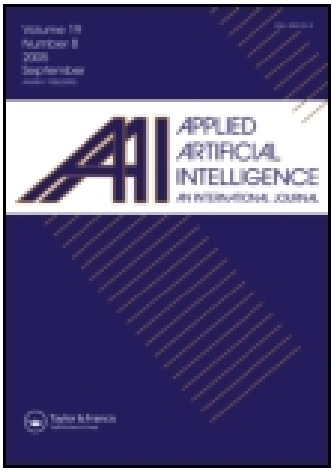

Applied Artificial Intelligence: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uaai20

Music Emotion Recognition with Standard and Melodic Audio Features

Renato Panda[a], Bruno Rocha[a] & Rui Pedro Paiva[a]

[a] CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

Published online: 18 Apr 2015.

To cite this article: Renato Panda, Bruno Rocha & Rui Pedro Paiva (2015) Music Emotion Recognition with Standard and Melodic Audio Features, Applied Artificial Intelligence: An International Journal, 29:4, 313-334, DOI: 10.1080/08839514.2015.1016389

To link to this article: http://dx.doi.org/10.1080/08839514.2015.1016389

Applied Artificial Intelligence, 29:313–334, 2015

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839514.2015.1016389

Taylor & Francis
Taylor & Francis Group

MUSIC EMOTION RECOGNITION WITH STANDARD AND MELODIC AUDIO FEATURES

Renato Panda, Bruno Rocha, and Rui Pedro Paiva
CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

□ *We propose a novel approach to music emotion recognition by combining standard and melodic features extracted directly from audio. To this end, a new audio dataset organized similarly to the one used in MIREX mood task comparison was created. From the data, 253 standard and 98 melodic features are extracted and used with several supervised learning techniques. Results show that, generally, melodic features perform better than standard audio. The best result, 64% f-measure, with only 11 features (9 melodic and 2 standard), was obtained with ReliefF feature selection and Support Vector Machines.*

INTRODUCTION

Since the beginning of mankind, music has been present in our lives, serving a myriad of both social and individual purposes. Music is used in fields as diverse as religion, sports, entertainment, health care, and even war, conveying emotions and perceptions to the listener, which vary among cultures and civilizations.

As a result of technological innovations in this digital era, a tremendous impulse has been given to the music distribution industry. Factors such as widespread access to the Internet and the generalized use of compact audio formats such as mp3 have contributed to that expansion. The frenetic growth in music supply and demand uncovered the need for more powerful methods for automatic retrieval of relevant songs in a given context from extensive databases.

Digital music repositories need, then, more advanced, flexible, and user-friendly search mechanisms, adapted to the requirements of individual users. In fact, "music's preeminent functions are social and psychological,"

Address correspondence to Renato Panda or Rui Pedro Paiva, CISUC, Department of Informatics Engineering, University of Coimbra, DEI, Polo 2, Pinhal de Marrocos, 3030-290 Coimbra, Portugal. E-mail: panda@dei.uc.pt; ruipedro@dei.uc.pt

and so "the most useful retrieval indexes are those that facilitate searching in conformity with such social and psycho-logical functions. Typically, such indexes will focus on stylistic, mood, and similarity information" (Huron 2000; p. 1). This is supported by studies on music information behavior that have identified emotional content of music as an important criterion for music retrieval and organization (Hevner 1936).

Research devoted to emotion analysis is relatively recent, although it has received increasing attention in recent years. Hence, many limitations can be found and several problems are still open. In fact, the present accuracy of those systems shows that there is still room for improvement. In the last Music Information Retrieval Evaluation eXchange (MIREX; an annual evaluation campaign for Music Information Retrieval [MIR] algorithms, coupled to the International Society for Music Information Retrieval [ISMIR] and its annual ISMIR conference), the best algorithm in the Music Mood Task achieved 67.8% accuracy in a secret dataset organized into five emotion categories.

Therefore, in this article we propose a system for music emotion recognition (MER) in audio, combining both standard and melodic audio features. To this date, most approaches based only on standard audio features, such as that followed in the past by our team (Panda and Paiva 2012), seem to have attained a so-called glass ceiling. We believe that combining the current features as well as melodic features directly extracted from audio, which were already successfully used in genre recognition (Salamon, Rocha, and Gómez 2012), might help improve current results.

The system is evaluated using a dataset proposed by our team (Panda and Paiva 2012) made of 903 audio clips, following the same organization of that used in the MIREX Mood Classification Task (i.e., five emotion clusters). We evaluate our approach with several supervised learning and feature selection strategies. Among these, best results were attained with an SVM classifier: 64% F-measure in the set of 903 audio clips, using a combination of both standard and melodic audio features.

We believe this work offers a number of relevant contributions to the MIR/MER research community:

- A MIREX-like audio dataset (903 samples);
- A methodology for automatic emotion data acquisition, resorting to the AllMusic[1] platform, an influential website and API providing 289 mood labels that are applied to songs and albums;
- A methodology for MER, combining different types of audio features, capable of significantly improving the results attained with standard audio features only.

[1] http://www.allmusic.com/moods.

To the best of our knowledge, this is the first study using melodic audio features in music emotion recognition.

The article is organized as follows. In "Literature Review," an overview of the work related to the subject is presented. "Methods" describes the methodology used in our MER approach. In "Experimental Results," the experimental results are analyzed and discussed. Conclusions and plans for future work are described in the final section.

LITERATURE REVIEW

Emotion Taxonomies

For a very long time, emotions have been a major subject of study by psychologists. Emotions are subjective and their perception varies from person to person and also across cultures. Furthermore, usually there are many different words describing them; some are direct synonyms whereas others represent small variations. Different persons have different perceptions of the same stimulus and often use some of these different words to describe similar experiences. Unfortunately, there is not one standard, widely accepted, classification model for emotions.

Several theoretical models have been proposed over the last century by authors in the psychology field. These models can be grouped into two major approaches: categorical models and dimensional models of emotion. This article is focused on categorical models. A brief overview of such models is presented in the following paragraphs.

Categorical models, also known as discrete models, classify emotions by using single words or groups of words to describe them (e.g., happy, sad, anxious). Dimensional models consist of a multidimensional space, mapping different emotional states to locations in that space. Some of the most adopted approaches use two dimensions, normally arousal, energy, or activity against valence or stress, forming four quadrants corresponding to distinct emotions (Russell 1980; Thayer 1989).

An example of this approach is the emotion model that can be derived from the basic emotions—anger, disgust, fear, happiness, sadness, and surprise—identified by Ekman (1992). These emotions are considered the basis on which all the other emotions are built. From a biological perspective, this idea is manifested in the belief that there might be neurophysiological and anatomical substrates corresponding to the basic emotions. From a psychological perspective, basic emotions are often held to be the primitive building blocks of other, nonbasic emotions. This notion of basic emotions is, however, questioned in other studies (Ortony and Turner 1990). Even so, the idea is frequently adopted in MER research, most likely due to the use of

specific words, offering integrity across different studies, and their frequent use in the neuroscience in relation to physiological responses (Ekman 1992).

Another widely known discrete model is Hevner's adjective circle (1936). Hevner, best known for research in music psychology, concluded that music and emotions are intimately connected, with music always carrying emotional meaning. As a result, the author proposed a grouped list of adjectives (emotions), instead of using single words. Hevner's list is composed of 67 different adjectives, organized into eight different groups in a circular way. These groups, or clusters, contain adjectives with similar meaning, used to describe the same emotional state.

In this article we use a categorical model of emotions following the organization employed in the MIREX Mood Classification Task,[2] an annual comparison of state-of-the-art MER approaches held in conjunction with the ISMIR conference.[3] This model classifies emotions into five distinct groups or clusters, each containing the following list of 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;

However, as will be discussed, the MIREX taxonomy is not supported by psychological studies.

Musical Features

Research on the relationship between music and emotion has a long history, with initial empirical studies starting in the 19th century (Gabrielsson and Lindström 2001). This problem was studied more actively in the 20th century, when several researchers investigated the relationship between emotions and particular musical attributes. As a result, a few interesting discoveries were made, for example, major modes are frequently related to emotional states such as happiness or solemnity, whereas minor modes are associated with sadness or anger (Laurier et al. 2009). Moreover, simple, consonant harmonies are usually happy, pleasant, or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension, or sadness, because they create instability in a musical piece (Laurier et al. 2009).

[2] http://www.music-ir.org/mirex/wiki/2013:Audio_Classification_%28Train/Test%29_Tasks#Audio_Mood_Classification.

[3] http://www.ismir.net/.

In a 2008 overview, Friberg (2008) described the following musical features as related to emotion: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, and rhythm. Other common features not included in that list are, for example, mode, loudness, vibrato, or musical form (Laurier et al. 2009; Laurier 2011). Many of these attributes were also identified as relevant by Meyers (2007): mode, harmony, tempo, rhythm, dynamics, and musical form. Several of these features have already been studied in the Musical Instrument Digital Interface (MIDI) domain (e.g., Cataltepe, Tsuchihashi, and Katayose 2007).

The following list contains many of the relevant features for music emotion analysis:

- Timing: tempo, tempo variation, duration contrast
- Dynamics: overall level, crescendo/decrescendo, accents
- Articulation: overall (staccato/legato), variability
- Timbre: spectral richness, onset velocity, harmonic richness
- Pitch (high/low)
- Interval (small/large)
- Melody: range (small/large), direction (up/down)
- Harmony (consonant/complex-dissonant)
- Tonality (chromatic-atonal/key-oriented)
- Rhythm (regular-smooth/firm/flowing-fluent/irregular-rough)
- Mode (major/minor)
- Loudness (high/low)
- Musical form (complexity, repetition, new ideas, disruption)
- Vibrato (extent, speed)

However, many of these listed features are often difficult to extract from audio signals. Also, several of them require further study from a psychological perspective. Schubert studied some of the interactions between such features and the emotional responses in the Russell's model of emotion (Schubert 1999). As a result, he identified some interesting nondirect relations between the variation of feature values and arousal-valence results, as well as hypothesized interactions among features, resulting in different emotional states. As an example, for minor modes, increasing tempo leads to increasing arousal and unchanged valence, whereas for major modes, increasing tempo leads also to increasing valence (Schubert 1999). The author concludes that "there are underlying principles which govern the relationship between musical features and emotional response. It is likely that the rules are bound by cultural norms, but whether the relationships be local or universal, a mass of relationships awaits discovery" (Schubert 1999; p. 391).

Previous MER Works

To the best of our knowledge, the first MER article was published in 1988 by Katayose, Imai, and Inokuchi (1988). There, a system for sentiment analysis based on audio features from polyphonic recordings of piano music was proposed. Music primitives such as melody, chords, key, and rhythm features were used to estimate the emotion with heuristic rules.

Long after 1988, a long period without active research in the field, Feng, Zhuang, and Pan (2003) proposed a system for emotion detection in music, using only features of tempo and articulation in a music piece to identify emotions. The used categorical model comprises only four emotions: happiness, sadness, anger, and fear (basic emotions). The classification is then performed using a neural network with three layers. Although results were high, between 75% and 86%, the test collection was very limited with only three songs in the fear category.

In the same year, Li and Ogihara (2003) studied the problem of emotion detection as a multi-label classification system, thus admitting that music can have more than one emotion. The musical database was composed of 499 songs, 50% used for training and the remaining 50% for testing. From these, acoustic features such as timbral texture, rhythm content (beat and tempo detection) and pitch content were extracted and classification was performed with Support Vector Machines (SVM), resulting in an F-measure of 44.9% (micro average) and 40.6% (macro average). One of the major problems with the article is related to the dataset, in which a single subject was used to classify the songs.

Still in 2003, Liu and Lu (2003) studied hierarchical versus nonhierarchical approaches to emotion detection in classical music. The used algorithms that rely on features such as root mean square value in each sub-band, spectral shape features such as centroid, rolloff, and spectral flux, and a Canny estimator used to detect beat. The results are apparently very good, with accuracy reaching values from 76.6% to 94.5% for the hierarchical framework, with the nonhierarchical reaching 64.7% to 94.2%. However, it is important to note that only classical music was used and only four possible emotions were considered.

In the next years, other researchers proposed interesting approaches. Li and Ogihara (2004) built on their previous system to study emotion detection and similarity search in jazz. Yang and Lee (2004) proposed a strategy for emotion rating to assist human annotators in the music emotion annotation process, using acoustic data to extract emotion intensity information, but also using song lyrics to distinguish among emotions by assessing valence.

In 2005, Carvalho and Chao (2005) studied the impact caused by the granularity of the emotional model and different classifiers in the emotion

detection problem. The results showed that the model granularity, using a binary problem against a more "fine-grained" problem (of five labels), has a much higher impact on performance (13.5% against 63.45% in error rate) than the used classifiers and learning algorithms, which made the results vary only within 63.45% and 67%.

In 2006, Lu, Liu, and Zhang (2006) proposed an approach for emotion detection on acoustic music data, building on their previous work (Liu and Lu 2003). Two distinct approaches were selected: hierarchical, organized in three layers similar to the feature groups, and nonhierarchical. Both frameworks classify music sets based on the following feature sets: (1) intensity (energy in each sub-band); (2) timbre, composed by mel-frequency cepstrum coefficients (MFCC), spectral shape features, and spectral contrast features; and (3) rhythm (rhythm strength, rhythm regularity, and tempo). The results showed an average precision on emotion detection of 86.3%, with average recall of 84.1%. Although the results were high, it is important to note that clips were classified in only four different categories and all the clips were classical music. One of the most interesting aspects of this study is the chosen feature sets.

Meyers (2007) proposed a tool to automatically generate playlists based on a desired emotion, using audio information extracted from songs' audio signals and lyrics.

More recently, Wang et al. (2010) proposed an audio classification system, in which posterior weighted Bernoulli mixture model (PWBMM; Wang, Lo, and Jeng 2010) is applied to each song's feature vectors (made of 70 features from the MIR Toolbox), transforming them into a semantic representation based on music tags. For each emotion class, a set of semantic representations of songs is used to train an ensemble classifier (SVM was used). Although the system was initially developed for music genre classification, it obtained the top score in the MIREX 2010 Mood Classification Task with 64.17%.[4]

McVicar and Freeman (2011) proposed a bimodal approach, combining the study of audio and lyrics to identify common characteristics between them. This strategy is founded on the authors' assumption that "the intended mood of a song will inspire the songwriter to use certain timbres, harmony, and rhythmic features, in turn affecting the choice of lyrics as well" (McVicar and Freeman 2011; p. 783). Using this method, the Pearson's correlation coefficient between each of the Arousal and Valence (AV) values of the audio features and lyrics were computed, finding many of the correlations to be statistically significant but below 0.2 in absolute value. Other bimodal approaches were also proposed recently (e.g., Yang et al. 2008; Hu and Downie 2010).

[4]http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results.

In 2012, our team presented an approach to categorical emotion classification in audio music (Panda and Paiva 2012). To this end, a freely available MIREX-like audio dataset based on the AllMusic database was created. Three frameworks—PsySound3, Marsyas, and MIR Toolbox—were used to extract audio features. A top result of 47.2% F-measure was attained using SVMs and feature selection.

Also in 2012, Song, Dixon, and Pearce (2012) evaluated the influence of musical features for emotion classification. To this end, 2904 clips tagged with one of the four words, "happy," "sad," "angry," or "relaxed," on the Last.FM website were used to extract features, and SVMs were the selected classifier. Some of the interesting results were that spectral features outperformed those based on rhythm, dynamics, and, to a lesser extent, harmony. The use of an SVM polynomial kernel led to better results and the fusion of different feature sets did not always lead to improved classification.

In addition to categorical solutions, other researchers have tackled MER based on the dimensional perspective. One of the most notable works was carried out by Yang and colleagues (2008). There, they propose a solution to MER in the continuous space following a regression approach using 114 audio features. Best results, measured using R^2 statistics, were obtained with support vector regression, attaining 58.3% for arousal and 28.1% for valence.

Some studies that followed have significantly improved the results using both standard and melodic audio features from the same dataset, attaining 67.8% arousal and 40.6% valence accuracy (Panda, Rocha, and Paiva 2013).

METHODS

Dataset Acquisition

To create the dataset, we built on the AllMusic knowledge base, organizing it in a similar way to the MIREX Mood Classification Task testbed. It contains the same five clusters with the same several emotional categories each as those mentioned in the "Introduction."

The MIREX taxonomy is employed because this is the only base of comparison generally accepted by the MER community. Although the MIREX campaign helps in comparing different state-of-the-art approaches, the datasets are not publicly available. Thus, we try to mimic the referred dataset, providing a public dataset that can be freely used to compare results outside of the MIREX campaign. To this end, we chose the AllMusic database because, unlike other popular databases such as Last.FM, annotations are performed by professionals instead of a large community of music listeners (as happens in Last.FM). Therefore, those annotations are likely more

reliable. However, the annotation process is not made public and, hence, we cannot critically analyze it.

The first step for acquiring the dataset consisted in accessing automatically the AllMusic API to obtain a list of songs with the MIREX mood tags and other meta-information, such as song identifier, artists, and title. To this end, a script was created to fetch existing audio samples from the same site, most being 30-second mp3 files.

The next step was to create the emotion annotations. To do so, the songs containing the same mood tags present in the MIREX clusters were selected. Because each song may have more than one tag, the tags of each song were grouped by cluster and the resulting song annotation was based on the most significant cluster (i.e., the one with more tags; for instance, a song with one tag from cluster 1 and three tags from cluster 5 is marked as cluster 5). A total of 903 MIREX-like audio clips, nearly balanced across clusters, were acquired.

Although ours and the original MIREX Mood Task dataset have similarities in organization, they still differ in important aspects such as the annotation process, and results must be analyzed or compared with this in mind. In the case of the MIREX mood dataset, songs were labeled based on the agreement among three experts (Hu et al. 2008). AllMusic songs are annotated by experts, but few details are provided about the process, which does not allow for a critical analysis of the annotation process.

Our proposed dataset is relatively balanced among clusters, with a slightly higher representation for clusters 3 and 4, as shown in Figure 1. Another relevant aspect of the dataset is that, as pointed out in a few studies, there is a semantic overlap (ambiguity) between clusters 2 and 4, and

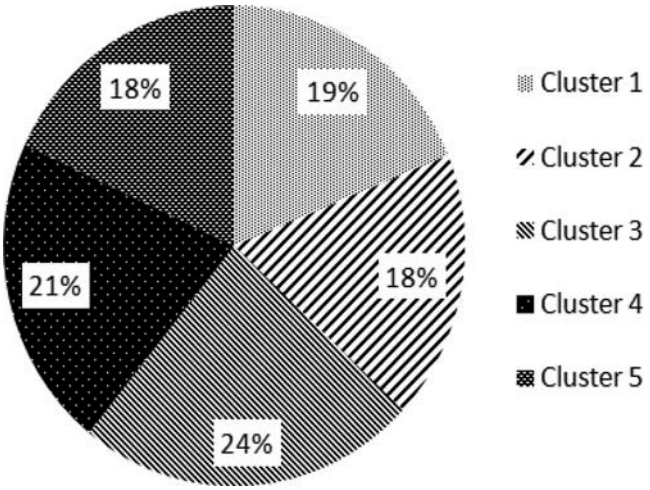

FIGURE 1 MIREX-like dataset audio clips distribution among the five clusters.

an acoustic overlap between clusters 1 and 5 (Laurier and Herrera 2007). For illustration, the word fun (cluster 2) and humorous (cluster 4) share the synonym amusing. As for songs from clusters 1–5, there are some acoustic similarities. Both are energetic, loud, and many use electric guitar (Laurier and Herrera 2007).

This dataset is available at our website[5] to any researchers wishing to use it in future research.

Audio Feature Extraction

In this work, we extract two distinct types of features from the audio samples. The first type, standard audio (SA) features, corresponds to features available in common audio frameworks. In addition, we also extract melodic audio (MA) features directly from the audio files. MA features were previously applied with success in genre recognition (Salamon et al. 2012) and are able to capture valuable information that is absent from SA features.

Standard Audio Features

As mentioned, various researchers have studied the most relevant musical attributes for emotion analysis. Several features and relations among them are now known to play an important part in the emotion present in music. Nonetheless, many of these musical characteristics are often difficult to extract from audio signals. Some are not fully understood yet and require further study from a psychological perspective.

Therefore, we follow the common practice and extract standard features available in common audio frameworks. Such descriptors aim to represent attributes of audio such as pitch, harmony, loudness, timbre, rhythm, tempo, and so forth. Some of those features, the so-called low-level descriptors (LLD), are generally computed from the short-time spectrum of the audio waveform (e.g., spectral shape features such as centroid, spread, bandwidth, skewness, kurtosis, slope, decrease, rolloff, flux, contrast or MFCCs). Other higher-level attributes such as tempo, tonality, or key are also extracted.

As mentioned, various audio frameworks are available and can be used to process audio files and extract features. These frameworks have several differences: the number and type of features available, stability, ease of use, performance, and the system resources they require. In this work, features from PsySound3,[6] MIR Toolbox,[7] and Marsyas[8] were used, and their results

[5] http://mir.dei.uc.pt/resources/MIREX-like_mood.zip.
[6] http://psysound.wikidot.com/.
[7] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox.
[8] http://marsyas.info/.

compared in order to study their importance and how their feature sets are valuable in MER.

PsySound3 is a MATLAB toolbox for the analysis of sound recordings using physical and psychoacoustical algorithms. It performs precise analysis using standard acoustical measurements, as well as implementations of psychoacoustical and musical models such as loudness, sharpness, roughness, fluctuation of strength, pitch, rhythm, and running interaural cross correlation coefficient (IACC). Although PsySound is cited in the literature (Yang et al. 2008) as having several emotionally relevant features, there are few works using this framework, possibly due to its slow speed and stability problems—some of the most interesting features, such as tonality, do not work properly, outputting the same value for all songs, or simply crashing the framework.

The MIR toolbox is an integrated set of functions written in MATLAB that are specific to the extraction of musical features such as pitch, timbre, tonality, and others (Lartillot and Toiviainen 2007). A high number of both low- and high-level audio features are available.

Marsyas (Music Analysis, Retrieval, and Synthesis for Audio Signals) is a software framework developed for audio processing with specific emphasis on MIR applications. Marsyas has been used for a variety of projects in both academia and industry, and it is known to be lightweight and very fast. One of the applications provided with Marsyas is the feature extraction tool used in previous editions of MIREX, extracting features such as tempo, MFCCs, and spectral features. Because the results of those editions are known, for comparison reasons we used the same tool, extracting 65 features.

A brief summary of the features extracted and their respective framework is given in Table 1. With regard to Marsyas, we set the analysis window for frame-level features to 512 samples. As for the MIR toolbox, we used the default window size of 0.05 seconds. These frame-level features are integrated into song-level features by the MeanVar model (Meng et al. 2007), which represents the feature by mean and variance (and also kurtosis and

TABLE 1 Frameworks Used for SA Features Extraction and Respective Features

Framework	Features
Marsyas (65)	Centroid, rolloff, flux, Mel frequency cepstral coefficients (MFCCs), and tempo.
MIR toolbox (177)	Among others: Root mean square (RMS) energy, rhythmic fluctuation, tempo, attack time and slope, zero crossing rate, rolloff, flux, high frequency energy, Mel frequency cepstral coefficients (MFCCs), roughness, spectral peaks variability (irregularity), inharmonicity, pitch, mode, harmonic change and key.
PsySound3 (11)	Loudness, sharpness, timbral width, spectral and tonal dissonances, pure tonalness, multiplicity.

skewness for MIR Toolbox). All extracted features were normalized to the [0, 1] interval.

Melodic Audio Features

The extraction of melodic features from audio resorts to a previous melody transcription step. To obtain a representation of the melody from polyphonic music excerpts, we employ the automatic melody extraction system proposed by Salamon and Gómez (2012). Figure 2 shows a visual representation of the contours output by the system for one excerpt.

Then, for each estimated predominant melodic pitch contour, a set of melodic features is computed. These features, explained in Rocha (2011) and Salamon, Rocha, and Gómez (2012), can be divided into three categories: pitch and duration, vibrato, and contour topology. Then, in global features, contour features are used to compute global per-excerpt features for use in the estimation of emotion.

Pitch and Duration Features. Three pitch features—mean pitch height, pitch deviation, pitch range—and interval (the absolute difference in cents between the mean pitch height of one contour and the previous one) are computed. The duration (in seconds) is also calculated.

Vibrato Features. Vibrato is a voice source characteristic of the trained singing voice. It corresponds to an almost sinusoidal modulation of the fundamental frequency (Sundberg 1987). When vibrato is detected in a contour, three features are extracted: vibrato rate (frequency of the variation, typical values 5–8 Hz); vibrato extent (depth of the variation, typical values 10–300 cents (Seashore 1967); vibrato coverage (ratio of samples with vibrato to total number of samples in the contour).

Contour Typology. Adams (1976) proposed a new approach to study melodic contours based on "the product of distinctive relationships among

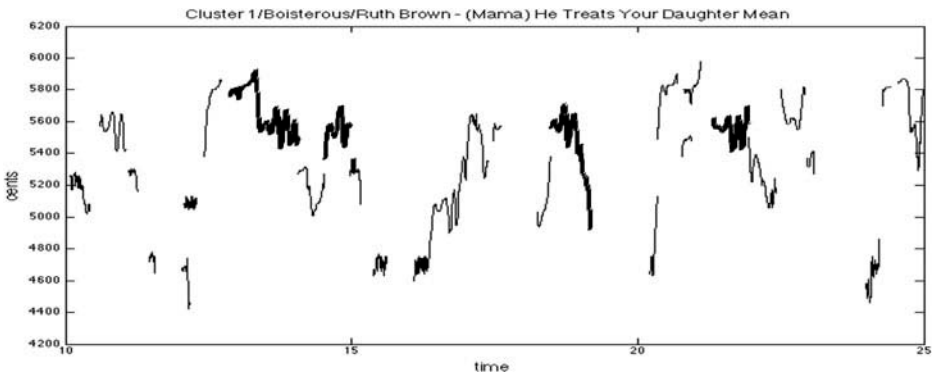

FIGURE 2 Melody contours extracted from an excerpt. A thicker line indicates the presence of vibrato.

their minimal boundaries" (Adams 1976; p. 195). By categorizing the possible relationship among a segment's initial (I), final (F), highest (H), and lowest (L) pitches, 15 *contour types* are defined. We adopt Adams' melodic contour typology and compute the type of each contour.

Global Features. The contour features are used to compute global excerpt features, which are used for the classification. For the pitch, duration, and vibrato features we compute the mean, standard deviation, skewness, and kurtosis of each feature over all contours. The contour typology is used to compute a type distribution describing the proportion of each contour type out of all the pitch contours forming the melody. In addition to these features, we also compute: the melody's highest and lowest pitches; the range between them; the ratio of contours with vibrato to all contours in the melody.

This gives us a total of 51 features. Initial experiments revealed that some features resulted in better classification if they were computed using only the longer contours in the melody. For this reason we computed for each feature (except for the interval features) a second value using only the top third of the melody contours when ordered by duration. This gives us a total of 98 features.

Applying these features to emotion recognition presents a few challenges. First, melody extraction is not perfect, especially when not all songs have clear melody. Second, these features were designed with a very different purpose in mind: to classify genre. As mentioned, emotion is highly subjective. Still, we believe melodic characteristics may be an important contribution to MER.

Emotion Classification and Feature Ranking

There are numerous machine learning algorithms available and usually applied in MER supervised learning problems. The goal of such algorithms is to predict the class of a test sample, based on a previous set of training examples used to create a model.

Thus, classifiers are used to train such models based on the feature vectors extracted from the dataset as well as the cluster labels gathered from the AllMusic database. These trained models can then be fed with new feature vectors, returning the predicted classes for them.

Various tests were run in our study with the following supervised learning algorithms: Support Vector Machines (SVM), K-nearest neighbors, C4.5, and Naïve Bayes. To this end, both Weka[9] (Hall et al. 2009), a data mining and machine learning platform, and MATLAB with libSVM were used.

[9]http://www.cs.waikato.ac.nz/ml/weka/.

In addition to classification, feature selection and ranking were performed in order to reduce the number of features and improve the results (both in terms of classification performance and computational cost). Both the ReliefF (Robnik-Šikonja and Kononenko 2003) and the CfsSubsetEval (Hall et al. 2009) algorithms were employed to this end, resorting to the Weka workbench. Regarding ReliefF, the algorithm outputs a weight for each feature, based on which the ranking is determined. After feature ranking, the optimal number of features was determined experimentally by evaluating results after adding one feature at a time, according to the obtained ranking. As for CfsSubsetEval, we kept only those features selected in all folds.

For both feature selection and classification, results were validated with 10-fold cross validation with 20 repetitions, reporting the average obtained accuracy. A grid parameter search was also carried out to retrieve the best values for parameters, for example, the γ and C (cost) parameters used in the radial basis function (RBF) kernel of the SVM model. To this end, 5-fold cross validation was used. The dataset was divided into five groups, using four to train the SVM model with candidate parameters, leaving one to test the model and measure accuracy. This was repeated to ensure that the five groups were used in testing. To find the most suitable parameters, the same procedure was repeated, varying both parameters between -5 and 15 for C and -15 and 3 for γ.

EXPERIMENTAL RESULTS

Several experiments were performed to assess the importance of the various features' sources and the effect of their combination in emotion classification.

Classification results for SA features and MA features separately, as well as the combination of both, are shown in Table 2. These results are

TABLE 2 Results for SA and MA Features, and Their Combination (F-Measure)

Classifier	SA	MA	SA+MA
NaïveBayes	37.0%	31.4%	38.3%
NaïveBayes*	38.0%	34.4%	44.8%
C4.5	31.4%	53.5%	55.9%
C4.5*	33.4%	56.1%	57.3%
KNN	38.9%	38.6%	41.7%
KNN*	40.8%	56.6%	56.7%
SVM	45.7%	52.8%	52.8%
SVM*	46.3%	60.9%	64.0%

presented in terms of F-measure using all features and feature selection (represented by *).

As shown, best results were achieved with SVM classifiers and feature selection. The commonly used standard audio features clearly lag behind the melodic features (46.3% against 60.9% F-measure). However, melodic features alone are not enough. In fact, combining SA and MA features, results improve even more, to 64%. To evaluate the significance of these results, statistical significance tests were performed. As both F-measure distributions were found to be Gaussian using the Kolmogorov–Smirnov test, the paired *t*-test was carried out. These results proved statistically significant (p-value $<$ 0.01).

Also important is that this performance was attained resorting to only 11 features (9 MA and 2 SA) from the original set of 351 SA + MA features. These results strongly support our initial hypothesis that the combination of both standard and melodic audio features is crucial in MER problems.

We also analyzed the impact of the number of features in classification performance. Table 3 shows the F-measure attained using different combinations of feature sets, feature selection algorithms, and classifiers. We excluded Naïve Bayes from this analysis because of its lower performance (as expected, because this algorithm is typically used for baseline comparison).

We see that increasing the number of features does not have a linear effect on the results. For the SA feature set, using SVM, there is no significant

TABLE 3 Results Using Different Combinations of Feature Sets, Feature Selection Algorithms and Classifiers

Feat. Set	Feat. Sel.	# Feat.	SVM	K-NN	C4.5
SA	Relief	10	36.7%	34.7%	32.6%
SA	Cfs	18	42.2%	37.6%	32.9%
SA	Relief	20	41.4%	39.1%	32.9%
SA	Relief	40	45.1%	41.0%	33.2%
SA	Relief	80	46.6%	40.8%	33.4%
SA	–	253	45.7%	38.9%	31.4%
MA	Relief	2	36.1%	45.9%	36.9%
MA	Relief	5	55.6%	56.1%	48.0%
MA	Cfs	6	56.9%	56.6%	48.0%
MA	Relief	10	60.9%	52.1%	56.1%
MA	Relief	20	57.7%	45.4%	53.7%
MA	–	98	52.8%	38.6%	53.5%
SA+MA	Cfs	8	57.1%	56.3%	52.1%
SA+MA	Relief	11	64.0%	56.7%	55.9%
SA+MA	Relief	20	59.2%	47.1%	52.9%
SA+MA	Relief	25	61.7%	49.4%	55.1%
SA+MA	Relief	40	59.8%	46.8%	57.3%
SA+MA	–	351	52.8%	41.7%	55.9%

difference in the accuracy once we reach 40 features (45.1% against 46.3%). For the MA feature set, k-nn produces the best results up to six features, decreasing substantially after that (56.6% for 6 features against 37.2% for 98 features); however, 60.9% accuracy is achieved using SVM and 10 features, still a much smaller number of features (when compared with SA's 40) yielding a much better result. It must also be noticed that we can achieve an accuracy of 45.9% using only two melodic features and a k-nn classifier. For the combination of feature sets, the best result (64%) is attained using SVM and only 11 features (2 SA + 9 MA), as mentioned previously.

In Table 4, the 11 features used to achieve the best result are listed.

As can be observed, vibrato features are particularly important (all the 9 MA features selected pertain to this category). As for SA features, one tonal and one harmony feature were selected. Hence, higher-level features are more relevant in this study than the commonly employed low-level descriptors (e.g., spectral features such as centroid, etc.).

In Table 5, we present the confusion matrix obtained with the best performing set. There are some disparities among clusters: although the performance for cluster 5 was significantly above average (76.1%), cluster 1 had a performance significantly under average (50%), with all the others

TABLE 4 Top Features of Each Feature Set

Feature Set	Feature Name
SA+MA	1. Vibrato Coverage (VC) (skew), 2. VC (kurt), 3. VC (avg), 4. Vibrato Extent (VE) (avg), 5. VE (kurt), 6. Tonal Centroid 4 (std), 7. Harmonic Change Detection Function (avg), 8. Vibrato Rate (VR) (std), 9. VC (std), 10. VR (avg), 11. VE (std)

Avg, std, skew, and kurt stand for average, standard deviation, skewness, and kurtosis, respectively.

TABLE 5 Confusion Matrix Obtained with the Best Feature Combination and libSVM

	C1	C2	C3	C4	C5
C1	50.0%	4.7%	2.9%	13.5%	28.8%
C2	1.2%	61.6%	10.4%	17.7%	9.1%
C3	0.5%	7.4%	66.0%	17.2%	8.8%
C4	0.5%	5.2%	12.6%	63.4%	18.3%
C5	0.6%	3.7%	3.1%	16.6%	76.1%

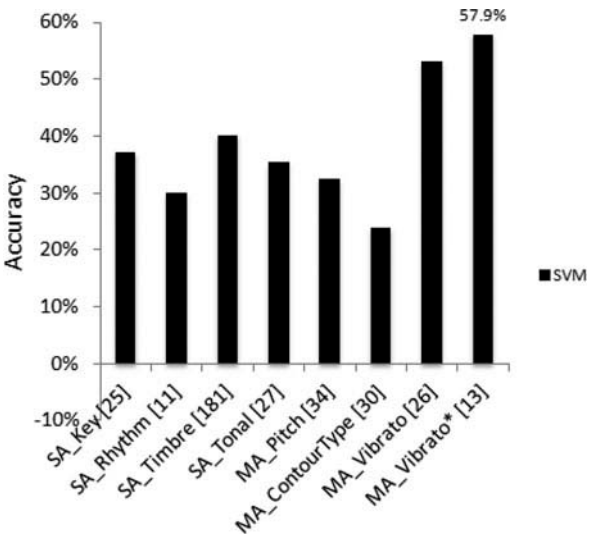

FIGURE 3 Accuracy per category of features for SA and MA using SVM.

attaining similar performance. This suggests cluster 1 may be more ambiguous in our dataset. Moreover, there are reports about a possible semantic and acoustic overlap in the MIREX dataset (Laurier 2007). Specifically, between clusters 1 and 5 and clusters 2 and 4. Because we follow the same organization, the same issue might explain the wrong classifications we observed among these clusters, especially clusters 1 and 5.

In addition, we have also experimented separating the feature sets into subsets by category. The MA set was split into pitch, contour type, and vibrato subsets, whereas four subsets were detached from the SA set: key, rhythm, timbre, and tonality. Figure 3 shows the results obtained for each subset using SVM (the number of features of each subset is between brackets; the subset signaled with * uses features computed using only the top third lengthier contours).

Looking at the results, it is interesting to notice how the MA_Vibrato* subset achieves almost the same result as the best of the MA sets with feature selection (57.9% against 60.9%). However, the weak result achieved by the rhythm subset (30%), which includes tempo features, requires special attention, because tempo is often referred to in the literature as an important attribute for emotion recognition.

As observed, melodic features, especially vibrato and pitch related, performed well in our MER experiments. The same features have already been identified as relevant in a previous study related to genre recognition (Salamon, Rocha, and Gomez 2012). Although melodic features, to the best of our knowledge, have not been used in other fields, some studies have

already uncovered the relations between emotions and vibrato in the singing voice. As an example, some results "suggest that the singers adjust fine characteristics of vibrato so as to express emotions, which are reliably transmitted to the audience even in short vowel segments" (Konishi, Imaizumi, and Niimi 2000; p. 1). Nonetheless, further research will be performed in order to understand the reasons behind such importance of vibrato in our work.

In addition to melodic, some standard features, such as tonality, contributed to improve these results. This accords with other works (Yang et al. 2008) that identified audio features such as tonality and dissonance as the most relevant for dimensional MER problems. With regard to tempo and other rhythmic features, which are normally mentioned as important for emotion (Schubert 1999; Kellaris and Kent 1993), the results were average. A possible justification might be the lack of accuracy in such features to measure the perceived speed of a song. A possible cause for this is given by Friberg and Hedblad (2011), noting that, although the speed of a piece is significantly correlated with many audio features, in their experiments none of these features was tempo. This leads the authors to the conclusion that the perceived speed has little to do with the musical tempo, referring the number of onsets per second as the most appropriate surrogate for perceptual speed. Nonetheless, some of our features were already related with onset density and none was particularly relevant to the problem. In the future we plan to further investigate this issue and research novel features related to event density, better suited to the problem.

Finally, tests using SA features split by audio framework were also conducted, showing MIR Toolbox features as achieving better results, with Marsyas close behind. While PsySound3 ranked third, it is important to note that it used a much smaller number of features when compared to the other two frameworks. A brief summary of these results is presented in Figure 4.

As a final note, we have participated in MIREX 2012 Mood Classification Task and our solution achieved the top score with 67.8% F-measure performance. Our solution used only SA features from the three frameworks tested here and SVM classifiers. The difference in results between our proposed dataset and this one might indicate that ours is more challenging than the MIREX one, although it is difficult to directly compare them. We also believe, based on this article's results, that an SA + MA solution might help improve the results achieved in the MIREX campaign.

CONCLUSIONS AND FUTURE WORK

We proposed an approach for emotion classification in audio music based on the combination of both standard and melodic audio features. To this end, an audio dataset with categorical emotion annotations, following the MIREX Mood Task organization was created (Panda and Paiva

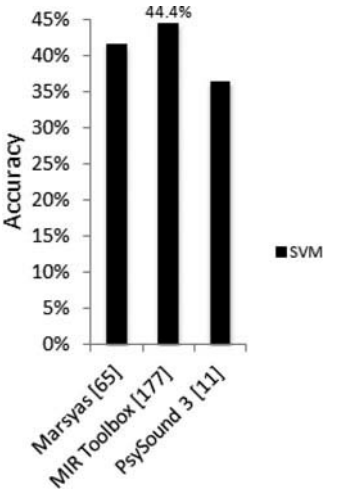

FIGURE 4 Best results (F-measure) obtained for each feature set of the SA frameworks.

2012). It is available to other researchers and might help in comparing different approaches and preparing submissions to the MIREX Mood Task competition.

The best results, 64% F-measure, were attained with only 11 features. This small set, 2 SA and 9 MA features, was obtained with the ReliefF feature selection algorithm. Using SA features alone, a result of 46.3% was achieved, whereas MA features alone scored 60.9%. The results obtained so far suggest that the combination of SA and MA helps raise the current glass ceiling in emotion classification. In order to further evaluate this hypothesis, we submitted our novel approach to the 2013 MIREX comparison. Although our submission with MA features was among the highest three, with an accuracy of 67.67%, its results were lower than expected. It is difficult to draw conclusions from this, given the fact that the MIREX dataset is secret and few details are known. Nonetheless, probable causes can be related to difficulties in the melodic transcription of the MIREX dataset, raising some questions about the robustness of these features. Another possible cause is the absence of vibrato in the dataset.

Finally, we plan to expand our dataset in the near future and explore other sources of information. Therefore, we will acquire a larger multimodal dataset containing audio excerpts and their corresponding lyrics and MIDI files. We believe that the most important factors for improving MER results overall are probably the integration of multimodal features, as well as the creation of novel features. Until now, most of the used features are SA features from the audio domain. The development of new, high-level features specifically suited to emotion recognition problems (from audio, MIDI, and

lyrics domains) is a problem with plenty of opportunity for research in the years to come.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their comments that helped improve the manuscript.

FUNDING

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) – Portugal, as well as the PhD Scholarship SFRH/BD/91523/2012, funded by the Fundação para Ciência e a Tecnologia (FCT), Programa Operacional Potencial Humano (POPH) and Fundo Social Europeu (FSE). This work was also supported by the RECARDI project (QREN 22997), funded by the Quadro de Referência Estratégica Nacional (QREN).

REFERENCES

Adams, C. 1976. Melodic contour typology. *Ethnomusicology* 20:179–215.

Carvalho, V. R., and C. Chao. 2005. Sentiment retrieval in popular music based on sequential learning. In *Proceedings of the 28th ACM SIGIR conference*. New York, NY: ACM.

Cataltepe, Z., Y. Tsuchihashi, and H. Katayose. 2007. Music genre classification using MIDI and audio features. *EURASIP Journal on Advances in Signal Processing* 2007(1): 275–279.

Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion* 6(3):169–200.

Feng, Y., Y. Zhuang, and Y. Pan. 2003. Popular music retrieval by detecting mood. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, 375–376. New York, NY: ACM.

Friberg, A. 2008. Digital audio emotions - an overview of computer analysis and synthesis of emotional expression in music. Paper presented at the 11th International Conference on Digital Audio Effects, Espoo, Finland, September 1–4.

Friberg, A., and A. Hedblad. 2011. A comparison of perceptual ratings and computed audio features. In *Proceedings of the 8th sound and music computing conference*, 122–127. SMC.

Gabrielsson, A., and E. Lindström. 2001. The influence of musical structure on emotional expression. In *Music and emotion: Theory and research*, ed. P. N. Juslin and J. A. Sloboda, 223–248. Oxford, UK: Oxford University Press.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11(1):10–18.

Hevner, K. 1936. Experimental studies of the elements of expression in music. *American Journal of Psychology* 48(2):246–268.

Hu, X., and J. S. Downie. 2010. When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of the 11th international society for music information retrieval conference* (*ISMIR 2010)* , 619–624. Utrecht, The Netherlands: ISMIR.

Hu, X., J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. 2008. The 2007 Mirex audio mood classification task: Lessons learned. In *Proceedings of the 9th international society for music information retrieval conference (ISMIR 2011)*, 462–467. Philadelphia, PA, USA: ISMIR.

Huron, D. 2000. Perceptual and cognitive applications in music information retrieval. Cognition 10(1):83–92.

Katayose, H., M. Imai, and S. Inokuchi. 1988. Sentiment extraction in music. In *Proceedings of the 9th international conference on pattern recognition*, 1083–1087. IEEE.

Kellaris, J. J., and R. J. Kent. 1993. An exploratory investigation of responses elicited by music varying in tempo, tonality, and texture. *Journal of Consumer Psychology* 2(4):381–401.

Konishi, T., S. Imaizumi, and S. Niimi. 2000. Vibrato and emotion in singing voice (abstract). In *Proceedings of the sixth international conference on music perception and cognition (ICMPC), August 2000* (CD-rom). Keele, UK: Keele University.

Lartillot, O., and P. Toiviainen. 2007. A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th international conference on digital audio effects*, 237–244. Bordeaux, France: ICDAFx-07.

Laurier, C. 2011. *Automatic classification of musical mood by content-based analysis* (PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain).

Laurier, C., and P. Herrera. 2007. Audio music mood classification using support vector machine. In MIREX task on Audio Mood Classification 2007. *Proceedings of the 8th international conference on music information retrieval*, September 23–27, 2007. Vienna, Austria.

Laurier, C., O. Lartillot, T. Eerola, and P. Toiviainen. 2009. Exploring relationships between audio features and emotion in music. In *Proceedings of the 7th triennial conference of European Society for the Cognitive Sciences of Music* (ESCOM 2009) Jyväskylä, Finland, 260–264.

Li, T., and M. Ogihara. 2003. Detecting emotion in music. In *Proceedings of the* 2003 *international symposium on music information retrieval (ISMIR 03)*, 239–240. ISMIR.

Li, T., and M. Ogihara, M. 2004. Content-based music similarity search and emotion detection. In *Proceedings of the* 2004 *IEEE international conference on acoustics, speech, and signal processing*, 5:V–705. IEEE.

Liu, D., and L. Lu. 2003. Automatic mood detection from acoustic music data. *International Journal on the Biology of Stress* 8(6):359–377.

Lu, L., D. Liu, and H.-J. Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing* 14(1):5–18.

McVicar, M., and T. Freeman. 2011. Mining the correlation between lyrical and audio features and the emergence of mood. In *Proceedings of the 12th international society for music information retrieval conference (ISMIR 2011)*, 783–788. Miami, FL, USA: ISMIR.

Meng, A., P. Ahrendt, J. Larsen, and L. K. Hansen. 2007. Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech and Language Processing* 15(5):275–279.

Meyers, O. C. 2007. *A mood-based music classification and exploration system* (Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, USA).

Ortony, A., and T. J. Turner. 1990. What's basic about basic emotions? *Psychological Review* 97(3):315–331.

Panda, R., and R. P. Paiva. 2012. Music emotion classification: Dataset acquisition and comparative analysis. Paper presented at the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK, October.

Panda, R., B. Rocha, and R. Paiva. 2013. Dimensional music emotion recognition: Combining standard and melodic audio features. Paper presented at the Computer Music Modelling and Retrieval - CMMR'2013. Marseille, France, October 15–18.

Robnik-Šikonja, M., and I. Kononenko. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning* 53(1–2):23–69.

Rocha, B. 2011. *Genre classification based on predominant melodic pitch contours* (Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain).

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.

Salamon, J., and E. Gomez. 2012. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio Speech and Language Processing* 20(6):1759–1770.

Salamon, J., B. Rocha, and E. Gómez. 2012. Musical genre classification using melody features extracted from polyphonic music signals. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Kyoto, Japan: IEEE.

Schubert, E. 1999. *Measurement and time series analysis of emotion in music* (PhD Thesis, School of Music and Music Education, University of New South Wales, Sydney, Australia).

Seashore, C. 1967. *Psychology of music*. New York, NY: Dover.

Song, Y., S. Dixon, and M. Pearce. 2012. Evaluation of musical features for emotion classification. In *Proceedings of the 13th international society for music information retrieval conference (ISMIR 2012)*, 523–528. Porto, Portugal. ISMIR.

Sundberg, J. 1987. *The science of the singing voice*. Dekalb, IL, USA: Northern Illinois University Press.

Thayer, R. E. 1989. *The biopsychology of mood and arousal*. New York, NY: Oxford University Press.

Wang, J., H. Lee, S. Jeng, and H. Wang. 2010. Posterior weighted Bernoulli mixture model for music tag annotation and retrieval. Paper presented at the APSIPA Annual Summit and Conference (ASC) 2010. December 14–17, 2010. Biopolis, Singapore.

Wang, J., H.-Y. Lo, S. Jeng, and H.-M. Wang. 2010. Audio classification using semantic transformation and classifier ensemble. In *Proceedings of the 6th International WOCMAT and New Media Conference (WOCMAT 2010)*, YZU, Taoyuan, Taiwan, November 12–13, 2010, 2–5.

Yang, D., and W. Lee. 2004. Disambiguating music emotion using software agents. In *Proceedings of the 5th international conference on music information retrieval*, 52–58. Barcelona, Spain: ISMIR.

Yang, Y.-H, Y.-C. Lin, H. Cheng, I. Liao, Y. Ho, and H. H. Chen. 2008. Toward multi-modal music emotion classification. In *Proceedings of the* 2008 *Pacific-rim conference on multimedia*, LNCS 5353:70–79. Berlin, Heidelberg: Springer.

Yang, Y.-H., Y.-C. Lin, Y.-F. Su, and H. H. Chen. 2008. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 16(2):448–457.