similarity (AMS) [4], as well as audio genre classification (AGC). However, as stated above, the results obtained by audio content-based systems are likely to be prone to a “glass ceiling” effect. In a recent study [58], 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 [71], psycho-musicology [23], and affective computing [48] 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 model the relation between features and emotion components as accurately as possible without taking into account the interpretability of the relationships, which is a disadvantage when trying to understand the underlying mechanisms [82]. Other emotion representation models - the appraisal models [48] - support the development of process models (see Sect. 2.4) which attempt to predict the association between appraisal and emotion components making it possible to interpret relationships.

With regard to machine learning techniques used in MER, the relatively low performance of classification approaches was commonly attributed to the weaknesses of the categorical emotion model discussed in Sect. 2.1 and 4.4. As a result, recent research focuses on the use of regression and attempt to estimate continuous valued coordinates in some emotion space, which may then be mapped to an emotion label or a broader category. Although these approaches seem to solve some of the problems related to classification, the decision between regression and classification is not yet straightforward, as both categorical and dimensional emotion models have strengths and weaknesses with regard to specific applications. Moreover, retrieving labels or categories given the estimated coordinates is often necessary, and requires a mapping between the dimensional and categorical models. This however may not be available for a given model, may not be psychologically validated in a given application, 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 can only be vanquished if both contextual features and features from different musical modalities are also considered.

Acknowledgments. This work was partly funded by the TSB project 12033-76187 "Making Musical Mood Metadata" (TS/J002283/1). The first author wishes to thank Christopher Jack for proofreading the article.

References

1. Ali, S.O., Peynirciogu, Z.F.: Songs and emotions: are lyrics and melodies equal partners? Psychology of Music 34(4), 511–534 (2006)
2. Arnold, M.B.: Emotion and personality. Columbia University Press, New York (1960)
3. Asmus, E.P.: Nine affective dimensions. Tech. rep., University of Miami (1986)
4. Aucouturier, J.J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences 1(1) (2004)
5. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636 (1996)
6. Barrington, L., Turnbull, D., Yazdani, M., Lanckriet, G.: Combining audio content and social context for semantic music discovery. In: Proc. of the ACM Special Interest Group on Information Retrieval, SIGIR (2009)
7. Berenzweig, A., Logan, B., Ellis, D., Whitman, B.: A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music Journal 28(2), 63–76 (2004)
8. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proc. of the ACM Conference on Information and Knowledge Management (CIKM), pp. 193–202 (2008)
9. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification - a hybrid approach. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 657–662 (2011)
10. Castellano, G., Caridakis, G., Camurri, A., Karpouzis, K., Volpe, G., Kollias, S.: Body gesture and facial expression analysis for automatic affect recognition. In: Scherer, K.R., Bänziger, T., Roesch, E.B. (eds.) Blueprint for Affective Computing: A Sourcebook, pp. 245–255. Oxford University Press, New York (2010)
11. Chion, M.: Audio-Vision: Sound On Screen. Columbia University Press (1994)
12. Cohen, A.J.: Music as a source of emotion in film. In: Music and Emotion Theory and Research, pp. 249–272. Oxford University Press (2001)
13. Cowie, R., McKeown, G., Douglas-Cowie, E.: Tracing emotion: an overview. International Journal of Synthetic Emotions 3(1), 1–17 (2012)
14. Dang, T.T., Shirai, K.: Machine learning approaches for mood classification of songs toward music search engine. In: Proc. of the International Conference on Knowledge and Systems Engineering (ICKSE), pp. 144–149 (2009)
15. Darwin, C.: The expression of the emotions in man and animals, 3rd edn. Harper-Collins (1998) (original work published 1872)
16. Davies, S., Allen, P., Mann, M., Cox, T.: Musical moods: a mass participation experiment for affective classification of music. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 741–746 (2011)
17. Eerola, T.: A comparison of the discrete and dimensional models of emotion in music. Psychology of Music 39(1), 18–49 (2010)
18. Eerola, T., Lartillot, O., Toiviainen, P.: Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: Proc. of the International Society for Music Information Retrieval (ISMIR) Conference (2009)
19. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1978)

20. Essid, S., Richard, G., David, B.: Musical instrument recognition by pairwise classification strategies. IEEE Trans. on Audio, Speech, and Language Proc. 14(4), 1401–1412 (2006)
21. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. In: Proc. ACM SIGIR, pp. 375–376 (2003)
22. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotions is not two-dimensional. Psychological Science 18(2), 1050–1057 (2007)
23. Gabrielsson, A.: The influence of musical structure on emotional expression, pp. 223–248. Oxford University Press (2001)
24. Han, B.J., Dannenberg, R.B., Hwang, E.: SMERS: music emotion recognition using support vector regression. In: Proc. of the 10th International Society for Music Information Retrieval (ISMIR) Conference, pp. 651–656 (2009)
25. Hevner, K.: Expression in music: a discussion of experimental studies and theories. Psychological Review 42(2), 186–204 (1935)
26. Hevner, K.: Experimental studies of the elements of expression in music. The American Journal of Psychology 48(2), 246–268 (1936)
27. Hu, X., Downie, J.S.: Exploring mood metadata: relationships with genre, artist and usage metadata. In: Proc. of the 8th International Conference on Music Information Retrieval (ISMIR), pp. 67–72 (2007)
28. Huq, A., Bello, J.P., Rowe, R.: Automated music emotion recognition: A systematic evaluation. Journal of New Music Research 39(3), 227–244 (2010)
29. Kim, J.H., Lee, S., Kim, S.M., Yoo, W.Y.: Music mood classification model based on Arousal-Valence values. In: Proc. of the 2nd International Conference on Advancements in Computing Technology (ICACT), pp. 292–295 (2011)
30. Kim, Y.E., Schmidt, E.M., Emelle, L.: Moodswings: A collaborative game for music mood label collection. In: Proc. of the International Society for Music Information Retrieval (ISMIR) Conference, pp. 231–236 (2008)
31. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G.: Music emotion recognition: a state of the art review. In: 11th International Society for Music Information Retrieval (ISMIR) Conference, pp. 255–266 (2010)
32. Kolozali, S., Fazekas, G., Barthet, M., Sandler, M.: Knowledge representation issues in musical instrument ontology design. In: 12th International Society for Music Information Retrieval Conference (ISMIR), Miami, USA, Florida, pp. 465–470 (2011)
33. Krumhansl, C.L.: An exploratory study of musical emotions and psychophysiology. Canadian Journal of Experimental Psychology 51(4), 336–353 (1997)
34. 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: Proc. of the LREC Workshop on Corpora for Research on Emotion and Affect, pp. 53–57. European Language Resources Association, Paris (2010)
35. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proc. of the Conference on Machine Learning and Applications (ICMLA), pp. 688–693 (2008)
36. LeDoux, J.E.: The emotional brain: the mysterious underpinnings of emotional life. Touchstone, New York (1998)
37. Lee, J.A., Downie, J.S.: Survey of music information needs, uses, and seeking behaviors: preliminary findings. In: Proc. of the 5th International Society for Music Information Retrieval (ISMIR) Conference, pp. 441–446 (2004)
38. Lee, S., Kim, J.H., Kim, S.M., Yoo, W.Y.: Smoodi: Mood-based music recommendation player. In: Proc. of the IEEE International Conference on Multimedia and Expo. (ICME), pp. 1–4 (2011)

39. Lesaffre, M., Leman, M., Martens, J.P.: A user oriented approach to music information retrieval. In: Proc. of the Content-Based Retrieval Conference (Published online), Daghstul Seminar Proceedings, Germany, Wadern (2006)
40. Li, T., Ogihara, M.: Detecting emotion in music. In: Proc. International Society of Music Information Retrieval Conference, pp. 239–240 (2003)
41. Lin, Y.C., Yang, Y.H., Chen, H.H.: Exploiting online music tags for music emotion classification. ACM Transactions on Multimedia Computing Communications and Applications 7S(1), 26:1–26:15 (2011)
42. Lin, Y.C., Yang, Y.H., Chen, H.H., Liao, I.B., Ho, Y.C.: Exploiting genre for music emotion classification. In: Proc. of the IEEE International Conference on Multimedia and Expo. (ICME), pp. 618–621 (2009)
43. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. on Audio, Speech, and Language Proc. 14(1), 5–18 (2006)
44. Mann, M., Cox, T.J., Li, F.F.: Music mood classification of television theme tunes. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 735–740 (2011)
45. McVicar, M., Freeman, T., De Bie, T.: Mining the correlation between lyrical and audio features and the emergence of mood. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 783–788 (2011)
46. Meyer, L.B.: Emotion and meaning in music. The University of Chicago press (1956)
47. MIREX: Audio mood classification (AMC) results (2009), `http://www.music-ir.org/mirex/wiki/2009:Audio_Music_Mood_Classification_Results`
48. Mortillaro, M., Meuleman, B., Scherer, R.: Advocating a componential appraisal model to guide emotion recognition. International Journal of Synthetic Emotions 3(1), 18–32 (2012)
49. Myint, E.E.P., Pwint, M.: An approach for multi-label music mood classification. In: 2nd International Conference on Signal Processing Systems (ICSPS), vol. VI, pp. 290–294 (2010)
50. Ogihara, M., Kim, Y.: Mood and emotional classification. In: Music Data Mining. CRC Press (2011)
51. Osgood, C.E., May, W.H., Miron, M.S.: Cross-Cultural Universals of Affective Meaning. University of Illinois Press, Urbana (1975)
52. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning. University of Illinois Press, Urbana (1957)
53. Parke, R., Chew, E., Kyriakakis, C.: Quantitative and visual analysis of the impact of music on perceived emotion of film. Computers in Entertainment (CIE) 5(3) (2007)
54. Raimond, Y., Abdallah, S., Sandler, M., Frederick, G.: The music ontology. In: Proc. of the 7th International Conference on Music Information Retrieval (ISMIR), Vienna, Austria, pp. 417–422 (2007)
55. Raimond, Y., Giasson, F., Jacobson, K., Fazekas, G., Gangler, T.: Music ontology specification (November 2010), `http://musicontology.com/`
56. Roseman, I.J., Smith, C.A.: Appraisal theory: Overview, assumptions, varieties, controversies. In: Scherer, K.R., Schorr, A., Johnstone, T. (eds.) Appraisal Processes in Emotion: Theory, Methods, Research, pp. 3–19. Oxford University Press, New York (2001)
57. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)

58. 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 Language Proc. 19(6), 1802–1812 (2011)
59. Sanden, C., Zhang, J.: An empirical study of multi-label classifiers for music tag annotation. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 717–722 (2011)
60. Scherer, K.R., Brosch, T.: Culture-specific appraial biases contribute to emotion disposition. European Journal of Personality 288, 265–288 (2009)
61. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal processes in emotion: Theory, methods, research. Oxford University Press, New York (2001)
62. Schlosberg, H.: The description of facial expressions in terms of two dimensions. Journal of Experimental Psychology 44, 229–237 (1952)
63. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions from audio. In: Proc. of the 11th International Society for Music Information Retrieval (ISMIR) Conference, pp. 465–470 (2010)
64. Schmidt, E.M., Kim, Y.E.: Prediction of time-varying musical mood distributions using Kalman filtering. In: Proc. of the 9th International Conference on Machine Learning and Applications (ICMLA), pp. 655–660 (2010)
65. Schmidt, E.M., Kim, Y.E.: Modeling musical emotion dynamics with conditional random fields. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 777–782 (2011)
66. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proc. of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval (MIR), pp. 267–273 (2010)
67. Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. Australian Journal of Psychology 51(3), 154–165 (1999)
68. Schubert, E.: Update of the Hevner adjective checklist. Perceptual and Motor Skills, pp. 117–1122 (2003)
69. Schubert, E.: Continuous self-report methods. In: Juslin, P.N., Sloboda, J.A. (eds.) Handbook of Music and Emotion, pp. 223–253. Oxford University Press (2010)
70. Schuller, B., Weninger, F., Dorfner, J.: Multi-modal non-prototypical music mood analysis in continous space: reliability and performances. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 759–764 (2011)
71. Sloboda, J.A., Juslin, P.N.: Psychological perspectives on music and emotion. In: Juslin, P.N., Sloboda, J.A. (eds.) Music and Emotion Theory and Research. Series in Affective Science, pp. 71–104. Oxford University Press (2001)
72. Thayer, J.F.: Multiple indicators of affective responses to music. Dissertation Abstracts International 47(12) (1986)
73. Thompson, W.F., Robitaille, B.: Can composers express emotions through music? Empirical Studies of the Arts 10(1), 79–89 (1992)
74. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. In: Proc. International Society of Music Information Retrieval Conference, pp. 325–330 (2008)
75. Tsunoo, E., Akase, T., Ono, N., Sagayama, S.: Music mood classification by rhythm and bass-line unit pattern analysis. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 265–268 (2010)
76. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Towards musical query by semantic description using the CAL500 data set. In: Proc. of the ACM Special Interest Group on Information Retrieval (SIGIR), pp. 439–446 (2007)

77. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling dynamic patterns for emotional content in music. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 747–752 (2011)
78. Vuoskoski, J.K.: Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. Musicae Scientiae 15(2), 159–173 (2011)
79. Waletzky, J.: Bernard Hermann: Music For the Movies. DVD Les Films d'Ici / Alternative Current (1992)
80. Wang, J., Anguerra, X., Chen, X., Yang, D.: Enriching music mood annotation by semantic association reasoning. In: Proc. of the International Conference on Multimedia, pp. 1445–1450 (2010)
81. Wang, X., Chen, X., Yang, D., Wu, Y.: Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp. 765–770 (2011)
82. Wehrle, T., Scherer, K.R.: Toward computational modelling of appraisal theories. In: Scherer, K.R., Schorr, A., Johnstone, T. (eds.) Appraisal Processes in Emotion: Theory, Methods, Research, pp. 92–120. Oxford University Press, New York (2001)
83. Whissell, C.M.: The dictionary of affect in language. In: Plutchik, R., Kellerman, H. (eds.) Emotion: Theory Research and Experience, vol. 4, pp. 113–131. Academic Press, New York (1989)
84. Wieczorkowska, A., Synak, P., Ras, Z.W.: Multi-label classification of emotions in music. In: Proc. of Intelligent Information Processing and Web Mining, pp. 307–315 (2006)
85. Yang, Y.H., Chen, H.H.: Ranking-based emotion recognition for music organisation and retrieval. IEEE Trans. on Audio, Speech, and Language Proc. 19(4), 762–774 (2010)
86. Yang, Y.H., Chen, H.H.: Music emotion recognition. In: Multimedia Computing. Communication and Intelligence Series. CRC Press (2011)
87. Yang, Y.H., Chen, H.H.: Prediction of the distribution of perceived music emotions using discrete samples. IEEE Trans. on Audio, Speech, and Language Proc. 19(7), 2184–2195 (2011)
88. 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 Language Proc. 16(2), 448–457 (2008)
89. Yang, Y.H., Liu, C.C., Chen, H.H.: Music emotion classification: A fuzzy approach. In: Proc. of the 14th Annual ACM International Conference on Multimedia, Santa Barbara, CA, USA, pp. 81–84 (2006)
90. Yoo, M.J., Lee, I.K.: Affecticon: emotion-based icons for music retrieval. IEEE Computer Graphics and Applications 31(3), 89–95 (2011)
91. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: Differentiation, classification, and measurement. Emotion 8(4), 494–521 (2008)
92. Zhao, Y., Yang, D., Chen, X.: Multi-modal music mood classification using co-training. In: International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1–4 (2010)
93. Zhao, Z., Xie, L., Liu, J., Wu, W.: The analysis of mood taxonomy comparison between Chinese and Western music. In: Proc. of the 2nd International Conference on Signal Processing Systems (ICSPS), vol. VI, pp. 606–610 (2010)