Machine Recognition of Music Emotion: A Review

YI-HSUAN YANG and HOMER H. CHEN, National Taiwan University

The proliferation of MP3 players and the exploding amount of digital music content call for novel ways of music organization and retrieval to meet the ever-increasing demand for easy and effective information access. As almost every music piece is created to convey emotion, music organization and retrieval by emotion is a reasonable way of accessing music information. A good deal of effort has been made in the music information retrieval community to train a machine to automatically recognize the emotion of a music signal. A central issue of machine recognition of music emotion is the conceptualization of emotion and the associated emotion taxonomy. Different viewpoints on this issue have led to the proposal of different ways of emotion annotation, model training, and result visualization. This article provides a comprehensive review of the methods that have been proposed for music emotion recognition. Moreover, as music emotion recognition is still in its infancy, there are many open issues. We review the solutions that have been proposed to address these issues and conclude with suggestions for further research.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Systems, modeling*

General Terms: Algorithms, Performance, Human Factors

Additional Key Words and Phrases: Music emotion recognition

ACM Reference Format:
Yang, Y.-H. and Chen, H. H. 2012. Machine recognition of music emotion: A review. ACM Trans. Intell. Syst. Technol. 3, 3, Article 40 (May 2012), 30 pages.
DOI = 10.1145/2168752.2168754 http://doi.acm.org/10.1145/2168752.2168754

1. INTRODUCTION

Music plays an important role in human history, even more so in the digital age. Never before has such a large collection of music been created and accessed daily. The popularity of the Internet and the use of compact audio formats with near-CD quality, such as MP3 (MPEG-1 Audio Layer 3), have expedited the growth of digital music libraries [Wieczorkowska et al. 2006]. The prevailing context in which we encounter music is now ubiquitous, including those contexts in which the most routine activities of life take place: waking up, eating, housekeeping, shopping, studying, exercising, driving, and so forth [Juslin and Sloboda 2001]. Music is everywhere. As the amount of content continues to explode, conventional approaches that manage music pieces based on catalogue metadata, such as artist name, album name, and song title, are longer

This work was supported in part by grants from the National Science Council of Taiwan under contracts NSC 100-2221-E-002-198-MY3 and NSC 100-2218-E-001-009 and in part by grants from the National Taiwan University under contracts 10R80919-5 and 101R7609-2.

Y.-H. Yang is currently affiliated with Academia Sinica.
Authors' addresses: Y.-H. Yang, Academia Sinica, Taiwan R.O.C., 10617; email: yang@citi.sinica.edu.tw; H. H. Chen, Graduate Institute of Communication Engineering, National Taiwan University, Taiwan R.O.C., 10617; email: homer@cc.ee.ntu.edu.tw.

DOI 10.1145/2168752.2168754 http://doi.acm.org/10.1145/2168752.2168754

sufficient. The way that music information is organized and retrieved has to evolve in order to meet the ever-increasing demand for easy and effective information access [Casey et al. 2008].

In response to this demand, music organization and retrieval by emotion has received growing attention in the past few years. Since the preeminent functions of music are social and psychological, and since almost every music piece is created to convey emotion, music organization and retrieval by emotion has been considered a reasonable way of accessing music information [Feng et al. 2003; Huron 2000]. It is generally believed that music cannot be composed, performed, or listened to without affection involvement [Sloboda and Juslin 2001]. Music can bring us to tears, console us when we are in grief, or drive us to love. Music information behavior studies have also identified emotion as an important criterion used by people in music seeking and organization. According to a study of social tagging on Last.fm[1], a popular commercial music website, emotion tag is the third most frequent type of tag (second to genre and locale) assigned to music pieces by online users [Lamere 2008]. Despite the idea of emotion-based music retrieval being new at the time, a survey conducted in 2004 showed that about 28.2% of participants identified emotion as an important criterion in music seeking and organization [Laurier et al. 2004; Lee and Downie 2004]. Consequently, emotion-based music retrieval has received increasing attention in both academia and the industry [Huq et al. 2009; Lu et al. 2006; Yang et al. 2008; Yang and Chen 2011c].

In academia, more and more multimedia systems that involve emotion analysis of music signals have been developed, such as Moodtrack [Vercoe 2006], LyQ [Hsu and Hsu 2006], MusicSense [Cai et al. 2007], Mood Cloud [Laurier and Herrera 2008; Laurier et al. 2009], Moody [Hu et al. 2008], and *i*.MTV [Zhang et al. 2008, 2009], just to name a few. In the industry, many music companies, such as AMG, Gracenote, MoodLogic, Musicovery, Syntonetic, and Sourcetone[2] use emotion as a cue for music retrieval. For example, the Gracenote mood taxonomy consists of over 300 highly specific mood categories, which are organized hierarchically with broader mood categories at the top-level that are made up of several more-specific sub-mood categories. Mood metadata is automatically derived by using Gracenote's proprietary content analysis and machine learning technologies, without any manual labeling or user input. This mood metadata provides an additional criterion by which users can organize and retrieve music in a content-based fashion. A user is able to organize their music collections by various mood categories represented by affective adjectives such as "peaceful," "romantic," "sentimental," "defiant," "fiery," and "easygoing."

Making computers capable of recognizing the emotion of music also enhances the way humans and computers interacts. It is possible to playback music that matches the user's moods detected from physiological, prosodic, or facial cues [Anderson and McOwan 2006; Jonghwa and Ande 2008; Lee and Narayanan 2005; Lin et al. 2009; Picard et al. 2001]. A portable device, such as an MP3 player or a cellular phone equipped with an automatic music emotion recognition (MER) function, can then play a song best suited to the emotional state of the user [Dornbush et al. 2005; Reddy and Mascia 2006]. A smart space (e.g., restaurant, conference room, residence) can play background music best suited to the people inside it [Jaimes and Sebe 2005; Lew et al. 2006]. For example, Wu et al. [2008] proposed an interactive content presenter based on the perceived emotion of multimedia content and the physiological feedback of the user. Multimedia content (photos, music, and Web blog articles) are automatically classified into eight emotion classes (happy, light, easy, touching, sad, sublime, grand, and

[1] http://www.last.fm/

[2] http://www.allmusic.com/, http://www.gracenote.com/, http://www.moodlogic.com, http://www.musicovery.com/, http://www.syntonetic.com/, http://www.sourcetone.com/

exciting) [Wu and Jeng 2008] and then organized in a tiling slideshow fashion [Chen et al. 2006] to create music videos (MVs) [Chen et al. 2008]. The user's preference of these MVs is detected from physiological signals, such as blood pressure and skin conductance, and then utilized to recommend the next MV. This retrieval paradigm is functionally powerful since people's criteria for music selection are often related to the emotional state at the moment of music selection [Juslin and Sloboda 2001].

A considerable amount of work has been done in the music information retrieval (MIR) community for automatic recognition of the perceived emotion of music.[3] A typical approach to MER categorizes emotions into a number of classes (such as happy, angry, sad, and relaxed) and applies machine learning techniques to train a classifier [Katayose et al. 1998; Kim et al. 2010; Laar 2006; Liu et al. 2006; Lu et al. 2006; Schuller et al. 2010]. Usually, timbre, rhythm, and harmony features of music are extracted to represent the acoustic property of a music piece. Typically, a subjective test is conducted to collect the ground truth needed for training the computational model of emotion prediction. Several machine learning algorithms have been applied to learn the relationship between music features and emotion labels, such as support vector machines [Bischoff et al. 2009; Li and Ogihara 2003; Hu et al. 2008; Wang et al. 2004], Gaussian mixture models [Liu et al. 2003], neural networks [Feng et al. 2003], boosting [Lu et al. 2010], and *k*-nearest neighbor [Wieczorkowska 2004; Yang et al. 2006]. After training, the automatic model can be applied to recognize the emotion of an input music piece.

Because of the multidisciplinary nature of MER and the wide variety of approaches that have been developed for MER, it is often difficult for a researcher to identify the structure of this field and gain insight into the state-of-the-art. The goal of this article is therefore to provide a comprehensive review of the MER literature and to discuss some possible directions of future research.

A central issue of MER is the conceptualization of emotion and the associated emotion taxonomy. There is still no consensus on which emotion model or how many emotion categories should be used. In addition, there is debate over whether emotions should be conceptualized as categories or continua. Different viewpoints on this issue have led to proposals of different ways of emotion annotation, model training, and result visualization.[4] As shown in Table I, existing work on MER can be classified into three approaches. The *categorical* approach to MER categorizes emotions into a number of discrete classes and applies machine learning techniques to train a classifier. The predicted emotion labels can be incorporated into a text-based or metadata-based music retrieval system. The *dimensional* approach to MER defines emotions as numerical values over a number of emotion dimensions (e.g., valence and arousal [Russell 1980]). A regression model is trained to predict the emotion values that represent the affective content of a song, thereby representing the song as a point in an emotion space. Users can then organize, browse, and retrieve music pieces in the

[3] In psychological studies, emotions are often divided into three categories: *expressed* emotion, *perceived* emotion, and *felt* (or evoked) emotion [Gabrielsson 2002]. The first one refers to the emotion the performer tries to communicate with the listeners, while the latter two refer to the emotional responses of the listeners. We may simply perceive an emotion being expressed in a song (emotion perception) or actually feel an emotion in response to the song (emotion induction). Both perceived emotion and felt emotion, especially the latter, are dependent on an interplay between the musical, personal, and situational factors [Gabrielsson 2002]. MIR researchers tend to focus on the perceived emotion, for it is relatively less influenced by the situational factors (environment, mood, etc.) of listening [Yang et al. 2008].

[4] While MIR researchers tend to use the terms "emotion" and "mood" interchangeably, a clear distinction of the two terms is often made by psychologists. Emotion is usually understood as a short experience in response to an object (here, music), whereas mood is a longer experience without specific object connection [Sloboda and Juslin 2001]. We fix our term of emotion according to the definition of the psychologists. Moreover, we use "music emotion" as a short term for "emotions that were perceived in music."

Table I. Comparison of Existing Work on Automatic Music Emotion Recognition

Methodology	Emotion conceptualization	Description
Categorical MER	Categorical	Predicting the *discrete emotion labels* of music pieces [Hu et al. 2008; Lu et al. 2006]
Dimensional MER	Dimensional	Predicting the *numerical emotion values* of music pieces [Eerola et al. 2009; Yang et al. 2008]
MEVD	Dimensional	Predicting the *continuous emotion variation* within a music piece [Korhonen et al. 2006; Schmidt et al. 2010]

emotion space, which provides a simple means for user interface. Finally, instead of predicting an emotion label or value that represents a song, *music emotion variation detection* (MEVD) focuses on the dynamic process of music emotion and makes emotion predictions for every short-time segment of a song, resulting in a series of emotion predictions. Users can then track the emotion variation of a song as time unfolds. To give the readers a sense of how emotions are usually conceptualized, we first discuss the emotion models that have been proposed by psychologists in Section 2. Section 3 is dedicated to music features that are often utilized to model music emotion. We then review existing works on the three methodologies of MER in Sections 4, 5, and 6, respectively.

Regardless of the approach taken, MER is a challenging task because of the following reasons. First, emotion perception is by nature subjective, and people can perceive different emotions for the same song. This subjectivity makes performance evaluation of an MER system fundamentally difficult because a common agreement on the recognition result is hard to obtain. Second, emotion annotations are usually difficult to obtain, especially when there is still no consensus on emotion taxonomy. Third, it is still inexplicable how music represents emotion. What intrinsic element of music, if any, creates a specific emotional response in the listener is still far from being well-understood. We discuss these challenges in detail and review the solutions that have been proposed to address these issues in Section 7. We conclude the article in Section 8 with suggestions for future research.

2. EMOTION CONCEPTUALIZATION

In the study of emotion conceptualization, psychologists often utilize people's verbal reports of emotion responses [Juslin and Sloboda 2001]. For example, the celebrated paper of Hevner [1935] studied the relationship between music and emotion through experiments in which subjects were asked to report the adjectives that came to their minds as the most representative part of a music piece played. From these empirical studies, a great variety of emotion models have been proposed, most of which belong to one of the following two approaches to emotion conceptualization: the categorical approach and the dimensional approach.

2.1. Categorical Conceptualization of Emotion

According to the categorical approach, people experience emotions as categories that are distinct from each other. Essential to this approach is the concept of basic emotions, that is, the idea that there is a limited number of universal and primary emotion classes, such as happiness, sadness, anger, fear, disgust, and surprise, from which all other secondary emotion classes can be derived [Ekman 1992; Picard et al. 2001]. Each basic emotion can be defined functionally in terms of a key appraisal of goal-relevant events that have occurred frequently during evolution. The basic emotions can be found in all cultures, and they are often associated with distinct patterns of physiological changes or emotional expressions. The notion of basic emotions, however,

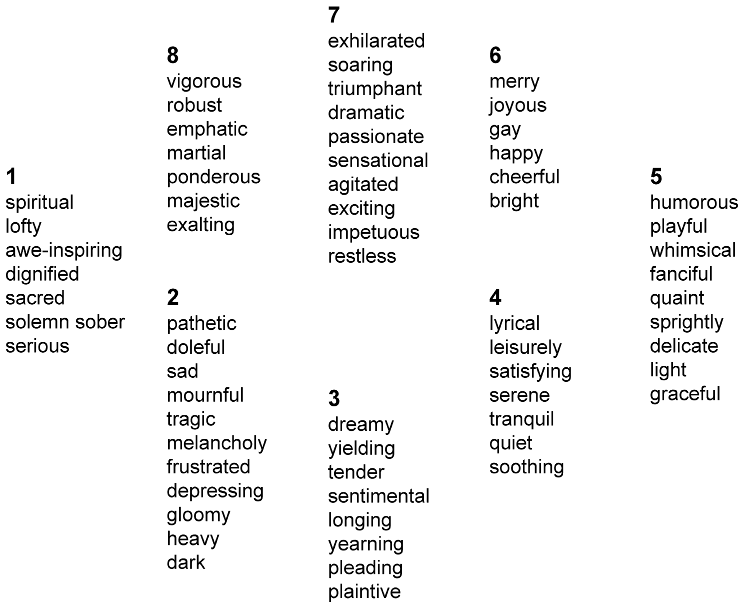


Fig. 1. Hevner's eight clusters of affective terms [Hevner 1935].

has been criticized on a number of grounds, most notably because different researchers have come up with different sets of basic emotions [Sloboda and Juslin 2001].

Another famous categorical approach to emotion conceptualization is Hevner's adjective checklist [Hevner 1935]. Through experiments, she discovered eight clusters of affective adjectives and laid them out in a circle, as shown in Figure 1. The adjectives within each cluster are similar, and the meaning of neighboring clusters varies in a cumulative way until reaching a contrast in the opposite position. The Hevner adjectives (proposed in 1935) were later refined and regrouped into ten adjective groups by Farnsworth [1954] and into nine adjective groups by Schubert [2003].

The major drawback of the categorical approach is that the number of primary emotion classes is too small in comparison with the richness of music emotion perceived by humans. Using a finer granularity, on other hand, does not necessarily solve the problem, because the language for describing emotions is inherently ambiguous and varies from person to person [Juslin and Laukka 2004]. Moreover, using a large number of emotion classes could overwhelm the subjects and is impractical for psychological studies [Sloboda and Juslin 2001].

2.2. Dimensional Conceptualization of Emotion

While the categorical approach focuses mainly on the characteristics that distinguish emotions from one another, the dimensional approach focuses on identifying emotions based on their placement on a small number of emotion dimensions with named axes, which are intended to correspond to internal human representations of emotion. These internal emotion dimensions are found by analyzing the correlation between affective terms. Subjects are asked to use a large number of rating scales of affective terms to describe the emotion of music stimulus. Factor analysis techniques are then employed to obtain a small number of fundamental factors (dimensions) from the correlations between the scales. Although differing in names, existing psychological studies give very similar interpretations of the resulting factors. Most of them correspond to the

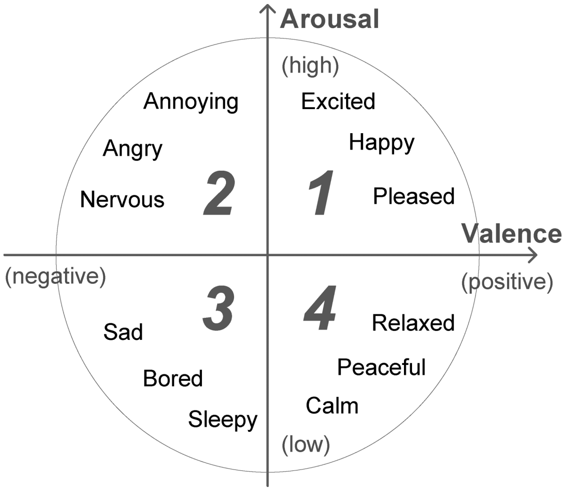


Fig. 2. The 2D valence-arousal emotion space [Russell 1980] (the position of the affective terms are only approximated, not exact).

following three dimensions of emotion: *valence* (or pleasantness; positive and negative affective states), *arousal* (or activation; energy and stimulation level), and *potency* (or dominance; a sense of control or freedom to act) [Osgood et al. 1957; Plutchik 1980; Remington et al. 2000; Thayer 1989; Whissell et al. 1986].

In the seminal work of Russell [1980], the *circumplex* model of emotion is proposed. The model consists of a two-dimensional, circular structure involving the dimensions of valence and arousal, as shown in Figure 2. Within this structure, emotions that are inversely correlated are placed across the circle from one another. Supportive evidence was obtained by scaling 28 affective terms in four different ways: Ross' technique [Ross 1938] for a circular ordering of variables, a multidimensional scaling procedure based on perceived similarity among the terms, a unidimensional scaling on hypothesized pleasure-displeasure and degree-of-arousal dimensions, and a principal component analysis [Duda et al. 2000] of 343 subjects' self-reports of their current affective states [Russell 1980]. All these methods result in a valence-arousal, circular-structure arrangement of the 28 terms. The circumplex model is also referred to as the two-dimensional emotion space (2DES).

One of the strengths of the circumplex model is that it suggests a simple yet powerful way of organizing different emotions in terms of their affect appraisals (valence) and physiological reactions (arousal), and it allows for direct comparison of different emotions on two standard and important dimensions. Juslin and Sloboda [2001] note the following.

> From a theoretical point of view one can argue that activation or arousal variation is one of the major distinctive features of emotion, and the valence dimension, the pervasive pleasant-unpleasant quality of experience, maps directly into the classic approach-avoidance action tendencies that have direct relevance for behavior. Recently, Russell even went as far as claiming that valence and arousal are the "core processes" of affect, constituting the raw material or primitive of emotional experience [Russell 2003].

Describing emotions by the two-dimensional model, however, is not free of criticism. It has been argued that it blurs important psychological distinctions and consequently obscures important aspects of the emotion process [Lazarus 1991]. For example, anger and fear are placed close to each other in the valence-arousal plane (both in the second quadrant), but they are very different in terms of their implications for the organism. In response to this deficiency, some researchers have advocated the use of potency (dominant–submissive) as the third dimension in order to obtain a more complete picture of emotion [Bigand et al. 2005; Collier 2007]. Following this idea, Eerola et al. [2009] build a regression-based computational model to predict the perceived emotion of movie soundtracks over the three-dimensional emotion space (3DES). This approach, however, requires subjects to annotate emotion in 3D, which is difficult to perform. It is also more difficult to visualize music in 3D rather than in 2D, especially on mobile devices. A two-dimensional model appears to offer a better balance between a parsimonious definition of emotion and limiting the complexity of the task [Sloboda and Juslin 2001].

3. MUSIC FEATURES

The experience of music listening is multidimensional. Different emotion perceptions of music are usually associated with different patterns of acoustic cues [Juslin 2000; Krumhansl 2002]. For example, while arousal is related to tempo (fast/slow), pitch (high/low), loudness level (high/low), and timbre (bright/soft), valence is related to mode (major/minor) and harmony (consonant/dissonant) [Gabrielsson and Lindström 2001]. It is also noted that emotion perception is rarely dependent on a single music factor but a combination of them [Hevner 1935; Rigg 1964]. For example, loud chords and high-pitched chords may suggest more positive valence than soft chords and low-pitched chords, irrespective of mode. See Gabrielsson and Lindström [2001] for an overview of the empirical research concerning the influence of different music factors on emotion perception. Next, we briefly review some features that have been utilized in MER.

3.1. Energy

The energy of a song is highly correlated to arousal perception [Gabrielsson and Lindström 2001]. The sound description toolbox can be employed to extract a number of energy-related features, including audio power, total loudness, and specific loudness sensation coefficients (SONE) [Benetos et al. 2007]. Audio power is simply the power of the audio signal. The extraction of total loudness and SONE is based on the perceptual models implemented in the MA Toolbox [Pampalk 2004], including an outer-ear model, the Bark critical-band rate scale (psycho-acoustically motivated critical bands), and spectral masking (by applying spreading functions). The resulting power spectrum, which better reflects human loudness sensation, is called the sonogram. SONE is the coefficients computed from the sonogram, which consists up to 24 Bark critical bands (the actual number of critical bands depends on the sampling frequency of the audio signal). Energy features have been utilized in Lu et al. [2006] to classify arousal.

3.2. Rhythm

Rhythm is the pattern of pulses/notes of varying strength. It is often described in terms of tempo, meter, or phrasing. A song with a fast tempo is often perceived as having high arousal. Besides, flowing/fluent rhythm is usually associated with positive valence, while firm rhythms with negative valence [Gabrielsson and Lindström 2001]. To describe the rhythmic property of music, the following five features are proposed in Lu et al. [2006]: rhythm strength, rhythm regularity, rhythm clarity, average onset frequency, and average tempo. Rhythm strength is the average onset strength in the onset detection curve, which can be computed based on the algorithm described in Klapuri [1999]. Rhythm regularity and clarity are computed by performing autocorrelation on the onset detection curve. If a music segment has an obvious and regular rhythm, the peaks of the corresponding autocorrelation curve will be obvious and strong as well. Onset frequency, or event density, is calculated as the number of note onsets per second, while tempo is estimated by detecting periodicity from the onset detection curve [Lartillot and Toiviainen 2007].

3.3. Melody

The MIR toolbox [Lartillot and Toiviainen 2007] can be employed to generate two pitch features (salient pitch and chromagram center) and three tonality features (key clarity, mode, harmonic change). MIR toolbox estimates the pitch, or the perceived fundamental frequency, of each short time frame (50 ms, 1/2 overlapping) based on the multipitch detection algorithm described in Tolonen and Karjalainen [2000]. The algorithm decomposes an audio waveform into two frequency bands (below and above 1 kHz), computes the autocorrelation function of the envelop in each subband, and finally produces pitch estimates by picking the peaks from the sum of the two autocorrelation functions. The pitch estimate corresponding to the highest peak is returned as the salient pitch. MIR toolbox also computes the wrapped chromagram, or the pitch class profile, for each frame (100 ms, 1/8 overlapping) and uses the centroid (center of gravity) of the chromagram as another estimate of the fundamental frequency. This feature is called the chromagram centroid. A wrapped chromagram projects the frequency spectrum onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. For example, frequency bins of the spectrum around 440 Hz (C4) and 880 Hz (C5) are all mapped to chroma C. Therefore, chromagram centroid may be regarded as a pitch estimate that does not consider absolute frequency. Each bin of the chromagram corresponds to one of the twelve semitone classes in the Western twelve-tone equal-temperament scale. By comparing a chromagram with the 24 major and minor key profiles [Gómez 2006], the strength of the frame in association with each key (e.g., C major) is estimated. The strength associated with the best key, that is, the one with the highest strength, is returned as the key clarity. The difference between the best major key and the best minor key in key strength is returned as the estimate of the musical mode. The PsySound toolbox [Cabrera 1999] can also be employed to compute pitch features estimated by the sawtooth waveform inspired pitch (SWIPE) estimator [Camacho 2007].

3.4. Timbre

A commonly used timbre feature is Mel-frequency cepstral coefficients (MFCC), the coefficients of the discrete cosine transform of each short-term log power spectrum expressed on a nonlinear perceptual-related Mel-frequency scale [Casey et al. 2008; Davis and Mermelstein 1980]. It represents the formant peaks of the spectrum. MFCC can be extracted by the MA toolbox [Pampalk 2004] or Marsyas [Tzanetakis and Cook 2002]. A drawback of MFCC, however, is that it averages the spectral distribution in each subband and therefore loses the relative spectral information. Octave-based spectral contrast is proposed in Jiang et al. [2002] to capture the relative energy distribution of the harmonic components in the spectrum . The feature considers the spectral peak, spectral valley, and their dynamics in each subband and roughly reflects the relative distribution of the harmonic and nonharmonic components in the spectrum.

Another timbre feature that is often used in previous MER work is DWCH, or the Daubechies wavelets coefficient histogram, which better ability in representing both the local and global information of the spectrum [Li and Ogihara 2003, 2004]. It is computed from the histograms of Daubechies wavelet coefficients at different frequency subbands with different resolutions.

One can also use the MIR toolbox [Lartillot and Toiviainen 2007] to extract three features related to the sensory dissonance of music (roughness, irregularity, inharmonicity). Roughness, or spectral dissonance, measures the noisiness of the spectrum; any peak of the spectrum that does not fall within the prevailing harmony is considered dissonant. The irregularity measures the degree of variation of the successive peaks of the spectrum. The inharmonicity estimates the amount of partials that departs from multiples of the fundamental frequency. The coefficient ranges from 0 (purely harmonic signal) to 1 (inharmonic signal).

The Marsyas software [Tzanetakis and Cook 2002] can also be employed to extract spectral flatness measures (SFM) and spectral crest factors (SCF), which are both related to the noisiness of audio signal [Allamanche et al. 2001]. SFM is the ratio between the geometric mean of the power spectrum and its arithmetic mean, whereas SCF is the ratio between the peak amplitude and the root-mean-square amplitude. They are extracted by computing the values in 12 subbands for each short time frame (23 ms, 1/2 overlapping) and then taking the means and standard deviations over a sliding texture window of one second. The sequence of feature vectors is then collapsed into a single vector representing the entire signal by taking again the mean and standard deviation [Tzanetakis and Cook 2002].

4. CATEGORICAL MUSIC EMOTION RECOGNITION

A great many efforts have been made by MIR researchers to automate MER, and the type of music under study has gradually shifted over the past few years from symbolic music [Katayose et al. 1998; Livingstone and Brown 2005; Wang et al. 2004; Yeh et al. 2006] to raw audio signals and from Western classical music [Korhonen et al. 2006; Li and Ogihara 2003; Livingstone and Brown 2005; Lu et al. 2006; Wang et al. 2004] to popular music. Western classical music is often chosen in the early studies partly because of the rich literature in musicology and psychology on classical music, and partly because it seems to be easier to gain agreement on the perceived emotion of a classical music selection. However, since the purpose of MER is to facilitate music retrieval and management in everyday music listening, and since it is popular music that dominates everyday music listening, analyzing the affective content of popular music has gained increasing attention lately.

The categorical approach to MER adopts the categorical conceptualization of emotions and categorizes music pieces by emotion classes. The major advantage of this approach is that it is easy to be incorporated into a text-based or metadata-based retrieval system. Similar to other music metadata, such as genres and instrumentations, emotion labels provide an atomic description of music that allows users to retrieve music through a few keywords. Many works have followed this direction and trained classifiers that predict the emotion class that best represent the affective content of a music signal [Dunker et al. 2008; Hu et al. 2008; Lu et al. 2006].

The major drawback of the categorical approach to MER is that the small number of primary emotion classes is too small in comparison with the richness of music emotion perceived by humans. Using a finer granularity, on other hand, does not necessarily solve the whole problem, because the language for categorizing emotion is inherently ambiguous and varies from person to person [Juslin and Laukka 2004]. Moreover, using a large number of emotion classes could overwhelm the subjects, so it is also not considered practical for psychological studies [Sloboda and Juslin 2001]. For example, for the affective terms *calm / peaceful*, *carefree*, *laid-back / mellow*, *relaxed*, and *soft*, we cannot simply quantify their similarity as zero just because they are different words or as one because they are synonyms [Shao et al. 2008]. This *ambiguity* and *granularity* issue of emotion description gives rise to the proposal of the dimensional approach to MER, as described in Section 5.

Table II. Comparison of Selected Work on MER

Approach	# emotion	# song	Genre	# subject per song
[Feng et al. 2003]	4	223	pop	N/A
[Li and Ogihara 2003]	13	499	pop	1
[Li and Ogihara 2004]	3	235	jazz	2
[Wang et al. 2004]	6	N/A	classical	20
[Wieczorkowska 2004]	13	303	pop	1
[Leman et al. 2005]	15	60	pop	40
[Tolos et al. 2005]	3	30	pop	10
[Wieczorkowska et al. 2006]	13	875	pop	1
[Yang et al. 2006]	4	195	pop	>10
[Lu et al. 2006]	4	250	classical	3
[Wu and Jeng 2006]	4	75	pop	60
[Skowronek et al. 2007]	12	1059	pop	6
[Hu et al. 2008]	5	1250	pop	<8
[Laurier et al. 2008]	4	1000	pop	from Last.fm
[Wu and Jeng 2008]	8	1200	pop	28.2
[Trohidis et al. 2008]	6	593	pop	3
[Lin et al. 2009]	12	1535	pop	from AMG
[Han et al. 2009]	11	165	pop	from AMG
[Hu et al. 2009]	18	4578	pop	from Last.fm
[Korhonen et al. 2006]	2DES	6	classical	35
[MacDorman and Ho 2007]	2DES	100	pop	85
[Yang et al. 2007, 2009]	2DES	60	pop	40
[Yang et al. 2008]	2DES	195	pop	>10
[Schmidt and Kim 2009]	2DES	120	pop	>20
[Eerola et al. 2009]	3DES	110	soundtrack	116
[Yang and Chen 2011b]	2DES	1240	pop	4.3

Next we review the methods of data preparation, subjective annotation, and model training of existing categorical MER works.

4.1. Data Collection

For lack of a common database, most existing works compile their own database [Skowronek et al. 2007; Yang and Lee 2004]. Because manual annotation is labor intensive, the size of the database of early works is usually less than 1,000. To make the database as general as possible, it is favorable to have a larger database that covers all sorts of music types, genres, or even songs of different languages.

There are many factors that impede the construction of a common database. First, there is still no consensus on which emotion model or how many emotion categories should be used. As Table II shows, the emotion taxonomy of existing work consist of three classes [Tolos et al. 2005], four classes [Feng et al. 2003], six classes [Wang et al. 2004], eight classes [Wu and Jeng 2007], and 13 classes [Li and Ogihara 2003], to name a few. Some taxonomy is based on the basic emotions proposed by psychologists, while some is derived from clustering affective terms or tags (e.g., Hu and Downie [2007] and Laurier et al. [2009]). Making comparisons with previous works which use different emotion categories and different datasets is virtually impossible. Second, due

Table III. Emotion Taxonomy Adopted in MIREX [Hu and Downie 2007]

Cluster	Description
1	passionate, rousing, confident, boisterous, rowdy
2	rollicking, cheerful, fun, sweet, amiable/good-natured
3	literate, poignant, wistful, bittersweet, autumnal, brooding
4	humorous, silly, campy, quirky, whimsical, witty, wry
5	aggressive, fiery, tense/anxious, intense, volatile, visceral

to copyright issues, the audio files cannot be distributed as freely as text documents or images [Goto et al. 2003; McKay et al. 2006]. Although the emotion annotations can be made publicly available, this is not the case for the audio files. The audio files are needed if a researcher wants to extract new music features that may be relevant to emotion perception.

In response to this need, the annual MIREX (music information retrieval evaluation exchange) Audio Mood Classification (AMC) task has been held since 2007, aiming at promoting MER research and providing benchmark comparisons [Hu et al. 2008].[5] The audio files are available to participants of the task who have agreed not to distribute the files for commercial purpose in order to get rid of the copyright issues. Being the only benchmark in the field of MER so far, this contest draws many participants every year. For example, six teams participated in AMC 2007 and 16 teams participated in AMC 2010. However, MIREX uses an emotion taxonomy that consists of five emotion clusters (see Table III) [Hu and Downie 2007][6], which have not been frequently used in existing MER works (cf. Table II). A more popular emotion taxonomy is to categorize emotions into four emotion classes, *happy*, *angry*, *sad*, and *relaxed*, partly because they are related to basic emotions studied in psychological theories and partly because they cover the four quadrants of the two-dimensional valence-arousal plane [Laurier et al. 2008]. Moreover, it has been pointed out that there is a semantic overlap (ambiguity) between clusters 2 and 4, and an acoustic overlap between clusters 1 and 5 [Laurier and Herrera 2007]. The issue on emotion taxonomy seems to remain open.

4.2. Data Preprocessing

To compare the music samples fairly, music pieces are normally converted to a standard format (e.g., 22,050 Hz sampling frequency, 16-bits precision, and mono channel). Moreover, since complete music pieces can contain sections with different emotions, a 20–30 second segment that is representative of the whole song is often selected to reduce the emotion variation within the segment and to lessen the burden of emotion annotation on the subjects [Hu et al. 2008; Lu et al. 2006]. This can be done by manually selecting the most representative part [Leman et al. 2005; Skowronek et al. 2006; Yang et al. 2008], by conducting music structure analysis to extract the chorus section [Cheng et al. 2009; Maddage et al. 2004], or simply by selecting the middle 30-second [Hu et al. 2008] or the 30-second segment starting from the 30th second of a song [Scaringella et al. 2006; Yang et al. 2007]. Few studies, if any, have been conducted to investigate the influence of music segmentation on emotion recognition.

[5] http://music-ir.org/mirexwiki

[6] The five-emotion taxonomy is determined via a statistical examination of the emotion tags obtained from AMG: http://www.allmusic.com/. See Hu and Downie [2007].

Regarding the length of the music segment, a good remark can be found in MacDorman and Ho [2007].

> In principle, we would like the segment to be as short as possible so that our analysis of the song's dynamics can likewise be as fine grained as possible. The expression of a shorter segment will also tend to be more homogeneous, resulting in higher consistency in an individual listener's ratings. Unfortunately, if the segment is too short, the listener cannot hear enough of it to make an accurate determination of its emotional content. In addition, ratings of very short segments lack ecological validity because the segment is stripped of its surrounding context.

Our literature survey reveals that using a 30-second segment seems to be common, perhaps because that corresponds to the typical length of a chorus section of popular music. As for classical music, Xiao et al. have empirically studied which length of music segments best presents the stable mood states of classical music and found that the use of a six-second or eight-second segment seems to be a good idea [Xiao et al. 2008].

4.3. Subjective Annotation

Because emotion is a subjective matter, collection of the ground truth data should be conducted carefully. Existing annotation methods can be grouped into two categories: expert-based or subject-based. The *expert-based* method employs only a few musical experts (often less than five) to annotate emotion (e.g., Li and Ogihara [2003]; Lu et al. [2006]; Trohidis et al. [2008]; Wieczorkowska [2004]). Music pieces whose emotions cannot gain consensus among experts are often simply abandoned. The *subject-based* method conducts a subjective test and employs a large number of untrained subjects to annotate emotion. The ground truth is often set by averaging the opinions of all subjects. Typically a song is annotated by more than ten subjects [Korhonen et al. 2006; MacDorman and Ho 2007; Wu and Jeng 2006; Yang et al. 2006].

As the annotation process can be very time consuming and labor costly, one needs to pay attention to the experiment design to reduce the human fatigue problem. Some common practices include:

- reducing the length of the music pieces [Skowronek et al. 2006; Yang et al. 2008];
- providing synonyms to reduce the ambiguity of the affective terms [Skowronek et al. 2006];
- using exemplar songs to better articulate what each emotion class means [Hu et al. 2008]. These exemplar songs are often selected as the songs whose emotion assignments are unanimously judged by a number of people.
- allowing the user to skip a song when none of the candidate emotion classes is appropriate to describe the affective content of the song [Hu et al. 2008];
- designing a user-friendly annotation interface [Yang et al. 2007].

Moreover, to enhance the reliability of the emotion annotations, the subjective annotation is rarely longer than an hour. The number of songs a subject is asked to annotate is accordingly limited. For example, in Yang et al. [2008] a subject is asked to annotate 15 songs.

Some lectures may also be needed to ensure the quality of the annotations, as MER may be new to the subjects. For example, in Yang et al. [2007] instructions regarding the purpose of MER, the meaning of valence and arousal, and the difference between perceived emotion and felt emotion are given to the subjects before annotation. In our empirical experience we also found that subjects are prone to misunderstand positive/negative valence as preferred/not preferred. A clear set of instructions is indeed important.

Because the perception of music emotion is multi-dimensional, the following questions may also deserve attention: Should we ask subjects to deliberatively ignore the lyrics? Should we use songs of a foreign language to the subjects to eliminate the influence of lyrics? Should we ask subjects to annotate songs they are familiar, or unfamiliar with? There seems to be no consensus on these issues so far.

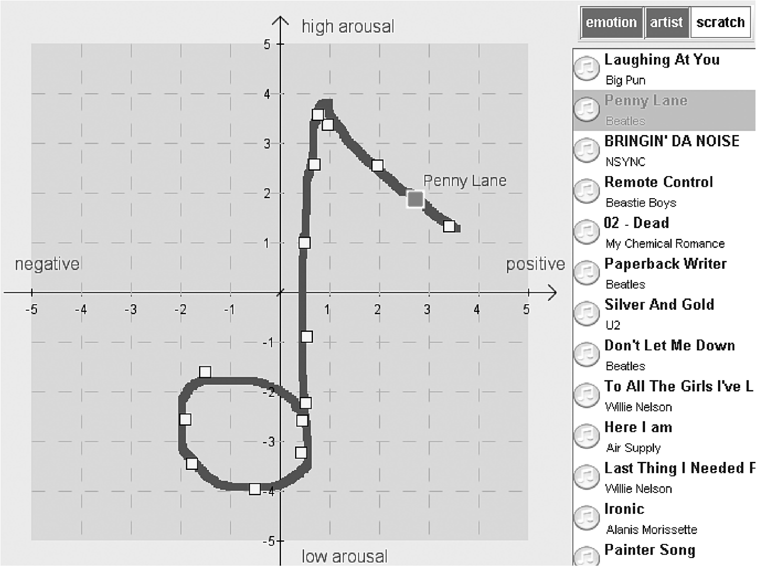
To mitigate the difficulty of subjective annotation, a recent trend is to obtain emotion tags directly from music websites such as AMG and Last.fm. Typically, this can be done by a simple script-based URL lookup. The advantage of this approach is that it is easy to obtain the annotation of a great number of songs (e.g., Bischoff et al. [2009]; Hu et al. [2009]; Laurier et al. [2009]; Lin et al. [2009]). However, the weakness is that the quality of such annotations is relatively lower than those collected through subjective annotation. For example, in AMG, the emotion labels are applied to artists and albums, not songs. In Last.fm, the tags may be incorrect because they are typically assigned by online users for their own personal use. An extensive study on social tagging of music pieces can be found in Lamere [2008].

The other trend is to harness so-called *human computation* to turn annotation into an entertaining task [Morton et al. 2010]. More specifically, the idea is to make users contribute emotion annotations as a by-product of playing Web-based games [Law et al. 2007; Mandel and Ellis 2007]. Such games are often designed as a collaborative game; that is, multiple users are playing the game at the same time to compete against one another. This practice could usually ensure the quality of the annotations. A famous example of such an online, multiplayer game is Turnbull et al.'s *Listen Game* [2007]. When playing the game, a player sees a list of semantically related words (e.g., instruments, emotions, usages, genres) and is asked to pick both the best and worst word to describe a song. Each player's score is determined by the amount of agreement between the player's choices and the choices of all other players and shown to each user immediately. Such games require little administration effort and typically obtain high-quality (and free) annotations.

4.4. Model Training

After obtaining the ground truth labeling and the musical features, the next step is to train a machine learning model to learn the relationship between emotion labels and music features. Music emotion classification is often carried out by existing classification algorithms, such as neural network, *k*-nearest neighbor (*k*-NN), maximum likelihood, decision tree, or support vector machine (SVM) [Cortes and Vapnik 1995]. For example, the best performing systems of MIREX AMC 2007–2009 are based on support vector machines [Chang and Lin 2001; Hu et al. 2008; Tzanetakis 2007], Gaussian mixture models [Peeters 2008], and the combination of SVM and GMM [Campbell et al. 2006; Cao and Li 2009], respectively. Many existing MER works also report that SVM tends to give the superior performance [Han et al. 2009; Laurier and Herrera 2007].

Considering the fact that a song may express more than one emotion, multilabel classification algorithms, such as multilabel SVM [Lewis et al. 2004] and ML*k*NN [Zhang and Zhou 2007], have been applied to assign multiple emotion labels to a song [Li and Ogihara 2003; Trohidis et al. 2008; Wieczorkowska et al. 2006]. Motivated by the fact that emotion perception is influenced by personal factors such as cultural background, generation, sex, and personality [Huron 2006], Yang et al. proposed a fuzzy approach that measures the strength of each emotion in association with the song under classification and provides a more objective measurement of emotion [Yang et al.



Laughing At You
Big Pun
Penny Lane
Beatles
BRINGIN' DA NOISE
NSYNC
Remote Control
Beastie Boys
02 - Dead
My Chemical Romance
Paperback Writer
Beatles
Silver And Gold
U2
Don't Let Me Down
Beatles
To All The Girls I've L
Willie Nelson
Here I am
Air Supply
Last Thing I Needed F
Willie Nelson
Ironic
Alanis Morissette
Painter Song

Fig. 3. In Yang et al. [2008], each song is represented as a point in the 2DES. A user can retrieve music of a certain emotion by simply specifying a point or a path in the plane.

2006]. Wu et al. also proposed a probabilistic approach that predicts the probability distribution of a music piece over the Hevner's eight emotion classes [Wu and Jeng 2008], using the probabilistic estimate of SVM [Platt 1999].

5. DIMENSIONAL MUSIC EMOTION RECOGNITION

The circumplex model is first adopted by MER researchers to track the emotion variation of a classical song (i.e., MEVD). The idea of representing the overall emotion of a popular song as a point in the emotion plane for music retrieval is studied in Yang et al. [2008] and MacDorman et al. [2007], under the assumption that the dominant emotion of a popular song undergoes less change than a classical song (this assumption is also made in the MIREX AMC contest [Hu et al. 2008]). The authors formulated MER as a regression problem [Sen and Srivastava 1990] and trained two independent regression models (regressors) to predict the valence and arousal (VA) values of a music piece. Associated with the VA values, each music piece is visualized as a point in the emotion plane, and the similarity between music samples can be estimated by computing the Euclidean distance in the emotion plane. This regression approach has a sound theoretical foundation and exhibits promising prediction accuracy. See Table II for a comparison of selected works on dimensional MER.

The attractions of this approach are the two-dimensional user interface and the associated emotion-based retrieval methods that can be created for mobile devices that have small display areas. For example, a user can specify a point in the plane to retrieve songs of a certain emotion or draw a trajectory to create a playlist of songs with various emotions corresponding to points on the trajectory (see Figure 3 for an illustration) [Yang et al. 2008]. In addition, because the emotion plane implicitly offers an infinite number of emotion descriptions, the granularity and ambiguity issue associated with emotion classes is alleviated.

Note that the use of valence and arousal as the two emotion dimensions, though largely inspired from the psychology domain, has also been empirically validated by MIR researchers. In Levy and Sandler [2007] and Laurier et al. [2009], researchers have investigated the semantic emotion space spanned from social music tags and found that the derived semantic space conforms to the valence-arousal emotion space. In another study [Leman et al. 2005], by applying factor analysis on emotion annotations of 15 bipolar affective terms, it was also found that the underlying 3D space is characterized by valence, activity, and interest (exciting–boring), which is fairly close to the valence-arousal-potency model.

Next we review the methods of subjective annotation, model training, and result visualization of existing dimensional MER works. The data preparation part of dimensional MER is skipped because it is similar to that of categorical MER.

5.1. Model Training

Dimensional MER is usually formulated as a regression problem [Sen and Srivastava 1990] by viewing the emotion values (i.e., VA values) as real values in [–1, 1]. Then a regression model can be trained to predict the emotion values. More specifically, given N inputs $(\mathbf{x}_i, y_i)$, $1 \leq i \leq N$, where $\mathbf{x}_i$ is a feature vector of the ith input sample and y_i is the real value to be predicted, a regression model (regressor) $f(\cdot)$ is created by minimizing the mismatch (i.e., mean squared difference) between the predicted and the ground truth values. Many good regression algorithms, such as support vector regression (SVR) [Schölkopf et al. 2000], Gaussian process regression [Rasmussen and Williams 2006], or AdaBoost.RT [Solomatine and Shrestha 2004] are readily available. Most existing works train two regressors for valence and arousal independently [MacDorman et al. 2007; Yang et al. 2008].

A standard metric for evaluating regressors [Sen and Srivastava 1990] is the R^2 statistics, or the coefficient of determination. It measures the proportion of the underlying data variation that is explained by the fitted regression model [Montgomery et al. 1998],

$$R^2(\mathbf{y}, f(\mathbf{x})) = \frac{\operatorname{cov}(\mathbf{y}, f(\mathbf{x}))^2}{\operatorname{var}(\mathbf{y})\operatorname{var}(f(\mathbf{x}))}. \tag{1}$$

$R^2 = 1$ means the model perfectly fits the data, while $R^2 = 0$ indicates no linear relationship between the ground truth and the estimate. It is generally observed that valence recognition is much more challenging than arousal recognition [Fornari and Eerola 2008; Korhonen et al. 2006; Lu et al. 2006]. For example, the R^2 reported in Yang et al. [2008] is 0.28 for valence and 0.58 for arousal, while that reported in Yang et al. [2007] is 0.17 for valence and 0.80 for arousal. This is partly because valence perception is more subjective and partly because it is computationally more difficult to reliably extract features relevant to valence perception, such as musical mode and articulation [Lu et al. 2006; Repp 1998].

Next we briefly describe SVR for its superior performance for dimensional MER [Huq et al. 2010; Schmidt and Kim 2009; Schmidt et al. 2010; Yang et al. 2008]. Since the nineties, support vector machines (SVMs) have been widely used in different classification and regression tasks [Cortes and Vapnik 1995]. SVM nonlinearly maps an input feature vector $\mathbf{x}$ to a higher dimensional feature space $\phi(\mathbf{x})$ by the so-called "kernel trick" and learns a nonlinear function by a linear learning machine in the kernel-induced feature space, where data are more separable [Cortes and Vapnik 1995]. For classification, we look for the optimal separating hyperplane that has the largest distance to the nearest training data points of any class. For regression, we look for a function $f(\mathbf{x}_s) = \mathbf{m}^\top \phi(\mathbf{x}_s) + b$ that has at most ε deviation from the ground truth y_s for all the training data and, meanwhile, is as flat as possible (i.e., $\mathbf{m}^\top\mathbf{m}$ is small) [Schölkopf et al. 2000]. In other words, we do not care about errors as long as they are less than ε but will not accept any deviation larger than this. Moreover, under the soft margin principle [Boyd and Vandenberghe 2004], we introduce slack variables ξ_s and ξ_s^* to allow the error to be greater than ε. Consequently, we have the following optimization problem,

$$\begin{aligned} \underset{\mathbf{m},b,\xi,\xi^*,\varepsilon}{\arg\min} \quad & \frac{1}{2}\mathbf{m}^\top\mathbf{m} + C(\nu\varepsilon + \frac{1}{N}\sum_{s=1}^{N}(\xi_s + \xi_s^*)), \\ \text{subject to} \quad & (\mathbf{m}^\top\phi(\mathbf{x}_s) + b) - y_s \leq \varepsilon + \xi_s, \\ & y_s - (\mathbf{m}^\top\phi(\mathbf{x}_s) + b) \leq \varepsilon + \xi_s^*, \\ & \xi_s, \xi_s^* \geq 0, s = 1, \ldots, N, \ \varepsilon \geq 0, \end{aligned} \tag{2}$$

where C controls the trade-off between the flatness of $f(\cdot)$ and the amount up to which deviations larger than ε are tolerated and $\nu \in [0, 1]$ controls the number of support vectors (the points lying on the boundaries). A common kernel function is the radial basis function (RBF): $K(\mathbf{x}_p, \mathbf{x}_q) \equiv \phi(\mathbf{x}_p)^\top\phi(\mathbf{x}_q) = \exp(-\gamma\,||\mathbf{x}_p - \mathbf{x}_q||^2)$, where γ is a scale parameter. Typically, the parameters of SVR are determined empirically by a grid search. The preceding quadratic optimization problem can be efficiently solved by known techniques [Boyd and Vandenberghe 2004]. A popular implementation of SVR is the LIBMSVM library [Chang and Lin 2001].

5.2. Subjective Annotation

Unlike its categorical counterpart, dimensional MER requires subjects to annotate the numerical VA values. This can be done using either a standard rating scale [MacDorman and Ho 2007; Yang et al. 2008] or a graphic rating scale [Cowie et al. 2000; Schubert 1999; Yang et al. 2007]. For example, in Yang et al. [2008], subjects were asked to rate the VA values from –1.0 to 1.0 in eleven ordinal levels. In MacDorman and Ho [2007], a seven-point scale was used, implemented as a radio button that consisted of a row of seven circles with an opposing semantic differential item appearing at each end. In Yang et al. [2007], subjects were asked to rate the VA values using a graphic interface called “AnnoEmo.” The VA values are annotated by clicking on the emotion plane displayed by computer. A rectangle is formed on the specified point so that the subject can directly compare the annotations of different music pieces. The subject can click on the rectangle to listen to the piece again or drag and drop the rectangle to modify the annotations. Regardless of the annotation method employed, each music piece is often annotated by multiple subjects, and the ground truth is set to the average rating as emotion perception is subjective. Algorithms such as the one described in Grubbs [1969] may be employed to remove outliers (annotations that are significantly different from others).

5.3. Result Visualization

Given the trained regressors, the VA values of a song are automatically predicted without further manual labeling. Associated with VA values, each music piece is visualized as a point in the emotion plane. Many novel retrieval methods can be realized in the emotion plane, making music information access much easier and more effective. For example, one can easily retrieve music pieces of a certain emotion without knowing the titles or browse personal collections in the emotion plane on mobile devices. One can also couple emotion-based retrieval with traditional keyword- or artist-based ones to retrieve songs similar (in the sense of perceived emotion) to a favorite piece or to select the songs of an artist according to emotion. We can also generate a playlist by drawing a free trajectory representing a sequence of emotions in the emotion plane. As the trajectory goes from one quadrant to another, the emotions of the songs in the playlist would vary accordingly, as shown in Figure 3. See Yang et al. [2008] for a technical demonstration.

Whether the emotions should be modeled as categories or continua has been a long debate in psychology [Collier 2007; Ekman 1992; Hevner 1935; Thayer 1989]. From an engineering perspective, the categorical approach and the dimensional approach offer different advantages that are complementary to each other. We can imagine a mobile device that employs both approaches to facilitate music retrieval.

6. MUSIC EMOTION VARIATION DETECTION

An important aspect of music that has thus far been neglected in this article is its temporal dynamics. Most research has focused on musical excerpts that are homogeneous with respect to emotional expression. However, as many styles of music (in particular classical music) also express or evoke different emotions as time unfolds, it is important to investigate the time-varying relationship between music and emotion. Techniques for continuous recording, sometimes combined with nonverbal responses, have been used to study emotion perception as a dynamic process since the seminal work of Nielsen [1986]. According to Schubert [1999], "Continuous response tools measure self-reported affective responses during the listening process and help researchers to better understand the moment-to-moment fluctuations in responses. The approach means that the music does not need to be broken into small portions to enable lucid response, nor does the listener have to compress their response by giving an overall impression at the end of the excerpt. A more realistic listening experience is possible. This realism [Hargreaves 1986] contributes to the ecological validity of experimental design at the expense of experimental control."

Because categorical responses require the subjects to choose an emotion term constantly, the dimensional approach to emotion conceptualization is found more useful for capturing the continuous changes of emotional expression [Gabrielsson 2002]. Usually subjects are asked to rate the VA values (typically by clicking a point in the emotion plane) every one second in response to the stimulus [Cowie et al. 2000; Lang 1995; Russell et al. 1989]. Psychologists then analyze the continuous recording to investigate the relationship between music and emotion [Eerola et al. 2002; Schubert 2001; Toiviainen and Krumhansl 2003].

Attempts have been made to automatically detect the music emotion variation, or music emotion variation detection (MEVD) [Schmidt et al. 2010]. Two approaches can be found in the literature. The first approach, such as the time series analysis method [Schubert 1999] and the system identification method [Korhonen et al. 2006], exploits the temporal information among the music segments while computing the VA values. For example, in Korhonen et al. [2006], system identification techniques [Ljung 1999] are utilized to model music emotion as a function of a number of musical features. The ground truth data are collected for every second of the music pieces, so the music pieces are also segmented every second for feature extraction. The dataset consists of six Western classical music pieces of various emotions. Results demonstrate that system identification provides a means to the generalization of the affective content of Western classical music.

The second approach neglects the temporal information underlying the music signals and makes a prediction for each music segment independently. For example, in Yang et al. [2006] a sliding window of ten seconds and 1/3 overlap is used to segment a music piece, whereas in Lu et al. [2006], the potential emotion change boundaries are identified first and then utilized to segment the music piece. The emotion of each segment is then predicted independently.

Besides music, the dimensional approach has been adopted to track the emotion variation within video sequences [Arifin and Cheng 2008; Dietz and Lang 1999;

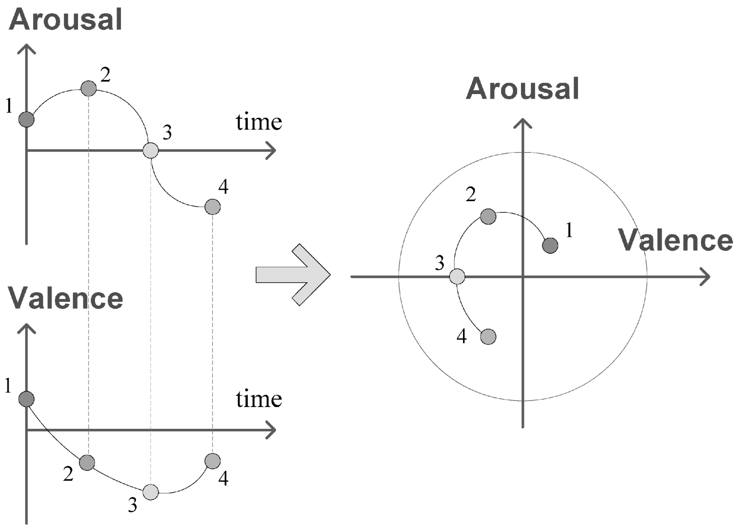


Fig. 4. With emotion variation detection, we can combine the valence and arousal curves (left) to form the affective curve (right), which represents the dynamic changes of the affective content of a video sequence or a music piece.

Hanjalic and Xu 2005; Wang and Cheong 2006] and speech signals [Giannakopoulos et al. 2009]. For example, *VA modeling* is proposed in Hanjalic and Xu [2005] to detect the emotion variation in movie sequences. The VA values are computed by the weighted sums of some component functions that are computed along the timeline. The component functions used for arousal are the motion vectors between consecutive video frames, the changes in shot lengths, and the energy of sound, whereas the component function used for valence is the sound pitch. The resulting valence and arousal curves are then combined to form an *affective curve* which makes it easy to trace the emotion variation of the video sequence and to identify the segments with high affective content [Hanjalic and Xu 2005]. See Figure 4 for an illustration. This work is later extended by Zhang et al. [2008, 2009], who modeled the emotion of music videos (MVs) and movies using 22 audio features (intensity, timbre, rhythm) and five visual features (motion intensity, shot switch rate, frame brightness, frame saturation, and color energy) and proposed an interface for affective visualization on time axis. They took the approach described in the previous paragraph and predicted the VA values for each short-time video clip independently. In their system, a movie sequence is segmented every 14 seconds, with an overlap of two seconds between consecutive segments.

Though both MEVD and dimensional MER view emotions from the dimensional perspective, they are different in the ways the computational problem is formulated and approached. MEVD computes the VA values of each short-time segment and represents a song as a series of VA values (points), whereas the dimensional MER computes the VA values of a representative segment (often 30 seconds) of the song and represents the song as a single point. However, it should be noted that a dimensional MER system can also be applied to MEVD if we neglect the temporal information and compute the VA values of each segment independently.

7. CHALLENGES

As MER is still in its infancy, there are many open issues. Some major issues and proposed solutions are discussed in this section.

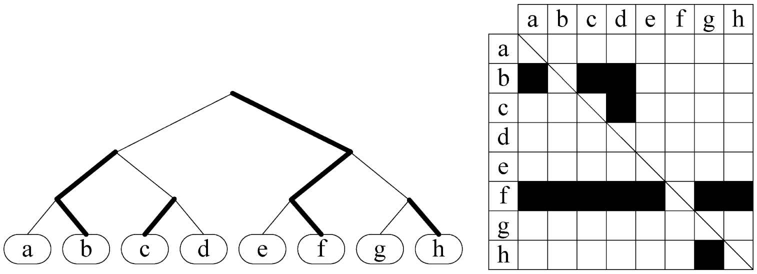


Fig. 5. The ranking-based emotion annotation method proposed in Yang and Chen [2011b], which groups eight randomly chosen music pieces in a tournament of seven matches. Users are asked to rank (by making pairwise comparisons) instead of rate music emotion with respect to emotion dimension, such as valence or arousal.

7.1. Difficulty of Emotion Annotation

To collect the ground truth needed for training an automatic model, a subjective test is typically conducted to invite human subjects to annotate the emotion of music pieces. Since the MER system is expected to be used in the everyday context, the emotion annotation should better be carried out by common people. The psychology literature suggests that each stimulus be annotated by more than 30 annotators for the annotation to be reliable [Cohen and Swerdlik 2002]. This requires a great many annotations to develop a large-scale dataset.

The difficulty of collecting emotion labels for training a categorical MER system has recently been alleviated with the surge of online tagging websites, such as AMG and Last.fm, as reviewed in Section 4.3. The emotion annotation process of dimensional MER, however, requires numerical emotion ratings that are not readily available from the online repository. Moreover, it has been found that rating emotion in a continuum usually imposes a heavy cognitive load on the subjects [Yang and Lee 2004]. It is also difficult to ensure a consistent rating scale between different subjects and within the same subject [Ovadia 2004]. As a result, the quality of the ground truth varies, which in turn degrades the accuracy of MER.

To address this issue, ranking-based emotion annotation is proposed [Yang and Chen 2011b]. A subject is asked to compare the affective content of two songs and determine, for example, which song has a higher arousal value, instead of the exact emotion values. Since it is a lengthy process to determine the straight order of n music pieces (requiring $n(n-1)/2$ comparisons), a music emotion tournament scheme is proposed to reduce the burden on subjects. As Figure 5 shows, the n music pieces can be grouped into $n-1$ matches, which form a hierarchy of $\log_2 n$ levels. A subject compares the emotion values of two music pieces in each match and decides whose emotion value is larger. The rankings of music emotion are then converted to numerical values by a greedy algorithm [Cohen et al. 1999]. Empirical evaluation shows that this scheme relieves the burden of emotion annotation on the subjects and enhances the quality of the ground truth. It is also possible to use an online game to harness the so-called human computation and make the annotation process more engaging [Kim et al. 2008].

7.2. Subjectivity of Emotional Perception

Music perception is intrinsically subjective and is under the influence of many factors, such as cultural background, age, gender, personality, training, and so forth [Huron 2006]. The interactions between music and listener may also involve the listener's familiarity with the music and his/her musical preferences [Jargreaves and North 1997].

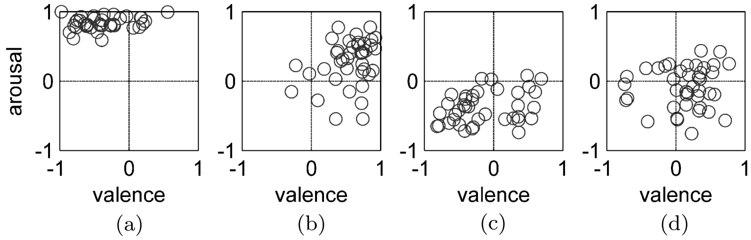


Fig. 6. Emotion annotations in the 2DES for four songs: (a) *Smells Like Teen Spirit* by Nirvana, (b) *A Whole New World* by Peabo Bryson and Regina Belle, (c) *The Rose* by Janis Joplin, (d) *Tell Laura I Love Her* by Ritchie Valens. Each circle corresponds to a subject's annotation of the song [Yang et al. 2007]. It can be observed that emotion perception is indeed subjective.

Because of this subjectivity issue, it is difficult to gain common consensus on which affective term best characterizes the affective content of a music piece. Therefore, typical categorical approaches that simply assign one emotion class to each music piece in a deterministic manner does not perform well in practice. The dimensional approach to MER also faces the subjectivity issue that people can response differently to the same song. For example, each circle in Figure 6 corresponds to a subject's annotation of the perceived emotion of a song [Yang et al. 2007]. We see that people often have different emotion perceptions, and the annotations of a music piece are sometimes fairly sparse.

Despite that the subjectivity nature of emotion perception is well recognized, little effort has been made to take the subjectivity into account. Most works either assume a common consensus can be achieved (particularly for classic music) [Wang et al. 2004], discard those songs upon which a common consensus cannot be achieved [Lu et al. 2006], or simply leave this as future work [Li and Ogihara 2003].

To address this issue, a fuzzy approach is proposed [Yang et al. 2006] to measure the strength of each emotion class in association with the song under classification. By assigning each music piece a soft label that indicates how likely a certain emotion would be perceived when listening to the piece, the prediction result becomes less deterministic. For example, a song could be 70% likely to be relaxed and 30% likely to be sad.

A different methodology that addresses the subjectivity issue is needed for the dimensional approach to MER, as emotions are not conceptualized as discrete classes but numerical values (e.g., VA values). In Yang et al. [2007], two personalization methods are proposed; the first trains a *personalized* MER systems for each individual specifically, whereas the second groups users according to some personal factors (e.g., gender, music experience, and personality) and then trains *group-wise* MER systems for each user group. Another two-stage personalization scheme is also studied [Yang et al. 2009]. In this approach, two models are created: one for predicting the general perception of a music piece, the other for predicting the difference (coined as *perceptual residual*) between general perception and the personal perception of a user. This simple method is effective because the music content and the individuality of the user are treated separately. These methods show that personalization is feasible, but they lack a solid computational framework.

Motivated by the observation that the perceived emotions of a song in fact constitute an *emotion distribution* in the emotion plane (cf. Figure 6), Yang and Chen [2011a] proposed a computational model to model the perceived emotions of a song as a probabilistic distribution in the emotion plane and to compute the probability of perceived emotion of a song—somewhat similar to the idea of soft labeling of the fuzzy approach [Yang et al. 2006]. More specifically, the computational model aims at predicting its *emotion mass* at discrete samples in the 2DES, with the values summed to one. Here,

Table IV. Top Three Performance of Audio Mood Classification (AMC) of MIREX (2007–2010)

Contest	Top Three Accuracy
AMC 2007	65.67%, 65.50%, 63.67%
AMC 2008	63.67%, 56.00%, 55.00%
AMC 2009	61.50%, 60.50%, 59.67%
AMC 2010	64.17%, 63.83%, 63.17%

Note: Retrieved from the website of MIREX.

the term emotion mass refers to the probability of the perceived emotion of a song being a specific point (discrete sample) in the 2DES. This computational framework provides a new basis for personalized emotion-based retrieval. An emotion distribution can be regarded as a collection of users' perceived emotions of a song, and the perceived emotion of a specific user can be regarded as a sample of the distribution.

7.3. Semantic Gap Between Low-Level Music Feature and High-Level Human Perception

The viability of an MER system largely lies in the accuracy of emotion recognition. However, due to the so-called semantic gap between the object feature level and the human cognitive level of emotion perception, it is difficult to accurately predict the emotion labels or values [Lu et al. 2006; Tolos et al. 2005; Yang et al. 2008]. What intrinsic element of music, if any, causes a listener to create a specific emotional perception is still far from being well-understood. Consequently, the performance of conventional methods that exploit only the low-level audio features seems to have reached a limit. For example, Table IV shows the top three performances (in terms of raw mean classification accuracy) of the MIREX AMC contest from 2007 to 2010. It can be observed that despite various low-level audio features and their combinations having been used [Hu et al. 2008], the classification accuracy seems to be bounded by 66%.[7]

Available data for MER are not limited to the raw audio signal. Complementary to music signal, lyrics are semantically rich and have profound impacts on human perception of music [Ali and Peynircioğu 2006]. It is often easy for us to tell from the lyrics whether a song expresses sadness or happiness. Incorporating lyrics into MER is feasible because most popular songs sold in the market come with lyrics [Fornäs 2006]. One can analyze lyrics using natural language processing to generate text feature descriptions of music, such as bag-of-words, part-of-speech [Sebastiani 2002], and latent semantic vectors [Hofmann 1999]. Several attempts have been made to augment the MER system with features extracted from the lyrics [Hu et al. 2009; Laurier et al. 2008; Lu et al. 2010; Meyers 2007; Yang and Lee 2004; Yang et al. 2008; van Zaanen and Kanters 2010]. It is often reported that the use of lyrics improves the accuracy of valence recognition.

Besides using lyrics, the use of other mid-level or high-level audio features, such as chord progression and genre metadata, has also been studied. Chord progression is automatically detected by a chord recognition system [Cheng et al. 2008], while genre metadata are obtained by applying an automatic genre classifier or by crawling the Internet [Lin et al. 2009]. Empirical evaluations show that the incorporation of these features improves the accuracy of emotion recognition. For example, Schuller et al. [2008, 2010] incorporated genre, ballroom dance style, chord progression, and lyrics in

[7] Note the performance is also influenced by the ambiguity issue inherent to the five-class taxonomy and the subjectivity issue inherent to emotion perception.

their MER system and found that many of them contribute positively to the prediction accuracy.

8. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Despite a great deal of effort, MER is still a fairly new research area with a lot of unknown or unsolved problems. For example, the accuracy of valence recognition of existing systems is still not satisfactory [Yang and Chen 2011b], and the subjectivity issue of emotion perception has not been resolved. In this section, we describe some possible future research directions. Like Klaus R. Scherer concluded in the foreword of *Music and Emotion: Theory and Research* [Juslin and Sloboda 2001], we hope this review article will inspire more multidisciplinary-minded researchers to study "a phenomenon that has intrigued mankind since the dawn of time."[8]

8.1. Exploiting Vocal Timbre for MER

Another source of information that is not yet fully exploited in the literature is the singing voice of the music. Typically, a pop music consists of the accompany music, lyrics, and the singing voice. The timbral of the singing voice, such as *aggressive*, *breathy*, *gravelly*, *high-pitched*, or *rapping* [Turnbull et al. 2008], is often directly related to our emotion perception. For example, a song with screaming and roaring voices usually conveys an angry emotion, whereas a song with sweet voices tends to be of positive emotion. Therefore, vocal timbre is important for valence perception and should be incorporated to MER. Speech features that have been shown useful for speech emotion recognition [Fernandez and Picard 2005; Giannakopoulos et al. 2009; Picard et al. 2001; Schuller et al. 2009; Ververidis and Kotropoulos 2006] and automatic singer identification [Fujihara et al. 2005; Nwe and Li 2007a; Shen et al. 2006; Tsai and Wang 2006], such as $\triangle$F0 [Fujihara and Goto 2007], vibrato, harmonics, attack-delay [Nwe and Li 2007b], voice source features [Fernandez and Picard 2005], and harmonics-to-noise ratio [Schuller et al. 2009], could be considered.

A key issue of vocal timbral recognition is the suppression or reduction of the accompanying music [Goto 2004]. The simplest way might be to apply a bandpass filter that preserves only the frequency components of the singing voice. A major drawback of this approach, however, is that many instruments also have frequency responses in the singing format and thus cannot be eliminated. Therefore, advanced techniques, such as predominant-F0 estimation [Goto 2004] or melodic source separation [Lagrange et al. 2008], may be needed.

8.2. Personalized Emotion-Based Music Retrieval

In Yang and Chen [2011a], the authors only focused on the indexing part of MER; that is, predicting the emotion distribution of a song $P(\mathbf{e}|d)$ such that we can organize and represent songs in the emotion plane. This methodology can be extended to address the retrieval part; that is, when a user u clicks on a point e^{ij}, return to the user a list of songs ranked in descending order of $P(d|e^{ij}, u)$, where $e^{ij} = [v^i, a^j]^\top$, and $[v^i, a^j]^\top \in [-1, 1]^2$ are VA values. We can consider an emotion distribution $P(\mathbf{e}|d)$ as a collection of users' perceived emotions of a song and the perceived emotion of the user $P(\mathbf{e}|d, u)$ as a sample of the distribution. Though the methodology of predicting $P(\mathbf{e}|d, u)$ remains to be developed, it is interesting because by doing so, music emotion recognition and emotion-based music retrieval would be studied under a unified probabilistic framework.

[8] Interested readers may also refer to Kim et al. [2010] for another state-of-the-art review of MER.

8.3. Connections Between Dimensional and Categorical MER

As we have described in Sections 4 and 5, the categorical approach and the dimensional approach to MER offer complementary advantages. The former offers an atomic description of music that is easy to incorporate into a conventional text-based retrieval system, whereas the latter offers a simple means for a 2D user interface. It is therefore interesting to combine the two approaches to construct a more effective and user-friendly emotion-based music retrieval system. We shall give two example directions.

From the model training perspective, as the emotion labels are easier to be obtained (e.g., by crawling AMG or Last.fm), it is interesting to develop a method that utilizes the categorical emotion labels as ground truth data for dimensional MER system. This can be approached, for example, by mapping the affective terms (e.g., *peaceful*, *romantic*, *sentimental*) to points in the emotion plane (cf. Figure 2) and regarding the corresponding VA values as the ground truth of the associated music pieces. In this way, constructing a large-scale database for training and evaluating dimensional MER would be easier.

From the music retrieval perspective, because users may be unfamiliar with the essence of the valence and arousal dimensions, when representing songs as points in the emotion plane it should be beneficial to add affective terms to guide the users. The users can choose the affective terms to be displayed. We could also allow the users to decide the position of the affective terms and utilize such information to personalize the MER system.

8.4. Considering the Situational Factors of Emotion Perception

According to psychological studies, our emotion response to music is dependent on an interplay between musical, personal, and situational factors [Gabrielsson 2002]. Indeed, under the influence of situational factors such as listening mood and listening environment, a person's emotion perception of the same song could vary a lot. For example, when we are in a sad mood, a happy song may be not so happy to us. Therefore, it would be great if the MER system could detect the listening mood (e.g., via prosodic cues, body movements, or physiological signals [Jaimes et al. 2006; Lee and Narayanan 2005; Lin et al. 2008, 2009; Picard et al. 2001]) or the listening environment (e.g., via monitoring the background volume) to modify the emotion predictions. On the other hand, the MER system can also utilize the information of the listening context to actively recommend music to the listeners.

REFERENCES

ALI, S. O. AND PEYNIRCIOĞU, Z. F. 2006. Songs and emotions: Are lyrics and melodies equal partners. *Psychol. Music 34,* 4, 511–534.

ALLAMANCHE, E., HERRE, J., HELMUTH, O., FRÖBA, B., KASTEN, T., AND CREMER, M. 2001. Content-based identification of audio material using MPEG-7 low level description. In *Proceedings of the International Conference on Music on Information Retrieval*. 197–204.

ANDERSON, K. AND MCOWAN, P. W. 2006. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. Syst. Man Cyber. 36,* 1, 96–105.

ARIFIN, S. AND CHEUNG, P. Y. K. 2008. Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Trans. Multimedia 10,* 7, 1325–1341.

BENETOS, E., KOTTI, M., AND KOTROPOULOS, C. 2007. Large scale musical instrument identification. In *Proceedings of the International Conference on Music Information Retrieval*. http://www.ifs.tuwien.ac.at/mir/muscle/del/audio_tools.html#SoundDescrToolbox.

BIGAND, E., VIEILLARD, S., MADURELL, F., MAROZEAU, J., AND DACQUET, A. 2005. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition Emotion 19,* 8, 1113–1139.

Bischoff, K., Firan, C. S., Paiu, R., Nejdl, W., Laurier, C., and Sordo, M. 2009. Music mood and theme classification—a hybrid approach. In *Proceedings of the International Conference on Music Information Retrieval*. 657–662.

Boyd, S. P. and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Cabrera, D. 1999. Psysound: A computer program for psycho-acoustical analysis. In *Proceedings of the Australian Acoustic Society Conference*. 47–54. `http://psysound.wikidot.com/`.

Cai, R., Zhang, C., Wang, C., Zhang, L., and Ma, W.-Y. 2007. MusicSense: Contextual music recommendation using emotional allocation modeling. In *Proceedings of the ACM International Conference on Multimedia*. 553–556.

Camacho, A. 2007. SWIPE: A sawtooth waveform inspired pitch estimator for speech and music. Ph.D. dissertation, University of Florida.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A. 2006. Support vector machines for speaker and language recognition. *Comput. Speech Lang. 20,* 2–3, 210–229.

Cao, C. and Li, M. 2009. Thinkit's submissions for MIREX2009 audio music classification and similarity tasks. In *Proceedings of the International Conference on Music Information Retreival*.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. 2008. Content-based music information retrieval: Current directions and future challenges. *IEEE 96,* 4, 668–696.

Chang, C.-C. and Lin, C.-J. 2001. LIBSVM: A library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, C.-H., Weng, M.-F., Jeng, S.-K., and Chuang, Y.-Y. 2008. Emotion-based music visualization using photos. In *Proceedings of the International Conference on Multimedia Modeling*. 358–368.

Chen, J.-C., Chu, W.-T., Kuo, J.-H., Weng, C.-Y., and Wu, J.-L. 2006. Tiling slideshows. In *Proceedings of the ACM International Conference on Multimedia*. 25–34.

Cheng, H.-T., Yang, Y.-H., Lin, Y.-C., and Chen, H.-H. 2009. Multimodal structure segmentation and analysis of music using audio and textual information. In *Proceedings of the IEEE International Symposium on Circuits and Systems*. 1677–1680.

Cheng, H.-T., Yang, Y.-H., Lin, Y.-C., Liao, I.-B., and Chen, H.-H. 2008. Automatic chord recognition for music classification and retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo.* 1505–1508.

Cohen, R. and Swerdlik, M. 2002. *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. Mayfield Publishing Company, Mountain View, CA.

Cohen, W. W., Schapire, R. E., and Singer, Y. 1999. Learning to order things. *J. Artificial Intell.Res. 10*, 243–270.

Collier, G. 2007. Beyond valence and activity in the emotional connotations of music. *Psychol. Music 35,* 1, 110–131.

Cortes, C. and Vapnik, V. 1995. Support vector networks. *Machine Learn. 20,* 3, 273–297.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröer, M. 2000. Feeltrace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion*. 19–24.

Davis, S. and Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing 28,* 4, 357–366.

Dietz, R. and Lang, A. 1999. Affective agents: Effects of agent affect on arousal, attention, liking and learning. In *Proceedings of the International Conference on Cognitive Technology.*

Dornbush, S., Fisher, K., Mckay, K., Prikhodko, A., and Segall, Z. 2005. XPOD: A human activity and emotion aware mobile music player. In *Proceedings of the International Conference on Mobile Technology, Applications and Systems*. 1–6.

Duda, R. O., Hart, P. E., and Stork, D. G. 2000. *Pattern Classification*. John Wiley & Sons, Inc., New York.

Dunker, P., Nowak, S., Begau, A., and Lanz, C. 2008. Content-based mood classification for photos and music. In *Proceedings of the International Conference on Multimedia Information Retrieval*. 97–104.

Eerola, T., Toiviainen, P., and Krumhansl, C. L. 2002. Real-time prediction of melodies: Continuous predictability judgments and dynamic models. In *Proceedings of the International Conference on Music Perception and Cognition*. 473–476.

EEROLA, T., LARTILLOT, O., AND TOIVIAINEN, P. 2009. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of the International Conference on Music Information Retrieval*. 621–626.

EKMAN, P. 1992. An argument for basic emotions. *Cognition Emotion 6,* 3, 169–200.

FARNSWORTH, P. R. 1954. A study of the Hevner adjective list. *J. Aesthetics Art Criticism 13*, 97–103.

FENG, Y., ZHUANG, Y., AND PAN, Y. 2003. Popular music retrieval by detecting mood. In *Proceedings of the International Conference on Information Retrieval*. 375–376.

FERNANDEZ, R. AND PICARD, R. W. 2005. Classical and novel discriminant features for affect recognition from speech. In *Proceedings of the INTERSPEECH Conference*.

FORNARI, J. AND EEROLA, T. 2008. The pursuit of happiness in music: Retrieving valence with high-level musical descriptors. In *Proceedings of the Computer Music Modeling and Retrieval*.

FORNÄS, J. 2006. Songs and emotions: Are lyrics and melodies equal partners. *Psychol. Music 34,* 4, 511–534.

FUJIHARA, H. AND GOTO, M. 2007. A music information retrieval system based on singing voice timbre. In *Proceedings of the International Conference on Music Information Retrieval*.

FUJIHARA, H., KITAHARA, T., GOTO, M., KOMATANI, K., OGATA, T., AND OKUNO, H. G. 2005. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the International Conference on Music Information Retrieval*.

GABRIELSSON, A. 2002. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae* (special issue), 123–147.

GABRIELSSON, A. AND LINDSTRÖM, E. 2001. The influence of musical structure on emotional expression. In *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda Eds., Oxford University Press, Oxford, UK.

GIANNAKOPOULOS, T., PIKRAKIS, A., AND THEODORIDIS, S. 2009. A dimensional approach to emotion recognition of speech from movies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 65–68.

GÓMEZ, E. 2006. Tonal description of music audio signal. Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona.

GOTO, M. 2004. A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication 43*, 311–329.

GOTO, M., HASHIGUCHI, H., NISHIMURA, T., AND OKA, R. 2003. RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the International Conference on Music Information Retrieval*. 229–230.

GRUBBS, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics 11,* 1, 1–21.

HAN, B.-J., RHO, S., DANNENBERG, R. B., AND HWANG, E. 2009. SMERS: Music emotion recognition using support vector regression. In *Proceedings of the International Conference on Music Information Retrieval*. 651–656.

HANJALIC, A. AND XU, L.-Q. 2005. Affective video content representation and modeling. *IEEE Trans. Multimedia 7,* 1, 143–154.

HARGREAVES, D. J. 1986. *The Developmental Psychology of Music*. Cambridge University Press, Cambridge, UK.

HEVNER, K. 1935. Expression in music: A discussion of experimental studies and theories. *Psychol. Review 48,* 2, 186–204.

HOFMANN, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM International Conference on Information Retrieval*. 50–57.

HSU, D. C.-W. AND HSU, J. Y.-J. 2006. LyQ: An emotion-aware music player. In *Proceedings of the AAAI Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness*.

HU, X. AND DOWNIE, J. S. 2007. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the International Conference on Music Information Retrieval*.

HU, X., DOWNIE, J. S., LAURIER, C., BAY, M., AND EHMANN, A. F. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the International Conference on Music Information Retrieval*. 462–467.

HU, X., SANGHVI, V., VONG, B., ON, P. J., LEONG, C., AND ANGELICA, J. 2008. Moody: A web-based music mood classification and recommendation system. In *Proceedings of the International Conference on Music Information Retrieval*.

HU, X., DOWNIE, J. S., AND EHMANN, A. F. 2009. Lyric text mining in music mood classification. In *Proceedings of the International Conference on Music Information Retrieval*.

HU, Y., CHEN, X., AND YANG, D. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of the International Conference on Music Information Retrieval*.

HUQ, A., BELLO, J. P., SARROFF, A., BERGER, J., AND ROWE, R. 2009. Sourcetone: An automated music emotion recognition system. In *Proceedings of the International Conference on Music Information Retrieval*.

HUQ, A., BELLO, J. P., AND ROWE, R. 2010. Automated music emotion recognition: A systematic evaluation. *J. New Music Res. 39,* 3, 227–244.

HURON, D. 2000. Perceptual and cognitive applications in music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*.

HURON, D. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA.

JAIMES, A. AND SEBE, N. 2005. Multimodal human computer interaction: A survey. In *Proceedings of the IEEE International Workshop on HCI in Conjuction with the Computer Vision*. 1–15.

JAIMES, A., SEBE, N., AND GATICA-PEREZ, D. 2006. Human-centered computing: A multimedia perspective. In *Proceedings of the ACM International Conference on Multimedia*. 855–864.

JARGREAVES, D. J. AND NORTH, A. C. 1997. *The Social Psychology of Music*. Oxford University Press, Oxford, UK.

JIANG, D. N., LU, L., ZHANG, H. J., TAO, J. H., AND CAI, L. H. 2002. Music type classification by spectral contrast features. In *Proceedings of the IEEE International Conference on Multimedia Expo.* 113–116.

JONGHWA, K. AND ANDE, E. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Machine Intell. 30,* 12, 2067–2083.

JUSLIN, P. N. 2000. Cue utilization in communication of emotion in music performance: Relating performance to perception. *J. Exp. Psychol.: Human Percep. Perform. 16,* 6, 1797–1813.

JUSLIN, P. N. AND LAUKKA, P. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *J. New Music Res. 33,* 3, 217–238.

JUSLIN, P. N. AND SLOBODA, J. A. 2001. *Music and Emotion: Theory and Research*. Oxford University Press, Oxford, UK.

KARL F. MACDORMAN, S. O. AND HO, C.-C. 2007. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *J. New Music Res. 36,* 4, 281–299.

KATAYOSE, H., IMAI, M., AND INOKUCHI, S. 1998. Sentiment extraction in music. In *Proceedings of the International Conference on Pattern Recognition*. 1083–1087.

KIM, Y. E., SCHMIDT, E., AND EMELLE, L. 2008. Moodswings: A collaborative game for music mood label collection. In *Proceedings of the International Conference on Music Information Retrieval*. 231–236.

KIM, Y. E., SCHMIDT, E. M., MIGNECO, R., MORTON, B. G., RICHARDSON, P., SCOTT, J., SPECK, J. A., AND TURNBULL, D. 2010. Music emotion recognition: A state of the art review. In *Proceedings of the International Conference on Music Information Retrieval*.

KLAPURI, A. 1999. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 3089–3092.

KORHONEN, M. D., CLAUSI, D. A., AND JERNIGAN, M. E. 2006. Modeling emotional content of music using system identification. *IEEE Trans. Syst. Man Cyber. 36,* 3, 588–599.

KRUMHANSL, C. 2002. Music: A link between cognition and emotion. *Current Directions Psychol. Sci. 11,* 2, 45–50.

LAAR, B. 2006. Emotion detection in music, a survey. In *Proceedings of the Twente Student Conference on IT*.

LAGRANGE, M., MARTINS, L., MURDOCH, J., AND TZANETAKIS, G. 2008. Normalized cuts for predominant melodic source separation. *IEEE Trans. Audio, Speech, Lang. Process. 16,* 2, 278–290.

LAMERE, P. 2008. Social tagging and music information retrieval. *J. New Music Res. 37,* 2, 101–114.

LANG, P. J. 1995. The emotion probe. *Amer. Psychol. 50,* 5, 372–290.

LARTILLOT, O. AND TOIVIAINEN, P. 2007. MIR in Matlab (II): A toolbox for musical feature extraction from audio. In *Proceedings of the International Conference on Music Information Retrieval*. 127–130. `http://users.jyu.fi/~lartillo/mirtoolbox/`.

LAURIER, C. AND HERRERA, P. 2007. Audio music mood classification using support vector machine. In *Proceedings of the International Conference on Music Information Retrieval*.

LAURIER, C. AND HERRERA, P. 2008. Mood cloud: A real-time music mood visualization tool. In *Proceedings of the Computer Music Modeling and Retrieval*.

LAURIER, C., SORDO, M., SERRÀ, J., AND HERRERA, P. 2004. Digital music interaction concepts: A user study. In *Proceedings of the International Conference on Music Information Retrieval*. 415–420.

LAURIER, C., GRIVOLLA, J., AND HERRERA, P. 2008. Multimodal music mood classification using audio and lyrics. In *Proceedings of the International Conference on Machine Learning and Applications*. 105–111.

LAURIER, C., SORDO, M., AND HERRERA, P. 2009. Mood cloud 2.0: Music mood browsing based on social networks. In *Proceedings of the International Conference on Music Information Retrieval*.

LAURIER, C., SORDO, M., SERRÀ, J., AND HERRERA, P. 2009. Music mood representations from social tags. In *Proceedings of the International Conference on Music Information Retrieval*. 381–386.

LAW, E. L. M., VON AHN, L., DANNENBERG, R. B., AND CRAWFORD, M. 2007. TagATune: A game for music and sound annotation. In *Proceedings of the International Conference on Music Information Retrieval*.

LAZARUS, R. S. 1991. *Emotion and Adaptation*. Oxford University Press, Oxford, UK.

LEE, J. H. AND DOWNIE, J. S. 2004. Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In *Proceedings of the International Conference on Music Information Retrieval*. 441–446.

LEE, C.-M. AND NARAYANAN, S. S. 2005. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process. 13,* 2, 293–303.

LEMAN, M., VERMEULEN, V., VOOGDT, L. D., MOELANTS, D., AND LESAFFRE, M. 2005. Prediction of musical affect using a combination of acoustic structural cues. *J. New Music Res. 34,* 1, 39–67.

LEVY, M. AND SANDLER, M. 2007. A semantic space for music derived from social tags. In *Proceedings of the International Conference on Music Information Retrieval*. 411–416.

LEW, M., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Trans. Multimedia Comput. Comm. Appl. 2*, 1–19.

LEWIS, D. D., YANG, Y., ROSE, T. G., AND LI, F. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res. 5*, 361–397.

LI, T. AND OGIHARA, M. 2003. Detecting emotion in music. In *Proceedings of the International Conference on Music Information Retrieval*. 239–240.

LI, T. AND OGIHARA, M. 2004. Content-based music similarity search and emotion detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 17–21.

LIN, Y.-P., WANG, C.-H., WU, T.-L., JENG, S.-K., AND CHEN, J.-H. 2008. Support vector machine for EEG signal classification during listening to emotional music. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*. 127–130.

LIN, Y.-C., YANG, Y.-H., AND CHEN, H.-H. 2009. Exploiting genre for music emotion classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo.* 618–621.

LIN, Y.-P., JUNG, T.-P., AND CHEN, J.-H. 2009. EEG dynamics during music appreciation. In *Proceedngs of the IEEE International Conference on Engineering in Medicine and Biology Society*.

LIN, Y.-P., WANG, C.-H., WU, T.-L., JENG, S.-K., AND CHEN, J.-H. 2009. EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 489–492.

LIU, D., LU, L., AND ZHANG, H.-J. 2003. Automatic music mood detection from acoustic music data. In *Proceedings of the International Conference on Music Information Retrieval*. 81–87.

LIU, C. C., YANG, Y.-H., WU, P.-H., AND CHEN, H. H. 2006. Detecting and classifying emotion in popular music. In *Proceedings of the Joint International Conference on Information Sciences*. 996–999.

LIVINGSTONE, S. R. AND BROWN, A. R. 2005. Dynamic response: A real-time adaptation for music emotion. In *Proceedings of the Australasian Conference on Interactive Entertainment*. 105–111.

LJUNG, L. 1999. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, NJ.

LU, L., LIU, D., AND ZHANG, H. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech Lang. Process. 14,* 1, 5–18.

LU, Q., CHEN, X., YANG, D., AND WANG, J. 2010. Boosting for multi-modal music emotion classification. In *Proceedings of the International Conference on Music Information Retrieval*.

MACDORMAN, K. F., OUGH, S., AND HO, C.-C. 2007. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *J. New Music Res. 36,* 4, 281–299.

MADDAGE, N. C., XU, C., KANKANHALLI, M. S., AND SHAO, X. 2004. Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the ACM International Conference on Multimedia*. 112–119.

MANDEL, M. I. AND ELLIS, D. P. W. 2007. A web-based game for collecting music metadata. In *Proceedings of the International Conference on Music Information Retrieval*.

MCKAY, C., MCENNIS, D., AND FUJINAGA, I. 2006. A large publicly accessible prototype audio database for music research. In *Proceedings of the International Conference on Music Information Retrieval*. 160–163.

MEYERS, O. C. 2007. A mood-based music classification and exploration system. M.S. thesis, Massachusetts Institute of Technology.

MONTGOMERY, D. C., RUNGER, G. C., AND HUBELE, N. F. 1998. *Engineering Statistics*. Wiley, New York, NY.

MORTON, B. G., SPECK, J. A., SCHMIDT, E. M., AND KIM, Y. E. 2010. Improving music emotion labeling using human computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 45–48.

NIELSEN, F. V. 1986. Musical 'tension' and related concepts. In *Semiotic Web*, 491–513.

NWE, T. L. AND LI, H. 2007a. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Trans. Audio, Speech, Lang. Process. 15,* 2, 519–530.

NWE, T. L. AND LI, H. 2007b. Singing voice detection using perceptually-motivated features. In *Proceedings of the ACM International Conference on Multimedia*. 309–312.

OSGOOD, C. E., SUCI, G. J., AND TANNENBAUM, P. H. 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL.

OVADIA, S. 2004. Ratings and rankings: Reconsidering the structure of values and their measurement. *Int. J. Social Res. Method. 7,* 5, 403–414.

PAMPALK, E. 2004. A Matlab toolbox to compute music similarity from audio. In *Proceedings of the International Conference on Music Information Retrieval*. http://www.ofai.at/~elias.pampalk/ma/.

PEETERS, G. 2008. A generic training and classification system for MIREX08 classification tasks: Audio music mood, audio genre, audio artist and audio tag. In *Proceedings of the International Conference on Music Information Retrieval*.

PICARD, R. W., VYZAS, E., AND HEALEY, J. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell. 23,* 10, 1175–1191.

PLATT, J. C. 1999. *Probabilities for Support Vector Machines*. MIT Press, Cambridge, MA.

PLUTCHIK, R. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York, NY.

RASMUSSEN, C. E. AND WILLIAMS, C. K. I. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. http://www.gaussianprocess.org/gpml/.

REDDY, S. AND MASCIA, J. 2006. Lifetrak: Music in tune with your life. In *Proceedings of the Human-Centered Multimedia*. 25–34.

REMINGTON, N. A., FABRIGAR, L. R., AND VISSER, P. S. 2000. Reexamining the circumplex model of affect. *J. Personality Social Psychol. 79*, 286–300.

REPP, B. H. 1998. A microcosm of musical expression. i. quantitative analysis of pianists' timing in the initial measures of chopin's etude in e major. *J. Acoustic. Soc. Amer. 104*, 1085–1100.

RIGG, M. G. 1964. The mood effects of music: A comparison of data from four investigators. *J. Psychol. 58*, 427–438.

ROSS, R. T. 1938. A statistics for circular scales. *J. Edu. Psychol. 29*, 384 – 389.

RUSSELL, J. A. 1980. A circumplex model of affect. *J. Personal. Social Psychol. 39,* 6, 1161–1178.

RUSSELL, J. A. 2003. Core affect and the psychological construction of emotion. *Psychol. Review 110,* 1, 145–172.

RUSSELL, J. A., WEISS, A., AND G. A, M. 1989. Affect grid: A single-item scale of pleasure and arousal. *J. Personal. Social Psychol. 57,* 3, 493–502.

SCARINGELLA, N., ZOIA, G., AND MLYNEK, D. 2006. Automatic genre classification of music content: A survey. *IEEE Signal Process. Mag. 23,* 2, 133–141.

SCHMIDT, E. M. AND KIM, Y. E. 2009. Projection of acoustic features to continuous valence-arousal mood. In *Proceedings of the International Conference on Music Information Retrieval*.

SCHMIDT, E. M., TURNBULL, D., AND KIM, Y. E. 2010. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*. 267–274.

SCHÖLKOPF, B., SMOLA, A. J., WILLIAMSON, R. C., AND BARTLETT, P. L. 2000. New support vector algorithms. *Neural Comput. 12*, 1207–1245.

SCHUBERT, E. 1999. Measurement and time series analysis of emotion in music. Ph.D. dissertion, School of Music Education, University of New South Wales, Sydney, Australia.

SCHUBERT, E. 2001. Correlation analysis of continuous response to music: Correcting for the effects of serial correlation. *Musicae Scientiae*, 213–236.

SCHUBERT, E. 2003. Update of the Hevner adjective checklist. *Perceptual Motor Skills 96*, 1117–1122.

SCHULLER, B., EYBEN, F., AND RIGOLL, G. 2008. Tango or waltz? Putting ballroom dance style into tempo detection. *EURASIP J. Audio, Speech, Music Process*. Article ID 846135.

SCHULLER, B., STEIDL, S., AND BATLINER, A. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proceedings of the INTERSPEECH Conference*.

SCHULLER, B., DORFNER, J., AND RIGOLL, G. 2010. Determination of nonprototypical valence and arousal in popular music: Features and performances. *EURASIP J. Audio, Speech, Music Process*. Article ID 735854.

SCHULLER, B., HAGE, C., SCHULLER, D., AND RIGOLL, G. 2010. Mister D. J., cheer me up!: Musical and textual features for automatic mood classification. *J. New Music Res. 39,* 1, 13–34.

SEBASTIANI, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surveys 34,* 1, 1–47.

SEN, A. AND SRIVASTAVA, M. 1990. *Regression Analysis: Theory, Methods, and Applications*. Springer, New York, NY.

SHAO, B., LI, T., AND OGIHARA, M. 2008. Quantify music artist similarity based on style and mood. In *Proceedings of the ACM Workshop on Web Information and Data Management*. 119–124.

SHEN, J., CUI, B., SHEPHERD, J., AND TAN, K.-L. 2006. Towards efficient automated singer identification in large music databases. In *Proceedings of the ACM International Conference on Information Retrieval*. 59–66.

SKOWRONEK, J., MCKINNEY, M. F., AND VAN DE PAR, S. 2006. Ground truth for automatic music mood classification. In *Proceedings of the International Conference on Music Information Retrieval*. 395–396.

SKOWRONEK, J., MCKINNEY, M. F., AND VAN DE PAR, S. 2007. A demonstrator for automatic music mood estimation. In *Proceedings of the International Conference on Music Information Retrieval*.

SLOBODA, J. A. AND JUSLIN, P. N. 2001. Psychological perspectives on music and emotion. In *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda Eds., Oxford University Press, Oxford, UK.

SOLOMATINE, D. AND SHRESTHA, D. 2004. AdaBoost.RT: A boosting algorithm for regression problems. In *Proceedings of the IEEE International Joint Conference Neural Networks*. 1163–1168.

THAYER, R. E. 1989. *The Biopsychology of Mood and Arousal*. Oxford University Press, Oxford, UK.

TOIVIAINEN, P. AND KRUMHANSL, C. L. 2003. Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception 32,* 6, 741–766.

TOLONEN, T. AND KARJALAINEN, M. 2000. A computationally efficient multipitch analysis model. *IEEE Trans. Speech Audio Process. 8,* 6, 708–716.

TOLOS, M., TATO, R., AND KEMP, T. 2005. Mood-based navigation through large collections of musical data. In *Proceedings of the IEEE Consumer Communications & Network Conference.* 71–75.

TROHIDIS, K., TSOUMAKAS, G., KALLIRIS, G., AND VLAHAVAS, I. 2008. Multi-label classification of music into emotions. In *Proceedings of the International Conference on Music Information Retrieval*. 325–330.

TSAI, W.-H. AND WANG, H.-M. 2006. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Trans. Audio, Speech Lang. Process. 14,* 1, 330–341.

TURNBULL, D., LIU, R., BARRINGTON, L., AND LANCKRIET, G. 2007. A game-based approach for collecting semantic annotations of music. In *Proceedings of the International Conference on Music Information Retrieval*.

TURNBULL, D., BARRINGTON, L., TORRES, D., AND LANCKRIET, G. 2008. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. Audio, Speech Lang. Process. 16,* 2, 467–476.

TZANETAKIS, G. 2007. MARSYAS submissions to MIREX 2007. In *Proceedings of the International Conference on Music Information Retrieval*.

TZANETAKIS, G. AND COOK, P. 2002. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process. 10,* 5, 293–302. http://marsyas.sness.net/.

VAN ZAANEN, M. AND KANTERS, P. 2010. Automatic mood classification using tf*idf based on lyrics. In *Proceedings of the International Conference on Music Information Retrieval*.

VERCOE, G. S. 2006. Moodtrack: practical methods for assembling emotion-driven music. M.S. thesis, MIT, Cambridge, MA.

VERVERIDIS, D. AND KOTROPOULOS, C. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Comm. 48,* 9, 1162–1181.

WANG, H. L. AND CHEONG, L. F. 2006. Affective understanding in film. *IEEE Trans. Circuits Syst. Video Technol. 16,* 6, 689–704.

WANG, M.-Y., ZHANG, N.-Y., AND ZHU, H.-C. 2004. User-adaptive music emotion recognition. In *Proceedings of the IEEE International Conference on Signal Processing*. 1352–1355.

Whissell, C. M., Fournier, M., Pelland, R., Weir, D., and Makarec, K. 21986. A dictionary of affect in language: IV. reliability, validity, and applications. *Perceptual Motor Skills 62*, 875–888.

Wieczorkowska, A. 2004. Towards extracting emotions from music. In *Proceedings of the International Workshop on Intelligent Media Technology for Communicative Intelligence*. 228–238.

Wieczorkowska, A., Synak, P., and Raś, Z. W. 2006. Multi-label classification of emotions in music. In *Proceedings of the International Conference on Intelligent Information Processing and Web Mining*. 307–315.

Wu, T.-L. and Jeng, S.-K. 2006. Automatic emotion classification of musical segments. In *Proceedings of the International Conference on Music Perception and Cognition*.

Wu, T.-L. and Jeng, S.-K. 2007. Regrouping of expressive terms for musical qualia. In *Proceedings of the International Workshop on Computer Music Audio Technology*.

Wu, T.-L. and Jeng, S.-K. 2008. Probabilistic estimation of a novel music emotion model. In *Proceedings of the International Multimedia Modeling Conference* 487–497.

Wu, T.-L., Wang, H.-K., Ho, C.-C., Lin, Y.-P., Hu, T.-T., Weng, M.-F., Chan, L.-W., Yang, C.-H., Yang, Y.-H., Hung, Y.-P., Chuang, Y.-Y., Chen, H.-H., Chen, H. H., Chen, J.-H., and Jeng, S.-K. 2008. Interactive content presenter based on expressed emotion and physiological feedback. In *Proceedings of the ACM International Conference on Multimedia*. 1009–1010.

Xiao, Z., Dellandrea, E., Dou, W., and Chen, L. 2008. What is the best segment duration for music mood analysis. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing*. 17–24.

Yang, D. and Lee, W.-S. 2004. Disambiguating music emotion using software agents. In *Proceedings of the International Conference onMusic Information Retrieval*.

Yang, Y.-H. and Chen, H. H. 2011a. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Trans. Audio, Speech Lang. Process 19,* 7, 2184–2196.

Yang, Y.-H. and Chen, H. H. 2011b. Ranking-based emotion recognition for music organization and retrieval. *IEEE Trans. Audio, Speech Lang. Process 19*, 4, 762–774.

Yang, Y.-H. and Chen, H. H. 2011c. *Music Emotion Recognition*. CRC Press, Boca Raton, FL.

Yang, Y.-H., Liu, C. C., and Chen, H. H. 2006. Music emotion classification: A fuzzy approach. In *Proceedings of the ACM International Conference on Multimedia*. 81–84.

Yang, Y.-H., Su, Y.-F., Lin, Y.-C., and Chen, H. H. 2007. Music emotion recognition: The role of individuality. In *Proceedings of the ACM International Workshop on Human-Centered Multimedia*. 13–21. `http://mpac.ee.ntu.edu.tw/~yihsuan/MER/hcm07/`.

Yang, Y.-H., Lin, Y.-C., and Chen, H. H. 2009. Personalized music emotion recognition. In *Proceedings of the ACM International Conference on Information Retrieval*. 748–749.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H. 2008. A regression approach to music emotion recognition. *IEEE Trans. Audio, Speech Lang. Process. 16,* 2, 448–457.

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., and Chen, H. H. 2008. Mr. Emo: Music retrieval in the emotion plane. In *Proceedings of the ACM International Conference on Multimedia*. 1003–1004. `http://www.youtube.com/watch?v=ra55x020UHU`.

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., and Chen, H.-H. 2008. Toward multimodal music emotion classification. In *Proceedings of the Pacific-Rim Conference on Multimedia*. 70–79.

Yeh, C.-C., Tseng, S.-S., Tsai, P.-C., and Weng, J.-F. 2006. Building a personalized music emotion prediction system. In *Proceedings of the Pacific-Rim Conference on Multimedia*. 730–739.

Zhang, M.-L. and Zhou, Z.-H. 2007. ML-knn: A lazy learning approach to multi-label learning. *Pattern Recogn. 40,* 7, 2038–2048.

Zhang, S., Huang, Q., Tian, Q., Jiang, S., and Gao, W. 2008. *i*.MTV – an integrated system for MTV affective analysis. In *Proceedings of the ACM International Conference on Multimedia*. 985–986.

Zhang, S., Tian, Q., Jiang, S., Huang, Q., and Gao, W. 2008. Affective MTV analysis based on arousal and valence features. In *Proceedings of the IEEE International Conference on Multimedia and Expo.* 1369–1372.

Zhang, S., Tian, Q., Huang, Q., Gao, W., and Li, S. 2009. Utilizing affective analysis for efficient movie browsing. In *Proceedings of the IEEE International Conference on Image Processing*. 1853–1856.

Received May 2010; revised August 2010; accepted October 2010