MUSIC EMOTION RECOGNITION: A STATE OF THE ART REVIEW

Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton
Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull[†]
Electrical and Computer Engineering, Drexel University
Computer Science, Ithaca College[†]
{ykim, eschmidt, rmigneco, bmorton,
patrickr, jjscott, jspeck}@drexel.edu
turnbull@ithaca.edu[†]

ABSTRACT

This paper surveys the state of the art in automatic emotion recognition in music. Music is oftentimes referred to as a "language of emotion" [1], and it is natural for us to categorize music in terms of its emotional associations. Myriad features, such as harmony, timbre, interpretation, and lyrics affect emotion, and the mood of a piece may also change over its duration. But in developing automated systems to organize music in terms of emotional content, we are faced with a problem that oftentimes lacks a well-defined answer; there may be considerable disagreement regarding the perception and interpretation of the emotions of a song or ambiguity within the piece itself. When compared to other music information retrieval tasks (e.g., genre identification), the identification of musical mood is still in its early stages, though it has received increasing attention in recent years. In this paper we explore a wide range of research in music emotion recognition, particularly focusing on methods that use contextual text information (e.g., websites, tags, and lyrics) and content-based approaches, as well as systems combining multiple feature domains.

1. INTRODUCTION

With the explosion of vast and easily-accessible digital music libraries over the past decade, there has been a rapid expansion of music information retrieval research towards automated systems for searching and organizing music and related data. Some common search and retrieval categories, such as artist or genre, are more easily quantified to a "correct" (or generally agreed-upon) answer and have received greater attention in music information retrieval research. But music itself is the expression of emotions, which can be highly subjective and difficult to quantify. Automatic recognition of emotions (or mood) [1] in music is still in its early stages, though it has received increasing attention in recent years. Determining the emotional content of music audio computationally is, by nature, a cross-disciplinary endeavor spanning not only signal processing and machine learning, but also requiring an understanding of auditory perception, psychology, and music theory.

Computational systems for music mood recognition may be based upon a model of emotion, although such representations remain an active topic of psychology research. Categorical and parametric models are supported through substantial prior research with human subjects, and these models will be described in further detail in the sections that follow. Both models are used in Music-IR systems, but the collection of "ground truth" emotion labels, regardless of the representation being used, remains a particularly challenging problem. A variety of efforts have been made towards efficient label collection, spanning a wide range of potential solutions, such as listener surveys, social tags, and data collection games. A review of methods for emotion data collection for music is also a subject of this paper.

The annual Music Information Research Evaluation eXchange (MIREX) is a community-based framework for formally evaluating Music-IR systems and algorithms [2], which included audio music mood classification as a task for the first time in 2007 [3]. The highest performing systems in this category demonstrate improvement each year using solely acoustic features (note that several of the systems were designed for genre classification and then appropriated to the mood classification task, as well). But emotion is not completely encapsulated within the audio alone (social context, for example, plays a prominent role), so approaches incorporating music metadata, such as tags and lyrics, are also reviewed here in detail.

For this state-of-the-art review of automatic emotion recognition in music, we first discuss some of the psychological research used in forming models of emotion, and then detail computational representations for emotion data. We present a general framework for emotion recognition that is subsequently applied to the different feature domains. We conclude with an overview of systems that combine multiple modalities of features.

[1] Emotion and mood are used interchangeably in the literature and in this paper.

2. PSYCHOLOGY RESEARCH ON EMOTION

Over the past half-century, there have been several important developments spanning multiple approaches for qualifying and quantifying emotions related to music. Such inquiry began well before the widespread availability of music recordings as a means of clinically repeatable musical stimuli (using musical scores), but recordings are the overwhelmingly dominant form of stimulus used in modern research studies of emotion. Although scores can provide a wealth of relevant information, score-reading ability is not universal, and our focus in this section and the overall paper shall be limited to music experienced through audition.

2.1 Perceptual considerations

When performing any measurement of emotion, from direct biophysical indicators to qualitative self-reports, one must also consider the source of emotion being measured. Many studies, using categorical or scalar/vector measurements, indicate the important distinction between one's perception of the emotion(s) *expressed* by music and the emotion(s) *induced* by music [4,5]. Both the emotional response and its report are subject to confound. Early studies of psychological response to environment, which consider the emotional weight of music both as a focal and distracting stimulus, found affective response to music can also be sensitive to the environment and contexts of listening [6]. Juslin and Luakka, in studying the distinctions between perceptions and inductions of emotion, have demonstrated that *both* can can be subject to not only the social context of the listening experience (such as audience and venue), but also personal motivation (i.e., music used for relaxation, stimulation, etc.) [5]. In the remainder of this paper, we will focus on systems that attempt to discern the emotion expressed, rather than induced, by music.

2.2 Perception of emotion across cultures

Cross-cultural studies of musical power suggest that there may be universal psychophysical and emotional cues that transcend language and acculturation [7]. Comparisons of tonal characteristics between Western 12-tone and Indian 24-tone music suggest certain universal mood-targeted melodic cues [8]. In a recent ethnomusicology study of people with no exposure to Western music (or culture), Mafa natives of Cameroon, categorized music examples into three categories of emotion in the same way as Westerners [9].

2.3 Representations of emotion

Music-IR systems tend to use either categorical descriptions or parametric models of emotion for classification or recognition. Each representation is supported by a large body of supporting psychology research.

2.3.1 Categorical psychometrics

Categorical approaches involve finding and organizing some set of emotional descriptors (tags) based on their relevance to some music in question. One of the earliest studies by Hevner, published in 1936, initially used 66 adjectives, which were then arranged into 8 groups [10]. While the adjectives used and their specific grouping and hierarchy have remained scrutinized and even disputed, many categorical studies conducted since Hevner's indicate such tagging can be intuitive and consistent, regardless of the listener's musical training [11,12].

In a recent sequence of music-listening studies Zenter *et al.* reduced a set of 801 "general" emotional terms into a subset metric of 146 terms specific for music mood rating. Their studies, which involved rating music-specificity of words and testing words in lab and concert settings with casual and genre-aficionado listeners, revealed that the interpretation of these mood words varies between different genres of music [13].

The recent MIREX evaluations for automatic music mood classification have categorized songs into one of five mood clusters, shown in Table 1. The five categories were derived by performing clustering on a co-occurrence matrix of mood labels for popular music from the All Music Guide[2] [3].

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

Table 1. Mood adjectives used in the MIREX Audio Mood Classification task [3].

2.3.2 Scalar/dimensional psychometrics

Other research suggests that mood can be scaled and measured by a continuum of descriptors or simple multi-dimensional metrics. Seminal work by Russell and Thayer in studying dimensions of arousal established a foundation upon which sets of mood descriptors may be organized into low-dimensional models. Most noted is the two-dimensional *Valence-Arousal* (V-A) space (See Figure 1), where emotions exist on a plane along independent axes of arousal (intensity), ranging high-to-low, and valence (an appraisal of polarity), ranging positive-to-negative [4]. The validity of this two-dimensional representation of emotions for a wide range of music has been confirmed in multiple studies [11,14].

Some studies have expanded this approach to develop three-dimensional spatial metrics for comparative analysis of musical excerpts, although the semantic nature of the third dimension is subject to speculation and disagreement [17]. Other investigations of the V-A model itself

[2] All Music Guide: http://www.allmusic.com

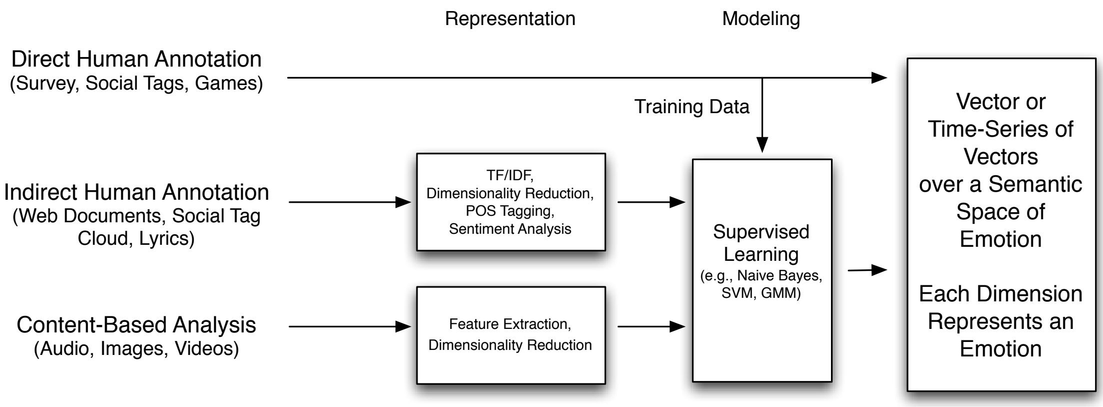
Representation
Modeling
Direct Human Annotation
(Survey, Social Tags, Games)
Training Data
Indirect Human Annotation
(Web Documents, Social Tag Cloud, Lyrics)
TF/IDF, Dimensionality Reduction, POS Tagging, Sentiment Analysis
Content-Based Analysis
(Audio, Images, Videos)
Feature Extraction, Dimensionality Reduction
Supervised Learning
(e.g., Naive Bayes, SVM, GMM)
Vector or Time-Series of Vectors over a Semantic Space of Emotion
Each Dimension Represents an Emotion

Figure 2. Overall model of emotion classification systems.

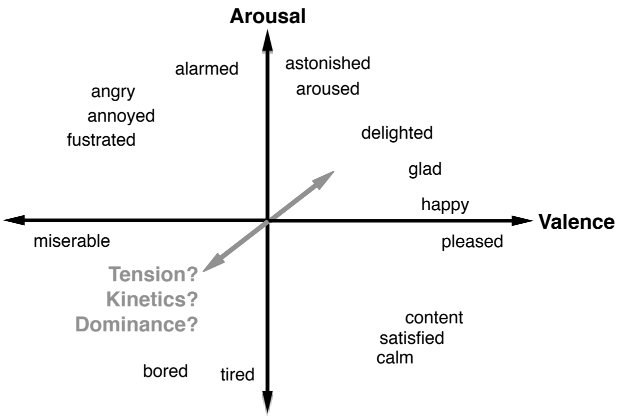

Figure 1. The Valence-Arousal space, labeled by Russell's direct circular projection of adjectives [4]. Includes semantic of projected third affect dimensions: "tension" [15], "kinetics" [16], "dominance" [6].

suggest evidence for separate channels of arousal (as originally proposed by Thayer) that are not elements of valence [18].

A related, but categorical, assessment tool for self-reported affect is the Positive and Negative Affect Schedule (PANAS), which asserts that all discrete emotions (and their associated labels) exist as incidences of positive or negative affect, similar to valence [19, 20]. In this case, however, positive and negative are treated as separate categories as opposed to the parametric approach of V-A.

3. FRAMEWORK FOR EMOTION RECOGNITION

Emotion recognition can be viewed as a multiclass-multilabel classification or regression problem where we try to annotate each music piece with a set of emotions. A *music piece* might be an entire song, a section of a song (e.g., chorus, verse), a fixed-length clip (e.g., 30-second song snipet), or a short-term segment (e.g., 1 second).

We will attempt to represent mood as either a single multi-dimensional vector or a time-series of vectors over a semantic space of emotions. That is, each dimension of a vector represents a single emotion (e.g., angry) or a bi-polar pair of emotions (e.g., positive/negative). The value of a dimension encodes the strength-of-semantic-association between the piece and the emotion. This is sometimes represented with a binary label to denote the presence or absence of the emotion, but more often represented as a real-valued score (e.g., Likert scale value, probability estimate). We will represent emotion as a time-series of vectors if, for example, we are attempting to track changes in emotional content over the duration of a piece.

We can estimate values of the emotion vector for a music piece in a number of ways using various forms of data. First, we can ask human listeners to evaluate the relevance of an emotion for a piece (see Section 4). This can be done, for example, using a survey, a social tagging mechanism, or an annotation game. We can also analyze forms of contextual meta-data in text form (see Section 5). This may include text-mining web-documents (e.g., artist biographies, album reviews) or a large collection of social tags (referred to as a *tag cloud*), and analyzing lyrics using natural language processing (e.g., sentiment analysis). We can also analyze the audio content using both signal processing and supervised machine learning to automatically annotate music pieces with emotions (see Section 6). Content-based methods can also be used to analyze other related forms of multimedia data such as music videos and promotional photographs [21]. Furthermore, multiple data sources, for example lyrics and audio, may be combined to determine the emotional content of music (see Section 7).

4. HUMAN ANNOTATION

A survey is a straightforward technique for collecting information about emotional content in music. All Music Guide has devoted considerable amounts of money, time and human resources to annotate their music databases with high-quality emotion tags. As such, they are unlikely to fully share this data with the Music-IR research community. To remedy this problem, Turnbull *et al.* collected the CAL500 data set of annotated music [22]. This data set contains one song from 500 unique artists each of which have been manually annotated by a minimum of three non-expert reviewers using a vocabulary of 174 tags, of which 18 relate to different emotions. Trohidis *et al.* have also created a publicly available data set consisting of 593 songs each of which have been annotated using 6 emotions by 3 expert listeners [23].

A second approach to directly collect emotion annotations from human listeners involves social tagging. For example, Last.fm [3] is a music discovery website that allows users to contribute *social* tags through a text box in their audio player interface. By the beginning of 2007, their large base of 20 million monthly users have built up an unstructured vocabulary of 960,000 free-text tags and used it to annotate millions of songs [24]. Unlike AMG, Last.fm makes much of this data available to the public through their public APIs. While this data is a useful resource for the Music-IR community, Lamere and Celma point out that there are a number of problems with social tags: sparsity due to the cold-start problem and popularity bias, ad-hoc labeling techniques, multiple spellings of tags, malicious tagging, etc. [25]

4.1 Annotation Games

Traditional methods of data collection, such as the hiring of subjects, can be flawed, since labeling tasks are time-consuming, tedious, and expensive [26]. Recently, a significant amount of attention has been placed on the use of collaborative online games to collect such ground truth labels for difficult problems, so-called “Games With a Purpose”. Several such games have been been proposed for the collection of music data, such as *MajorMiner* [27], *Listen Game* [28], and *TagATune* [29]. These implementations have primarily focused on the collection of descriptive labels for a relatively short audio clip. Screenshots of a few, select games are shown in Figure 3.

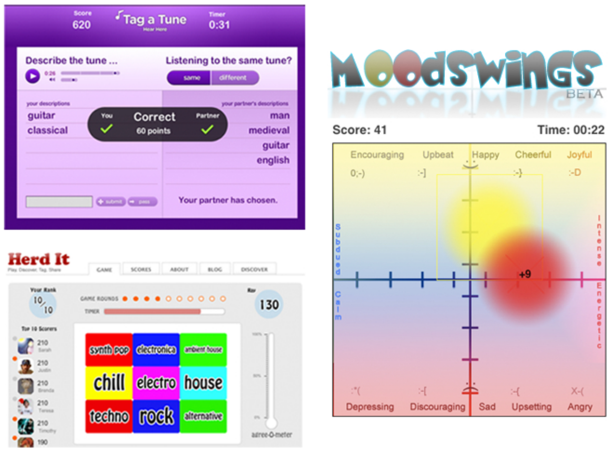

Figure 3. Examples of “Games With A Purpose” for ground truth data collection of musical data. Top left: *TagATune*. Right: *MoodSwings*. Bottom left: *Herd It*.

MoodSwings is another game for online collaborative annotation of emotions based on the arousal-valence model [30, 31]. In this game, players position their cursor within the V-A space while competing (and collaborating) with a partner player to annotate 30-second music clips where scoring is determined by the overlap between players’ cursors (encouraging consensus and discouraging nonsense labels). Using a similar parametric representation, Bachorik *et al.* concluded that most music listeners require 8 seconds to evaluate the mood of a song, a delay that should be considered when collecting such time-varying annotations [32]. *Herd It* combines multiple types of music annotation games, including valence-arousal annotation of clips, descriptive labeling, and music trivia [33].

5. CONTEXTUAL TEXT INFORMATION

In this section, we discuss web documents, social tag clouds and lyrics as forms of textual information that can be analyzed in order to derive an emotional representation of music. Analysis of these data sources involves using techniques from both text-mining and natural language processing.

5.1 Web-Documents

Artist biographies, album reviews, and song reviews are rich sources of information about music. There are a number of research-based Music-IR systems that collect such documents from the Internet by querying search engines [34], monitoring MP3 blogs [35], or crawling a music website [36]. In all cases, Levy and Sandler point out that such web mined corpora can be *noisy* since some of the retrieved webpages will be irrelevant, and in addition, much of the text content on relevant webpages will be useless [37].

Most of the proposed web mining systems use a set of one or more documents associated with a song and convert them into a single document vector (e.g., Term Frequency-Inverse Document Frequency (TF-IDF) representation) [38,39]. This *vector space* representation is then useful for a number of Music-IR tasks such as calculating music similarity [39] and indexing content for a text-based music retrieval system [38]. More recently, Knees *et al.* have proposed a promising new web mining technique called *relevance scoring* as an alternative to the vector space approaches [34].

5.2 Social Tags

Social tags have been used to accomplish such Music-IR tasks as genre and artist classification [40] as well as assessment of musical mood. Some tags such as “happy” and “sad” are clearly useful for emotion recognition and can be applied directly in information retrieval systems. Research has also shown that other tags, such as those related to genre and instrumentation, can also be useful for this task. Using ground truth mood labels from AMG, Bischoff *et al.* used social tag features from Last.fm to perform emotion classification based on MIREX mood categories as well as the V-A model [41]. They experimented with SVM, Logistic Regression, Random Forest, GMM, K-NN, Decision Trees, and Naive Bayes Multinomial classifiers, with

[3] Last.fm: http://www.last.fm

the Naive Bayes Multinomial classifier outperforming all other methods.

Other research involving the analysis of social tags has focused on clustering tags into distinct emotions and validating psychometric models. Making each tag a unique class would yield an unmanageable number of dimensions and fails to take into account the similarity of many terms used to describe musical moods. For example, the terms "bright", "joyful", "merry" and "cheerful" describe similar variants of happiness. Similarly, the tokens "gloomy", "mournful", "melancholy" and "depressing" are all related to sadness [10]. Recent efforts have demonstrated that favorable classification results can be obtained by grouping like descriptors into similarity clusters [14].

A number of approaches exist to arrange tags together into homogeneous groups. Manual clustering involves grouping of tags into pre-established mood categories, but given the size and variety of existing tag databases, this approach is not scalable. A straightforward automated clustering method, derived from the TF-IDF metric used often in text mining, looks for co-occurrences within the mood tags and forms clusters until no more co-occurrences are present. The co-occurrence method compares a threshold to the ratio of the number of songs associated with two tags to the minimum number of songs associated with either individual tag [42].

Another established method for automatically clustering labels is Latent Semantic Analysis (LSA), a natural language processing technique that reduces a term-document matrix to a lower rank approximation [43]. The term-document matrix in this case is a sparse matrix which describes the number of times each song is tagged with a given label. For a data set with several thousand songs and over 100 possible mood tags, the term-document matrix generated will have very high dimensionality. After some modifications, performing a singular value decomposition (SVD) on the modified term-document matrix yields the left and right singular vectors that represent the distance between terms and documents respectively. Initial work by Levy and Sandler applied a variation called Correspondence Analysis to a collection of Last.fm social tags to derive a semantic space for over 24,000 unique tags spanning 5700 tracks.

Tags can also be grouped by computing the cosine distance between each tag vector and using an unsupervised clustering method, such as Expectation Maximization (EM), to combine terms. In recent work, Laurier *et al.*, using a cost function to minimize the number of clusters to best represent over 400,000 unique tags, found that just four clusters yielded optimal clustering of the mood space [14]. The resulting clusters are somewhat aligned with Russell and Thayer's V-A model. Furthermore, both Levy & Sandler and Laurier *et al.* demonstrate that application of a self organizing map (SOM) algorithm to their derived semantic mood spaces yields a two-dimensional representation of mood consistent with the V-A model.

5.3 Emotion recognition from lyrics

In comparison to tag-based approaches, relatively little research has pursued the use of lyrics as the sole feature for emotion recognition (although lyrics have been used as features for artist similarity determination [44]). Lyric-based approaches are particularly difficult because feature extraction and schemes for emotional labeling of lyrics are non-trivial, especially when considering the complexities involved with disambiguating affect from text. Lyrics have also been used in combination with other features, work that is detailed in Section 7.

5.3.1 Lyrics feature selection

Establishing "ground-truth" labels describing the emotion of interconnected words is a significant challenge in lyric-based emotion recognition tasks. Mehrabian and Thayer proposed that environmental stimuli are linked to behavioral responses by emotional responses described by pleasure (valence), arousal and dominance (PAD) [6]. To this end, Bradley developed the Affective Norms for English Words (ANEW), which consists of a large set of words labeled with PAD values. A large number of subjects were used to label the words by indicating how the word made them feel in terms of relative happiness, excitement and situational control, which correspond to the pleasure, arousal and dominance dimensions, respectively. A distribution of the pleasure and arousal labels for words in ANEW show that they are well-distributed according to the V-A model [45]. Hu *et al.* used Bradley's ANEW to develop a translation called Affective Norms for Chinese Words (ANCW), operating under the assumption that the translated words carry the same affective meaning as their English counterparts [46].

Such affective dictionaries do not take into account multi-word structure. For lyrics features, most approaches employ a Bag of Words (BOW) approach, accounting for frequency of word usage across the corpus (e.g., TF-IDF), but not the specific order of words. One initial approach by Chen *et al.* utilized vector space model (VSM) features that consisted of all the words comprising the lyric [47]. However, more recently Xia *et al.* refined the feature vectors by only including sentiment and sentiment-related words, which they refer to as a *sentiment*-VSM (s-VSM) [48]. The focus on sentiment-related words is intended to capture the effect of modifying terms strengthening or weakening the primary sentiments of the lyrics and further reduces feature dimensionality.

5.3.2 Systems for emotion recognition from lyrics

Meyers' *Lyricator* system provides an emotional score for a song based on its lyrical content for the purpose of mood-based music exploration [49]. Feature extraction for Lyricator consists of obtaining PAD labels for the words comprising a songs lyric. Songs receive an overall emotional score in one of the four quadrants of the P-A model based on a summation of the PAD values for all the words in the lyric. While this approach is straightforward, it is not a machine learning system, nor does it make use of natural language processing (NLP) to disambiguate emotion from lyrics.

Vector space approaches by Xia and Chen utilize support vector machine (SVM) classifiers for training and testing data [47, 48]. Xia's corpus consists of 2600 Chinese pop songs, 60% of which are hand-labeled as "light-hearted" and the remainder labeled as "heavy-hearted". Their sentiment-VSM feature set scores above 73% in terms of precision, recall, and F-1 measure.

Hu *et al.* utilize a fuzzy clustering technique to determine the main emotion from a lyric. The clusters are then weighted using grammatical information, which specify confidence and weights for individual sentences based on factors such as tense and inter-sentence relationships. The cluster with the greatest overall weight is considered the main emotion of the song and is characterized by its mean V-A values into one of the four quadrants of the V-A model. They demonstrate that their system outperforms the baseline Lyricator system in multiple categories [46].

Y. Yang explored the use of bi-gram BOW features, using pairs of words to examine the effects of negation terms (e.g., "not happy" differs from "happy") and Probabilistic LSA (PLSA) to model song topics using word frequencies [50]. Bi-gram BOW features demonstrated negligible increases in classification valence, but PLSA proved to be much more robust to a reduction of training set size.

6. CONTENT-BASED AUDIO ANALYSIS

Clearly, many human assessments of musical mood are derived from the audio itself (after all, tags are most often generated by people listening to the music). Contextual information for a music piece may be incomplete or missing entirely (i.e., for newly composed music). Given the rapid expansion of digital music libraries, including commercial databases with millions of songs, it is clear that manual annotation methods will not efficiently scale. Thus, the appeal of content-based systems is obvious and the recognition of emotions from audio has been a longstanding goal for the Music-IR research community (the corresponding MIREX task focused on systems driven by music audio).

6.1 Acoustic Features

Emotions can be influenced by such attributes as tempo, timbre, harmony, and loudness (to name only a few), and much prior work in Music-IR has been directed towards the development of informative acoustic features. Although some research has focused on searching for the most informative features for emotion classification, no dominant single feature has emerged. An overview of the most common acoustic features used for mood recognition is given in Table 2.

Type	Features
Dynamics	RMS energy
Timbre	MFCCs, spectral shape, spectral contrast
Harmony	Roughness, harmonic change, key clarity, majorness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo, beat histograms
Articulation	Event density, attack slope, attack time

Table 2. Common acoustic feature types for emotion classification.

In searching for the most informative emotion and expressive features to extract from audio, Mion and De Poli investigated a system for feature selection and demonstrated it on an initial set of single-dimensional features, including intensity and spectral shape as well as several music-theoretic features [16]. Their system used sequential feature selection (SFS), followed by principal component analysis (PCA) on the subset to identify and remove redundant feature dimensions. The focus of their research, however, was monophonic instrument classification across nine classes spanning emotion and expression, as opposed to musical mixtures. Of the 17 tested features the most informative overall were found to be roughness, notes per second, attack time, and peak sound level.

MacDorman *et al.* examined the ability of multiple acoustic features (sonogram, spectral histogram, periodicity histogram, fluctuation pattern, and mel-frequency cepstral coefficients–MFCCs [51, 52]) to predict pleasure and arousal ratings of music excerpts. They found all of these features to be better at predicting arousal than pleasure, and the best prediction results when all five features were used together [53].

Schmidt *et al.* investigated the use of multiple acoustic feature domains for music clips both in terms of individual performance as well as in combination in a feature fusion system [54, 55]. Their feature collection consisted of MFCCs, chroma [56], statistical spectrum descriptors (including centroid, flux, and rolloff, depicted in Figure 4) [57], and octave-based spectral contrast [58]. The highest performing individual features were spectral contrast and MFCCs, but again the best overall results were achieved using combinations of features.

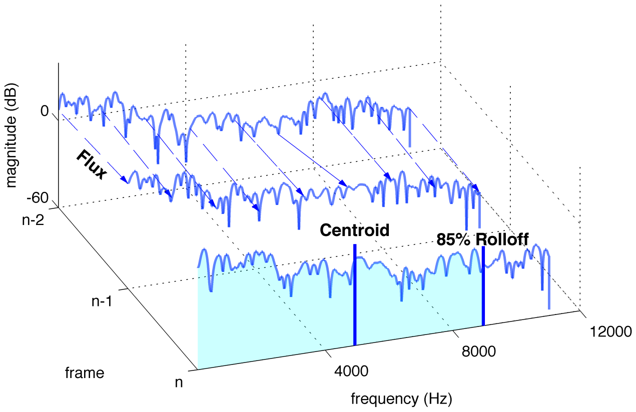

Figure 4. Graphical depiction of statistical spectrum descriptors.

Eerola *et al.*, also the developers of the open-source feature extraction code, *MIRtoolbox*,[4] have developed a specific subset of informative audio features for emotion. These features are aggregated from a wide range of domains including dynamics, timbre, harmony, register, rhythm, and articulation [15]. Other recent approaches have adopted a generalized approach to feature extraction, compiling multiple feature sets (resulting in high dimensional spaces) and employing dimensionality reduction techniques [50,59]. Regardless of feature fusion or dimensionality reduction methods, the most successful systems combine multiple acoustic feature types.

6.2 Audio-based systems

Audio content-based methods for emotion recognition use either a categorical or parametric model of emotion. The former results in a classification task and the latter in a regression task, with all recent systems employing one of these methods.

6.2.1 Categorical emotion classification

In one of the first publications on this topic, Li and Ogihara used acoustic features related to timbre, rhythm, and pitch to train support vector machines (SVMs) to classify music into one of 13 mood categories [60]. Using a hand-labeled library of 499 music clips (30-seconds each) from a variety of genres spanning ambient, classical, fusion, and jazz, they achieved an accuracy of 45%.

Lu *et al.* pursued mood detection and tracking using a similar variety of acoustic features including intensity, timbre, and rhythm [59]. Their classifier used Gaussian Mixture Models (GMMs) for the four principal mood quadrants on the V-A representation. The system was trained using a set of 800 classical music clips (from a data set of 250 pieces), each 20 seconds in duration, hand labeled to one of the 4 quadrants. Their system achieved an overall accuracy of 85%, although it is also unclear how the multiple clips extracted from the same recording were distributed between training and testing sets.

Proposing a guided scheme for music recommendation, Mandel *et al.* developed active learning systems, an approach that can provide recommendations based upon any musical context defined by the user [61]. To perform a playlist retrieval, the user would present the system with a set of "seed songs," or songs representing the class of playlist desired. The system uses this data, combined with verification data from the user, to construct a binary SVM classifier using MFCC features. When tested on 72 distinct moods from AMG labels, the system achieved a peak performance of 45.2%.

Skowronek *et al.* developed binary classifiers for each of 12 non-exclusive mood categories using a data set of 1059 song excerpts. Using features based on temporal modulation, tempo and rhythm, chroma and key information, and occurrences of percussive sound events they trained quadratic discriminant functions for each mood, with accuracy ranging from 77% (carefree-playful) to 91% (calming-soothing), depending on the category [62].

As mentioned in the introduction, MIREX first included audio music mood classification as a task in 2007 [3]. In 2007, Tzanetakis achieved the highest percentage correct (61.5%), using only MFCC, and spectral shape, centroid, and rolloff features with an SVM classifier [63]. The highest performing system in 2008 by Peeters demonstrated some improvement (63.7%) by introducing a much larger feature corpus including, MFCCs, Spectral Crest/Spectral Flatness, as well as a variety of chroma based measurements [64]. The system uses a GMM approach to classification, but first employs Inertia Ratio Maximization with Feature Space Projection (IRMFSP) to select the most informative 40 features for each task (in this case mood), and performs Linear Discriminant Analysis (LDA) for dimensionality reduction. In 2009, Cao and Li submitted a system that was a top performer in several categories, including mood classification (65.7%) [65]. Their system employs a "super vector" of low-level acoustic features, and employs a Gaussian Super Vector followed by Support Vector Machine (GSV-SVM). It's worth noting that the best performers in each of the three years of the evaluation were general systems designed to perform multiple MIREX tasks.

6.2.2 Parametric emotion regression

Recent work in music emotion prediction from audio has suggested that parametric regression approaches can outperform labeled classifications using equivalent features. Targeting the prediction of V-A coordinates from audio, Yang *et al.* introduced the use of regression for mapping high-dimensional acoustic features to the two-dimensional space [50]. Support vector regression (SVR) [66] and a variety of ensemble boosting algorithms, including AdaBoost.RT [67], were applied to the regression problem, and one ground-truth V-A label was collected for each of 195 music clips. As this work focused primarily on labeling and regression techniques, features were extracted using publicly available extraction tools such as PsySound [68] and Marsyas [69], totaling 114 feature dimensions. To reduce the data to a tractable number of dimensions principal component analysis (PCA) was applied prior to regression. This system achieves an R^2 (coefficient of determination) score of 0.58 for arousal and 0.28 for valence.

Schmidt *et al.* and Han *et al.* each began their investigation with a quantized representation of the V-A space and employed SVMs for classification [54, 70]. Citing unsatisfactory results (with Schmidt obtaining 50.2% on a four-way classification of V-A quadrants and Han obtaining 33% accuracy in an 11-class problem), both research teams moved to regression-based approaches. Han reformulated the problem using regression, mapping the projected results into the original mood categories, employing SVR and Gaussian Mixture Model (GMM) regression methods. Using 11 quantized categories with GMM regression they obtain a peak performance of 95% correct classification.

[4] MIRtoolbox: http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

Eerola *et al.* introduced the use of a three-dimensional emotion model for labeling music; fully committing themselves to regression [15]. In their work they investigated multiple regression approaches including Partial Least-Squares (PLS) regression, an approach that considers correlation between label dimensions. They achieve R^2 performance of 0.72, 0.85, and 0.79 for valence, activity, and tension, respectively, using PLS and also report peak R^2 prediction rates for 5 basic emotion classes (angry, scary, happy, sad, and tender) as ranging from 0.58 to 0.74.

Noting that quantization by quadrant is inconsistent with the continuous nature of their collected V-A labels, Schmidt *et al.* also approached the problem using both SVR and Multiple Linear Regression (MLR). Their highest performing system obtained 13.7% average error distance in the normalized V-A space [54]. In more recent work, Schmidt *et al.* have introduced the idea of modeling the collected human response labels to music in the V-A space as a parametrized stochastic distribution, noting that the labels for most popular music segments of a reasonably small size can be well represented by a single two-dimensional Gaussian [55]. They first perform parameter estimation in order to determine the ground truth parameters, $\mathcal{N}(\mu, \Sigma)$ and then employ MLR, PLS, and SVR to develop parameter prediction models. An example regression is shown in Figure 5, which employs only spectral contrast features and MLR.

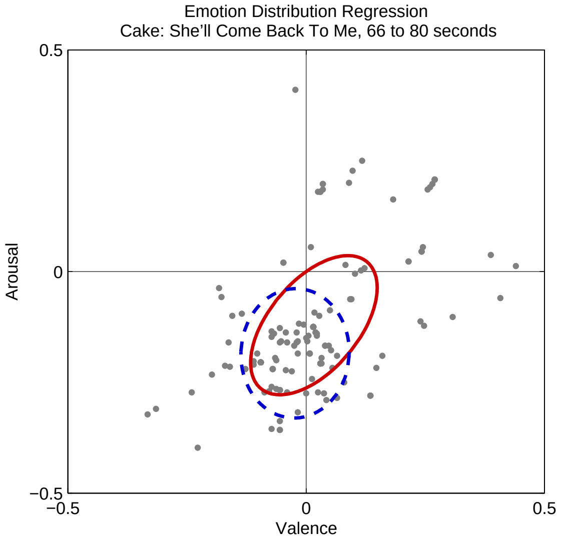

Figure 5. Collected V-A labels and distribution projections resulting from regression analysis. V-A labels: second-by-second labels per song (gray •), Σ of labels (solid red ellipse) and Σ of MLR projection from acoustic features (dashed blue ellipse) [55].

6.3 Emotion Recognition Over Time

As few other Music-IR tasks are subject to dynamic (time varying) “ground truth”, it can be argued that accounting for the time varying nature of music is perhaps more important for emotion recognition than most other tasks. Because of this variation, systems relying on a single mood label to refer to an entire song or lengthy clip are subject to high classification uncertainty. Lu *et al.* pursued mood tracking across the four principal V-A quadrants, detecting mood changes at 1 second resolution. They report precision and recall for mood boundary detection at 84.1% and 81.5%, respectively on a corpus of 9 movements from classical works [59].

Using second-by-second V-A labels collected via the *MoodSwings* game, Schmidt *et al.* also investigated tracking the emotional content of music over time [54, 55]. Their time-varying analyses remain formulated as a regression to develop a mapping from short-time high-dimensional acoustic features to time-localized emotion space coordinates. Figure 6 shows the V-A projections of a song clip obtained from spectral contrast features and MLR prediction [54]. In this example, a 15-second clip has been broken down into three five-second segments (projections) demonstrating the overall movement in the V-A space. A version of their time-varying regression system is implemented back into *MoodSwings* as a simulated “AI” partner for single-player games.

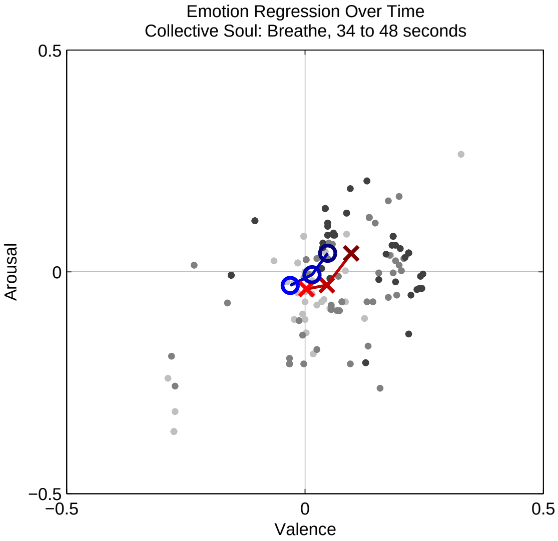

Figure 6. V-A labels and projections over time for a 15s segment (markers become darker over time): second-by-second labels per song (gray •), mean of the collected labels over 5-second intervals (red X), and projection from acoustic features in 5-second intervals (blue O) [54].

7. COMBINING MULTIPLE FEATURE DOMAINS

It is clear that some aspects of music data (e.g., social factors, such as “Christmas music”, or songs considered to be “one hit wonders”) are not revealed in the audio, and results using only acoustic (spectral) features for Music-IR tasks have led many to believe there is an upper limit to performance using these features alone [71]. This has led to a growing amount of research towards combining multiple feature domains to improve recognition in Music-IR systems. The earliest attempts at classification of emotion using multi-modal approaches (combining features from disparate domains) were applied to speech, using combined analysis of audio and facial expressions [72–74]. These methods have inspired other multi-modal approaches to several other Music-IR classification tasks, such as genre recognition, but such combined approaches to classifying the emotional content of music have emerged only within the past few years.

7.1 Combining Audio & Lyrics

In spite of the importance of audio, some musical genres (e.g., "Christmas songs") are much easier to detect using text. This was the motivation for Neumayer's work utilizing multiple combinations of audio and textual features for musical genre classification [75]. Similarly, many of the studies described below are motivated by the idea that some emotions conveyed by music are better detected using a combination of audio and lyrics. Some systems report relatively modest performance gains, but it is often in tasks where baseline performance using audio features alone has been high. The most compelling results show, in certain cases, improvement over audio features alone, demonstrating that information contained within the two feature modalities can be highly complementary.

7.1.1 Systems combining audio & lyrics

The first system to employ both audio and lyrics for emotion classification (D. Yang and Lee) used lyric text and a wide range of audio features, such as beats per minute and 12 low-level MPEG-7 descriptors (e.g., spectral centroid, rolloff, and flux), on a set of 145 30-second song clips [76]. Each clip was hand-labeled with one of 11 emotional categories, based on PANAS labels [19, 20]. While strong correlations were found between particular lyrics and emotional categories (hostility, sadness, and guilt), the addition of lyrics increased classification performance by only 2.1% (82.8% vs. 80.7% using only audio) on their relatively small data set.

A more recent system (Y. Yang *et al.*) combined lyrics and audio features using a database of 1240 Chinese pop songs [50]. The songs were hand-labeled according to one of the four valence-arousal quadrants of the Russell-Thayer model, and acoustic features (MFCC and spectral descriptors) were computed for a 30-second segment drawn from the middle of each song. Lyrics (assumably from the entire song) were text-analyzed using the BOW approach. This work compared three different methods of combining audio and text features, obtaining the best results by first using audio and text separately to classify arousal and valence, respectively, using SVMs and then merging the results to determine a full V-A classification. On an 80/20 training-testing split of the data with 1000-fold cross-validation, acoustic features alone resulted in a baseline of 46.6% correct classification, while combining audio and lyrics yielded 57.1% accuracy (a relative performance increase of 21%).

Other recent work by Laurier, Grivolla, and Herrera explored the use of audio and multiple lyrics features to classify emotions of songs in the four-quadrant V-A space [77]. Songs were labeled using Last.fm tags and filtered using the lexical database WordNet-Affect to remove synonyms (with subsequent validation by human listeners). The corpus consisted of 1000 songs, with equal distribution across the four quadrants, and a combination of timbral, rhythmic, tonal, and temporal features were computed for each song. Three different methods, lyric similarity, LSA, and Language Model Differences (LMD) were investigated for deriving lyrics features. LMD compares the difference in frequency of terms between the language models of various mood categories and is used here to select the 100 most discriminative terms, resulting in significantly higher classification performance vs. LSA. The audio and text features were combined into a single vector for joint classification, improving performance over audio features alone in two quadrants by 5%: "happy" (81.5% to 86.8%) and "sad" (87.7% to 92.8%). The other quadrants were essentially unchanged, but already had high classification accuracy using only audio features: "relaxed" (91.4% to 91.7%) and "angry" (98.1% to 98.3%).

Hu *et al.* also combined audio and lyrics for emotion recognition on a relatively large database of nearly 3000 songs [78]. Eighteen emotion classes were formed based on social tags from Last.fm, using WordNet-Affect to remove tags irrelevant to emotion recognition and significant refinement by human judges. For lyrics features, this system uses a BOW approach with TF-IDF weighting. The BOW stemmed (prefixes and suffixes removed), TF-IDF weighted (BSTI) features are directly concatenated with 63 spectrum-derived audio features for training an SVM classifier. BSTI features were selected using different methods, including the LMD selection method (described in the previous system [77]), varying the number of feature dimensions to identify optimal performance. Interestingly, using BSTI (lyrics) features alone outperforms audio alone for 12 of the 18 mood classes. Audio + lyric approaches demonstrate the highest performance in recognizing 13 of the 18 mood classes (audio alone is highest for 3 classes, "happy", "upbeat", and "desire", while lyrics alone perform best for "grief" and "exciting"). Audio with LMD feature selection (63 dimensions, equivalent to the number of audio features) performed the highest in 5 categories ("calm", "sad", "anger", "confident", and "earnest") and improved performance in half of the cases where audio alone outperforms lyrics and vice versa, demonstrating the utility of combining features.

Recently, Schuller *et al.* investigated using audio features, lyrics, and metadata to automatically label music in a discretized version of the V-A space [79]. For a database of 2648 pop songs, each was rated by four listeners who selected one of 5 discrete labels. Their classification task, is ultimately reduced to two independent three-class problems. Their best performing system made use of feature selection algorithms and label filtering, achieving 64.1% and 60.9% on valence and arousal, respectively.

7.2 Combining Audio & Tags

Music-IR researchers have also focused on multimodal approaches incorporating tags and low-level audio features for classification tasks. Turnbull *et al.* explore the problem of tag classification by combining semantic information from web documents, social tags and audio analysis on the CAL500 data set [80]. They compare a number of algorithms (e.g., Calibrated Score Averaging, RankBoost, Kernel Combination SVM) and find that multimodal approaches significantly outperform unimodal approaches.

Bischoff *et al.* combine social tag information and audio content-based analysis specifically for the task of emotion recognition [41]. For each of 4,737 songs, they collect social tags from Last.fm and generate a 240-dimensional audio feature vectors (including MFCCs, chroma features, and other spectral features). They then train a naive Bayes classifier for the social tags and an SVM for the audio feature vectors, combining them using a simple weighted combination approach. In one experiment, this approach is used to predict one of the five MIREX mood clusters [3]. In the second experiment, the approach is used to predict one of the four quadrants of the V-A mood space. Ground truth for each track is based on a manually-determined (ad-hoc) mapping between one of 178 mood tags to a MIREX mood cluster and to a V-A quadrant. In both experiments, the multimodal approach demonstrates better performance than either the tag-based and audio-based approaches alone.

7.3 Combining Audio & Images

Analysis of audio features of music in combination with associated images (album covers, artist photos, etc.) for Music-IR tasks has very recently sparked research interest. Dunker *et al.* investigated methods to combine music and image domains for mood classification [81]. Their work investigates both classifying audio paired with images in a multi-modal media tagging approach as well as trying to pair audio and images that have been tagged separately. More recently, Lībeks and Turnbull analyze promotional photographs of artists using content-based image annotation [21]. Although they label these artists with genre tags (provided by Last.fm), it would be straightforward to adapt their approach to use emotion tags.

8. CONCLUSION

Recognizing musical mood remains a challenging problem primarily due to the inherent ambiguities of human emotions. Though research on this topic is not as mature as some other Music-IR tasks, it is clear that rapid progress is being made. In the past 5 years, the performance of automated systems for music emotion recognition using a wide range of annotated and content-based features (and multimodal feature combinations) have advanced significantly. As with many Music-IR tasks open problems remain at all levels, from emotional representations and annotation methods to feature selection and machine learning.

While significant advances have been made, the most accurate systems thus far achieve predictions through large-scale machine learning algorithms operating on vast feature sets, sometimes spanning multiple domains, applied to relatively short musical selections. Oftentimes, this approach reveals little in terms of the underlying forces driving the perception of musical emotion (e.g., varying contributions of features) and, in particular, how emotions in music change over time. In the future, we anticipate further collaborations between Music-IR researchers, psychologists, and neuroscientists, which may lead to a greater understanding of not only mood within music, but human emotions in general. Furthermore, it is clear that individuals perceive emotions within music differently. Given the multiple existing approaches for modeling the ambiguities of musical mood, a truly personalized system would likely need to incorporate some level of individual profiling to adjust its predictions.

This paper has provided a broad survey of the state of the art, highlighting many promising directions for further research. As attention to this problem increases, it is our hope that the progress of this research will continue to accelerate in the near future.

9. REFERENCES

[1] C. C. Pratt, *Music as the language of emotion.* The Library of Congress, December 1950.

[2] J. S. Downie, “The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research,” *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[3] X. Hu, J. Downie, C. Laurier, M. Bay, and A. Ehmann, “The 2007 MIREX audio mood classification task: Lessons learned,” in *Proc. of the Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, 2008.

[4] J. A. Russell, “A circumspect model of affect,” *Journal of Psychology and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[5] P. Juslin and P. Luakka, “Expression, perception, and induction of musical emotions: A review and questionnaire study of everyday listening,” *Journal of New Music Research*, vol. 33, no. 3, p. 217, 2004.

[6] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology.* MIT Press, 1974.

[7] C. McKay, “Emotion and music: Inherent responses and the importance of empirical cross-cultural research,” Course Paper. McGill University, 2002.

[8] L.-L. Balkwill and W. F. Thompson, “A cross-cultural investigation of the perception of emotion in music,” *Music Perception*, vol. 17, no. 1, pp. 43–64, 1999.

[9] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, “Universal recognition of three basic emotions in music,” *Current Biology*, vol. 19, no. 7, pp. 573 – 576, 2009.

[10] K. Hevner, “Experimental studies of the elements of expression in music,” *American Journal of Psychology*, vol. 48, no. 2, pp. 246–267, 1936.

[11] P. Juslin and J. Sloboda, *Music and Emotion: theory and research.* Oxford Univ. Press, 2001.

[12] E. Schubert, “Update of the Hevner adjective checklist,” *Perceptual and Motor Skills*, vol. 96, pp. 1117–1122, 2003.

[13] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement." *Emotion*, vol. 8, p. 494, 2008.

[14] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representation from social tags," in *Proc. of the Intl. Society for Music Information Conf.*, Kobe, Japan, 2009.

[15] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. of the Intl. Society for Music Information Conf.*, Kobe, Japan, 2009.

[16] L. Mion and G. D. Poli, "Score-independent audio features for description of music expression," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 458–466, 2008.

[17] E. Bigand, "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts," *Cognition and Emotion*, vol. 19, no. 8, p. 1113, 2005.

[18] U. Schimmack and R. Reisenzein, "Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation," *Emotion*, vol. 2, no. 4, p. 412, 2002.

[19] D. Watson and L. Clark, *PANAS-X: Manual for the Positive and Negative Affect Schedule*, expanded form ed., University of Iowa, 1994.

[20] A. Tellegen, D. Watson, and L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999.

[21] J. Lībeks and D. Turnbull, "Exploring artist image using content-based analysis of promotional photos," in *Proc. of the Int. Computer Music Conf.*, 2010.

[22] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, 2008.

[23] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel classification of music into emotion," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, 2008.

[24] F. Miller, M. Stiksel, and R. Jones, "Last.fm in numbers," *Last.fm press material*, February 2008.

[25] P. Lamere and O. Celma, "Music recommendation tutorial notes," ISMIR Tutorial, September 2007.

[26] L. von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006.

[27] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Vienna, Austria, 2007, pp. 365–366.

[28] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet, "A game-based approach for collecting semantic annotations of music," in *Proc. Intl. Conf. on Music Information Retrieval*, Vienna, Austria, 2007, pp. 535–538.

[29] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford, "TagATune: A game for music and sound annotation," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Vienna, Austria, 2007.

[30] Y. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, September 2008.

[31] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, "Improving music emotion labeling using human computation," in *HCOMP 2010: Proc. of the ACM SIGKDD Workshop on Human Computation*, Washington, D.C., 2010.

[32] J. P. Bachorik, M. Bangert, P. Loui, K. Larke, J. Berger, R. Rowe, and G. Schlaug, "Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli," *Music Perception*, vol. 26, no. 4, pp. 355–364, April 2009.

[33] L. Barrington, D. Turnbull, D. O'Malley, and G. Lanckriet, "User-centered design of a social game to tag music," *ACM KDD Workshop on Human Computation*, 2009.

[34] P. Knees, T. Pohle, M. Schedl, D. Schnitzer, and K. Seyerlehner, *A Document-Centered Approach to a Natural Language Music Search Engine*. Springer Berlin / Heidelberg, 2008, pp. 627–631.

[35] O. Celma, P. Cano, and P. Herrera, "Search sounds: An audio crawler focused on weblogs," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Victoria, Canada, 2006.

[36] B. Whitman and D. Ellis, "Automatic record reviews," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 470–477.

[37] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Vienna, Austria, 2007.

[38] P. Knees, T. Pohle, M. Schedl, and G. Widmer, "A music search engine built upon audio-based and web-based similarity measures," in *ACM SIGIR*, 2007.

[39] B. Whitman, "Combining musical and cultural features for intelligent style detection," in *Proc. of the Intl. Conf. on Music Information Retrieval*, Paris, France, 2002, pp. 47–52.

[40] P. Knees, E. Pampalk, and G. Widmer, "Artist classification with web-based data," in *Proc. of the Intl. Symposium on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 517–524.

[41] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," in *Proc. of the Intl. Society for Music Information Retrieval Conf.*, Kobe, Japan, 2009.

[42] K. Bischoff, C. Firan, W. Nejdl, and R. Paiu, "How do you feel about 'Dancing Queen'?: Deriving mood & theme annotations from user tags," in *Proc. of the ACM/IEEE-CS Joint Conf. on Digital Libraries*, Austin, TX, Jan 2009.

[43] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[44] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *Proc. IEEE Intl. Conf. on Multimedia and Expo*, Taipei, Taiwan, June 27-30 2004.

[45] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW)," The NIMH Center for the Study of Emotion and Attention,University of Florida, Tech. Rep., 1999.

[46] Y. Hu, X. Chen, and D. Yang, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method," in *Proc. of the Intl. Society for Music Information Conf.*, Kobe, Japan, 2009.

[47] R. H. Chen, Z. L. Xu, Z. X. Zhang, and F. Z. Luo, "Content based music emotion analysis and recognition," in *Proc. of the Intl. Workshop on Computer Music and Audio Technology*, 2006.

[48] Y. Xia, L. Wang, K. Wong, and M. Xu, “Sentiment vector space model for lyric-based song sentiment classification,” in *Proc. of the Association for Computational Linguistics*. Columbus, Ohio, U.S.A: ACL-08, 2008, pp. 133–136.

[49] O. C. Meyers, “A mood-based music classification and exploration system,” Master’s thesis, Massachusetts Institute of Technology, June 2007.

[50] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. Chen, *Advances in Multimedia Information Processing - PCM 2008*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, December 2008, ch. 8, pp. 70–79.

[51] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357– 366, 1980.

[52] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in *Proc. of the Intl. Symposium on Music Information Retrieval*, Plymouth, MA, September 2000.

[53] K. F. MacDorman, S. Ough, and C.-C. Ho, “Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison,” *Journal of New Music Research*, vol. 36, no. 4, pp. 281–299, 2007.

[54] E. M. Schmidt, D. Turnbull, and Y. E. Kim, “Feature selection for content-based, time-varying musical emotion regression,” in *MIR ’10: Proc. of the Intl. Conf. on Multimedia Information Retrieval*, Philadelphia, PA, 2010, pp. 267–274.

[55] E. M. Schmidt and Y. E. Kim, “Prediction of time-varying musical mood distributions from audio,” in *Proc. of the Int. Society for Music Information Conf.*, Utrecht, Netherlands, August 2010.

[56] T. Fujishima, “Realtime chord recognition of musical sound: A system using Common Lisp music.” in *Proc. of the Intl. Computer Music Conf.*, 1999.

[57] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293– 302, 2002.

[58] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai, “Music type classification by spectral contrast feature,” in *Proc. Intl. Conf. on Multimedia and Expo*, vol. 1, 2002, pp. 113–116.

[59] L. Lu, D. Liu, and H. J. Zhang, “Automatic mood detection and tracking of music audio signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[60] T. Li and M. Ogihara, “Detecting emotion in music,” in *Proc. of the Intl. Conf. on Music Information Retrieval*, Baltimore, MD, October 2003.

[61] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, “Support vector machine active learning for music retrieval,” *Multimedia Systems*, vol. 12, no. 1, pp. 3–13, Aug 2006.

[62] J. Skowronek, M. McKinney, and S. van de Par, “A demonstrator for automatic music mood estimation,” in *Proc. Intl. Conf. on Music Information Retrieval*, Vienna, Austria, 2007.

[63] G. Tzanetakis, “Marsyas submissions to MIREX 2007,” MIREX 2007.

[64] G. Peeters, “A generic training and classification system for MIREX08 classification tasks: Audio music mood, audio genre, audio artist and audio tag,” MIREX 2008.

[65] C. Cao and M. Li, “Thinkit’s submissions for MIREX2009 audio music classification and similarity tasks.” ISMIR, MIREX 2009.

[66] A. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[67] D. Shrestha and D. Solomatine, “Experiments with AdaBoost.RT, an improved boosting scheme for regression,” *Neural Computation*, vol. 18, no. 7, pp. 1678–1710, 2006.

[68] D. Cabrera, S. Ferguson, and E. Schubert, “Psysound3: software for acoustical and psychoacoustical analysis of sound recordings,” in *Proc. of the Intl. Conf. on Auditory Display*, Montreal, Canada, June 26-29 2007, pp. 356–363.

[69] G. Tzanetakis and P. Cook, “Marsyas: a framework for audio analysis,” *Organized Sound*, vol. 4, no. 3, pp. 169–175, 1999.

[70] B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, “SMERS: Music emotion recognition using support vector regression,” in *Proc. of the Intl. Society for Music Information Conf.*, Kobe, Japan, 2009.

[71] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky?” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[72] J. Cohn and G. Katz, “Bimodal expression of emotion by face and voice,” in *ACM Intl. Multimedia Conf.*, 1998.

[73] L. Silva and P. Ng, “Bimodal emotion recognition,” in *Proc. of the IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 332–335.

[74] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, 2009, pp. 39–58.

[75] R. Neumayer and A. Rauber, “Integration of text and audio features for genre classification in music information retrieval,” in *Proc. of the European Conf. on Information Retrieval*, 2007.

[76] D. Yang and W. Lee, “Disambiguating music emotion using software agents,” in *Proc. of the Intl. Conf. on Music Information Retrieval*. Barcelona, Spain: Universitat Pompeu Fabra, October 2004.

[77] C. Laurier, J. Grivolla, and P. Herrera, “Multimodal music mood classification using audio and lyrics,” in *Proc. of the Intl. Conf. on Machine Learning and Applications*. Universitat Pompeu Fabra, 2008, pp. 1–6.

[78] X. Hu, J. S. Downie, and A. F. Ehmann, “Lyric text mining in music mood classification,” in *Proc. of the Intl. Society for Music Information Retrieval Conf.*, Kobe, Japan, 2009.

[79] B. Schuller, J. Dorfner, and G. Rigoll, “Determination of nonprototypical valence and arousal in popular music: Features and performances,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–20, Jan 2010.

[80] D. Turnbull, L. Barrington, M.Yazdani, and G. Lanckriet, “Combining audio content and social context for semantic music discovery,” *ACM SIGIR*, 2009.

[81] P. Dunker, S. Nowak, A. Begau, and C. Lanz, “Content-based mood classification for photos and music: A generic multimodal classification framework and evaluation approach,” in *Proc. of the Intl. Conf. on Multimedia Information Retrieval*. ACM, 2008.