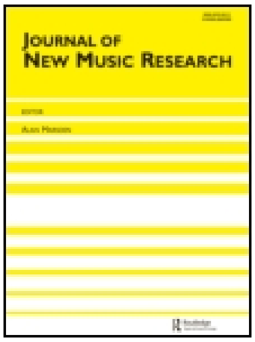

Routledge
Taylor & Francis Group

Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: https://www.tandfonline.com/loi/nnmr20

Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres

Tuomas Eerola

To cite this article: Tuomas Eerola (2011) Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres, Journal of New Music Research, 40:4, 349-366, DOI: 10.1080/09298215.2011.602195

To link to this article: https://doi.org/10.1080/09298215.2011.602195

Published online: 25 Nov 2011.

Journal of New Music Research
2011, Vol. 40, No. 4, pp. 349–366

Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres

Tuomas Eerola

University of Jyväskylä, Finland

Abstract

Empirical studies of emotions in music have described the role of individual musical features in recognizing particular emotions. However, no attempts have been made as yet to establish if there is a link between particular emotions and a specific genre. Here this is investigated by analysing nine separate datasets that represent categories ranging from classical (three sets), and film music (two), to popular music (two), and mixed genre (two). A total of 39 musical features were extracted from the audio. Models were then constructed from these to explain self-reports of valence and arousal, by using multiple and Random Forest regression. The models were fully validated across the datasets, suggesting low generalizability between the genres for valence (16% variance was accounted for) and moderately good generalizability between the genres for arousal (43%). In contrast, the generalizability within genres was considerably higher (43% and 62% respectively), which suggests that emotions, especially those that express valence, operate differently depending on the musical genre. The most reliable musical features of affects across genres were identified, yielding a ranked set of features most likely to operate across the genre. In conclusion, the implications of the findings, and the genre-specificity of emotions in music are discussed.

1. Introduction

Music has the ability to convey powerful emotional meanings to listeners. This process is undoubtedly complex, as it is related to aspects in the overall communication of emotions such as their perception or induction (Juslin & Västfjäll, 2008), models of emotions (Eerola & Vuoskoski, 2011), the personalities of listeners (Kallinen & Ravaja, 2004), and musical expectations (Huron, 2006). Music itself could be considered as the single most important source for emotional communication in terms of all these separate factors. It is therefore no surprise that there has been a wealth of research, over the past nine decades, into the individual elements of music that trigger certain emotions (e.g. Gabrielsson & Lindström, 2010). For the most part, this body of work has succeeded in determining the main musical elements that play such a role, however there are a number of shortcomings in a significant portion of the studies which either (a) deal with musical features in isolation (Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Ilie & Thompson, 2006), (b) are based on artificial stimulus materials (Bresin & Friberg, 2000; Gomez & Danuser, 2004; Vieillard et al., 2008), (c) are overly focused on the symbolic representation of music (Juslin, 1997b; Gagnon & Peretz, 2000; Lindström, 2003), or (d) rely excessively on classical music (approx. 50% according to a review by Eerola and Vuoskoski (2011), and Juslin and Laukka (2003)). Studies which avoid these shortcomings are few and far between.

Significant advances have been made over the past decade in signal processing, and there have been corresponding changes to perceptual and cognitive models of music processing (see recent reviews in Purwins, Herrera, et al., 2008; Purwins, Grachten, et al., 2008). In view of this it makes sense to prioritize

Correspondence: Dr Tuomas Eerola, University of Jyväskylä, Department of Music, Seminaarinkatu 35, Jyväskylä, FI-40014 Finland. E-mail: tuomas.eerola@jyu.fi

http://dx.doi.org/10.1080/09298215.2011.602195 © 2011 Taylor & Francis

approaches based on musical features that are extracted directly from the audio, since it allows the researcher to focus on more subtle and sophisticated elements such as timbre, a feature which has also proved crucial in research on how emotions are expressed in speech (Banse & Scherer, 1996; Juslin & Laukka, 2003). What was previously a trickle of studies using audio-extracted features, has grown into quite a stream of publications over the last five years (e.g. Schubert, 2004; Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Lu, Liu, & Zhang, 2006; MacDorman, 2007; Yang, Lin, Su, & Chen, 2007; Coutinho & Cangelosi, 2009; Eerola, Lartillot, & Toiviainen, 2009). These studies avoid most of the shortcomings mentioned earlier, but they have other drawbacks: (a) many rely on small datasets (about 2/3 of all music and emotion studies have less than 20 stimuli, Eerola & Vuoskoski, 2011), (b) some models are constructed without using cross-validation with an external dataset (e.g. Schubert, 2004; Leman et al., 2005), (c) many cover an insufficient number of different musical genres or ignore such aspects altogether (Schubert, 2004; Leman et al., 2005; Coutinho & Cangelosi, 2009), and (d) often they employ widely disparate musical features making it difficult to infer whether any underlying core set of features exists for emotions in music.

The present study concentrates on musical genre, and whether it has a part to play in the way emotions are conveyed. By carrying out a large scale evaluation of the relevant musical features across different musical genres, and by using real musical excerpts and a computational method of feature extraction from audio excerpts, many of the limitations that previous studies encountered should be removed.

2. Emotions and music

The field of music and emotion research has grown rapidly and diversified in many respects during the last decade. Not only has it become more popular, but it has become increasingly interdisciplinary, bringing together topics from music cognition, psychology and neuroscience. These disciplines have brought a wide range of approaches, models and settings to the field (summary in Zentner & Eerola, 2010). Emotion research in general recognizes the problems inherent in defining emotions in a way that would be acceptable to most researchers (Frijda, 2007; Izard, 2007). One such difficulty relates to the fact that music may be assumed to express certain emotions in music (i.e. perceived emotions), but whether or not it induces those emotions in a listener (induced emotions) is another matter. Sometimes the emotion induced is the same as that perceived, as these two processes bear some resemblance to each other (Evans & Schubert, 2008; Vieillard et al., 2008), but this is not always the case. In terms of computational modelling, the emotions perceived in music provide a better object of modelling, since they rely more on the musical features than an individual's felt experiences (induced emotions), which are prone to subjective associations and contextual effects (Gabrielsson, 2002; Juslin & Västfjäll, 2008). For this reason, the emphasis of the present study is clearly on perceived emotions in music, just as it is for most of the previous empirical efforts relevant to this study.

Three types of model have been used to represent perceived emotions—categorical, dimensional, and music-specific (summarized in Eerola & Vuoskoski, 2011). According to the categorical emotion model, all emotions can be derived from a limited number of innate basic emotions such as fear, anger, disgust, sadness and happiness. In music-related studies however, even though many of these have been found to be appropriate (Juslin & Laukka, 2004), certain emotions (i.e. disgust) have often been replaced by more appropriate ones (e.g. tenderness or peacefulness, Juslin, 1997b; Bresin & Friberg, 2000; Lindström, 2003; Balkwill, Thompson, & Matsunaga, 2004). Perhaps the best known example of the dimensional model is the two-dimensional circumplex (Russell, Weiss, & Mendelsohn, 1989), which proposes that all affective states arise from independent neurophysiological systems—one set related to valence and the other to arousal. This particular model has received a great deal of attention in music and emotion studies, and is also convenient in the analysis of large sets of mood labels (Hu & Downie, 2007). Recently, Zentner and his colleagues have proposed a more detailed model for music-induced emotions—called the Geneva Emotion Music Scale, or GEMS, which has nine-factors (Zentner, Grandjean, & Scherer, 2008). However, since it only deals with induced emotions, and there is not yet enough data to begin comprehensive modelling of it, the emphasis here will therefore be on evaluating studies which use the two-dimensional model of emotions.

2.1 The features of music that contribute to perceived emotional expression

Since the 1930s, there have been scientific attempts to pin down the specific features of music that communicate emotion (Hevner, 1936, 1937). These studies have mostly focussed on how emotional expression is perceived rather than induced, although the boundary between these two processes can be somewhat fuzzy. A recent summary of this research can be found in Gabrielsson and Lindström (2010), and it suggests that, from among all the musical features that have been studied to date, the most potent are mode, tempo, dynamics, articulation, timbre, and phrasing. Happiness and sadness provide a good example of two typical emotions expressed in music. They have received considerable attention in the literature since they can be fairly clearly distinguished in terms of tempo, pitch height, and mode. Happiness is expressed by using faster tempi, a higher pitch range, and a major rather than minor mode, and these features tend toward the opposite extreme in musical expressions of sadness (Wedin, 1972; Gerardi & Gerken, 1995; Peretz, Gagnon, & Bouchard, 1998; Dalla Bella et al., 2001). Certain typical constellations of musical features have also been observed for other emotions, such as anger, fear, tenderness and peacefulness, whether they are represented as discrete emotions (Bresin & Friberg, 2000; Balkwill et al., 2004; Vieillard et al., 2008), or on a dimensional model of emotion (e.g. Ilie & Thompson, 2006; Livingstone, Muhlberger, Brown, & Thompson, 2010).

From such summaries it is possible to draw up a list of candidate features that will be featured in a computational model of emotions. Although symbolic representations of music (MIDI, **kern, etc.) may offer easier analysis of certain musical features such as exact pitches, velocities and durations, they usually have little information about other more timbral sound qualities. As many of the more recent studies focus on timbral aspects of music, they therefore tend to use audio-based analysis, as this is the only format which faithfully represents timbre in all its richness.

2.2 Audio as a basis for modelling expressed emotions in music

Most of the past computational work on music and emotions has dealt with predicting emotional categories. For instance, Li and Ogihara (2003) used acoustic features related to timbre, rhythm, and pitch to train Support Vector Machines (SVMs) to classify music into 13 mood categories. They used half the database for training the machines and the other half for testing it, but achieved an accuracy of only 45%. Lu et al. (2006) based their system for classifying quadrants in the valence-arousal affective space on a variety of acoustic features related to intensity, timbre, and rhythm, constructing Gaussian Mixture Models (GMMs), and achieved an accuracy of 85% (cf. also Feng, Zhuang, & Pan, 2003). The accuracy rate has increased in more recent work using acoustic features for mood classification (Skowronek, McKinney, & Par, 2007; Hu, Downie, Laurier, Bay, & Ehmann, 2008). Here it was found that certain emotion categories (e.g. angry, calming) were much easier to predict, with correct classification exceeding 90%, than others (e.g. carefree, loving) where the accuracy was 75–80%. In these studies, the emotions have been defined according to discrete categories within the affect space, typically called emotion quadrants (Ritossa & Rickard, 2004; Rawlings & Leow, 2008).

Whereas modelling emotion categories has been a common choice among the MIR community (e.g. Lu et al., 2006; Skowronek et al., 2007; Yang et al., 2007), the underlying emotions are perhaps captured in a more subtle and realistic way by modelling them in terms of continua within and between the emotions (in this case, dimensions), since participants have made subtle decisions on Likert scales and have not had to reduce their emotional experiences into categories. Moreover, discrete categories of emotions (basic emotions or emotion quadrants) are not an optimal form of data from a behavioural point of view since the forced-choice paradigm (Zentner & Eerola, 2010) is notorious for creating inflated recognition rates if too few response categories are provided (Russell, 1994). A full representation of the affective space and linear (or non-linear) models of them will take precedence here, since they preserve the scalar nature of the responses (Schubert, 2004; MacDorman, 2007; Eerola et al., 2009).

A handful of studies have gone a step further than taking just one static rating per musical excerpt, and mapped continuous ratings on a dimensional model of valence and arousal, or explored these affects within multiple musical genres (Schubert, 2004; Leman et al., 2005; MacDorman, 2007; Yang et al., 2007; Coutinho & Cangelosi, 2009; Eerola et al., 2009). In a study by Schubert (2004), four excerpts of classical music were chosen for a continuous rating task. Forty-three percent of variance in the valence ratings could be explained using a linear combination of loudness, tempo, melody, and texture. For arousal, this figure was higher, at 58%, and the dominant musical features were found to be a combination of spectral centroid, loudness, and tempo. In 2005, Leman and his colleagues obtained results similar in magnitude with a larger and more varied set of musical examples (60 excerpts spanning 10 genres): 30% of variance in valence could be explained in terms of onsets, IOIs and articulation, whereas 46% of variance was linked to spectral centroid and pitch commonality. In 2007, MacDorman modelled valence and arousal using 100 excerpts covering 15 genres using spectrum and periodicity histograms, fluctuation patterns, and MFCCs. They also applied dimensionality reductions (PCA and ISOMAP) to trim down the number of components in the features and obtained impressive results ($R^2 > 0.95$ for both dimensions). In the same year, Yang and his colleagues (2007) also modelled valence and arousal by having listeners rate 60 excerpts of popular music. These ratings were predicted with 45 features extracted from 19 timbral, six rhythmic, and five pitch content features, and obtained 17% of variance for valence and 79.9% for arousal. Coutinho and his colleague (2009) predicted continuous ratings of valence and arousal in six excerpts of classical music with 15 features spanning dynamics, timbre and pitch, using a spatiotemporal neural network. They obtained an extraordinarily high degree of success ($R^2 > 00.95$), although the actual makeup of the model is difficult to infer due to the neural network architecture of the model. Finally, Eerola and his colleagues (2009) constructed several models with various types of regression techniques using 110 film music excerpts. They succeeded in explaining 51% of variance in valence ratings in terms of RMS variation, mode, and key clarity, and 77% of variance in arousal ratings in terms of spectral centroid, key clarity, and register.

In sum, several recent studies have successfully demonstrated how the valence-arousal affect space can be predicted using a set of musical and acoustic features. In all the studies, arousal ratings seem to be the easiest to account for, whereas valence ratings, although semi-successfully explained by the models to date, are thought to be more subjective and elusive in terms of their musical properties. There is a reasonable concern over the results with respect to their generalization accuracy, core features of emotions and dependence on particular kinds of dataset. Although almost all of these studies used internal cross-validation, the resulting models were not then tested with external datasets and nor were the genre-specific qualities of the features discussed. Also, widely different features of emotions were implicated across the studies. In fact, actually comparing the musical features used across all of them is impractical, since these varied widely in each instance, as did the musical genres used, and the methodologies for constructing the models.

Interestingly, no systematic studies have yet been conducted to assess the relative genre-specificity of musical features that communicate certain emotions, even though it seems intuitive that genre should play a major role in determining what those features might be, particularly since the high success rate of genre-recognition algorithms show that genres do indeed have distinct musical features (Tzanetakis & Cook, 2002b; Bergstra, Casagrande, Erhan, Eck, & Kögl, 2006; Silla Jr, Kaestner, & Koerich, 2007). What is meant in this study by the term genre—such as jazz, pop, and classical—should be clarified at this point. Genre remains the most popular kind of tag for describing musical content, and is used to organize libraries and online collections of music. Often identified from extremely brief excerpts (Gjerdingen & Perrott, 2008), genre is equally the subject of automatic recognition systems (e.g. Marques, Langlois, Gouyon, Lopes, & Sordo, 2011). However, the genres themselves—also called musical styles in the analysis and history literature—are not easy to define, since no natural genre taxonomies exist and they are defined as cultural constructs rather than by any intrinsic musical properties (e.g. Aucouturier & Pachet, 2003). Also worth exploring are the claims from cross-cultural studies of emotion recognition, that there may also be an underlying core set of emotional cues that operate across culture and genre, and not just culture or genre-specific cues (Balkwill & Thompson, 1999; Balkwill et al., 2004).

3. Aim of the study

The aim of the study is explore the reliance of emotion models and their features upon musical genre. For this, a computational modelling of valence and arousal, using audio as the underlying representation of music, will be taken as the starting point. Two particularly important analysis strategies will be used to gauge the possible genre-specificities of emotions: (1) an exhaustive cross-validation scheme using a planned variety of datasets spanning classical, film, and pop music as well as mixed genre datasets, and (2) an independent feature evaluation paradigm. The first strategy aims to tackle the question of how genre-dependent models for valence and arousal actually are, by comparing the model's prediction accuracy within and between genres. The latter paradigm should enable us to identify whether there is a possible core set of musical features that will contribute to emotions irrespective of the genre of music.

4. Method

The external validity of a predictive system is established by being tested and found accurate across diverse settings (Justice, Covinsky, & Berlin, 1999). In order to carry out a large-scale comparison of the way individual emotions are expressed across musical genres, using a wide range of musical features, examples representing several separate musical styles are necessary. These materials will all need to be annotated or rated by listeners for compatible notions of emotions (in terms of valence and arousal). This method of selection was chosen as previous work has shown valence and arousal to be the most non-redundant way of representing emotion ratings when making a direct comparison of discrete and dimensional models of emotions (Eerola & Vuoskoski, 2011). Moreover, these two dimensions are the most prevalent in previous studies of music and emotions (Zentner & Eerola, 2010), even if the merits of the model itself are still debatable (Zentner et al., 2008).

To compare these datasets, a comprehensive array of musical features needs to be assembled. To provide mutually compatible and transparent models, we will resort to regression analyses. However, to avoid the problems of high dimensionality, which is particularly demanding for regression (in terms of the number of observations needed for predictors), a data reduction will be carried out. The modelling will take place within an exhaustive cross-validation scheme. Finally, the contribution of independent features across musical genres will be studied separately, using Random Forest regression.

4.1 Musical materials

Several previous studies exist which could be used for evaluating emotions across different musical genres (e.g. Altenmüller, Schuermann, Lim, & Parlitz, 2002; Kallinen, 2005; Gomez & Danuser, 2007). However, the emotion concepts rated in these studies are in many cases incompatible, or the examples overly familiar to most listeners. This study takes nine datasets for analysis, which all rely on ratings of valence and arousal, and which, almost without exception, use only real music examples of a music (listed in Table 1). Three of these feature classical music (Dibben, 2004; Gomez & Danuser, 2004; Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005), and will be referred to hereafter as studies C1, C2 and C3; two studies (F1 & F2) were based on film soundtracks (Gosselin et al., 2005; Eerola & Vuoskoski, 2011); two others (P1 & P2) investigate popular music, (Yang et al., 2007; Schuller, Dorfner, & Rigoll, 2010); and, in order to have a comparison with genre-specific collections, two datasets (M1 & M2) were chosen because they had a mixture of genres (Nater, Abbruzzese, Krebs, & Ehlert, 2006; Wu & Jeng, 2006).

With the exception of one heavy metal excerpt in C1, which was used to represent an example of extreme negative valence combined with a high arousal condition (Gomez & Danuser, 2004), the C datasets consist, almost exclusively, of classical music examples: C2 uses a combination of Mozart and Haydn symphonies and Mozart quintets (Dibben, 2004), C1 has a piano concerto by Ravel, the first symphony by Mahler, and Salut d'Amour by Elgar (Gomez & Danuser, 2004), and C3 has symphonies by Wagner, Beethoven, Mendelssohn, and Shostakovitch (Bigand et al., 2005). In the latter study, Bigand et al. convincingly argue that the ratings of valence and arousal can be extracted and turned into their respective dimensions by making a scaling solution of the similarities for each excerpt.

Table 1. List of materials included (*N* refers to number of stimuli, *L* to median stimulus length is seconds).

Study	Genre	Abbr.	*N*	*L*
Gomez and Danuser (2004)	Classical	C1	16	30
Bigand et al. (2005)	Classical	C2	27	37
Dibben (2004)	Classical	C3	16	35
Eerola and Vuoskoski (2011)	Film	F1	110	15
Gosselin et al. (2005)	Film	F2	40	12
Yang et al. (2007)	Pop	P1	60	30
Schuller et al. (2010)	Pop	P2	2598	30
Nater et al. (2006)	Mix	M1	20	29
Wu and Jeng (2006)	Mix	M2	75	12

The two film music datasets have some important characteristics which should be mentioned. F1 (Eerola & Vuoskoski, 2011) features examples from unfamiliar soundtracks, and contains moderate or ambiguous examples as well as the more customary best examples of the expressed emotions.[1] The 40 excerpts in F2 differ in one sense from the rest of the datasets, in that they are not real recorded musical performances but artificial compositions, made in the style of film music, and rendered in MIDI (Gosselin et al., 2005). Nevertheless, since these stimuli are frequently used in music and emotion studies (Gosselin et al., 2006; Gosselin, Peretz, Johnsen, & Adolphs, 2007; Vieillard et al., 2008; Fritz et al., 2009), and consist nonetheless of polyphonic, orchestrated examples with different instrumentations, the stimulus set is considered realistic enough to be included here in the comparison.

Dataset P1 (Yang et al., 2007) uses well known examples of popular music (e.g. 'Smells Like Teen Spirit' by Nirvana, 'Dancing Queen' by ABBA, 'Hells Bells' by AC/DC, 'Imagine' by John Lennon, etc.), while P2 (Schuller et al., 2010) is more extensive, comprising a total of 2648 songs from the *Now That's What I Call Music!* —double CD collections (volumes 1–69). For the purposes of the present study, the author purchased 98% of these titles (2598 songs). In Schuller et al.'s study, a sample of 30 s was taken from the middle of each song, and this was then rated by a panel of three experts. However, the rating data was in the end based on an average of only two of these experts, since one evaluator was discarded due to low intersubject correlation with the other two.

To contrast with the previous datasets, which predominantly contain examples from a single genre, two datasets with mixed genres were included (M1 and M2). Examples of the dataset M1 (Nater et al., 2006) consists of the following genres: pop (e.g. 'Discovery' by Chris De Burgh), film soundtracks (e.g. 'La Libertad' from *Alibi* by Ennio Morricone), classical music (e.g. 'Hör ich die Stimmen im Traum' by Georges Bizet), and heavy metal (e.g. 'Left Hand Path' by Entombed). M2 (Wu & Jeng, 2006) also contains such musical genres as classical music (e.g. 'Peer Gynt' by Grieg), pop (e.g. 'Fight in the Cave' by Nathan Wang and Gary Chase), latin (e.g. 'Puerto de Soller' by Mario Berger), film soundtracks (e.g. *Brothers* by Hans Zimmer), and easy listening (e.g. 'Out of Macchu Pichu' by Georg Gabler).

For the valence and arousal ratings in all datasets, the mean ratings of the excerpts were normalized by z-score transformation. A summary of the datasets including the number of stimuli and the median stimulus lengths

[1] All sound examples available at http://www.jyu.fi/music/coe/materials/emotion/soundtracks/

for each collection can be found in Table 1. Meanwhile, Figure 1 displays the mean ratings for all the datasets across the affect space.

4.2 Feature extraction

The computational extraction of musical features has become a far more widespread phenomenon, due to the wealth of new research in Music Information Retrieval (MIR). A number of efficient tools have been created that allow for the rich processing of audio (Tzanetakis & Cook, 2000; Leman, Lesaffre, & Tanghe, 2001; Lartillot & Toiviainen, 2007). Consequently, hundreds of features may now be extracted from audio (MFCCs, and various descriptive statistics of frame-based analysis of rhythm, pitch and more).

Constructing linear models via regression with a large number of candidate features is, however, problematic since there has to be at least 10 or 20 times more cases than predictors (Hair, Black, Babin, Anderson, & Tatham, 2006). Such large numbers of observations are rarely available, and of the present nine datasets, only one would allow the use of hundreds of features (P2). This stresses the importance of first theoretically selecting and empirically reducing the features, in order to avoid overfitting the model with unnecessary features which would render it unstable.

The theoretical selection of features was made by first basing them on traditional descriptors of musical elements (dynamics, articulation, rhythm, timbre, pitch, tonal content, and form) and then by representing them with a few, non-redundant (i.e. low-correlating) features. Table 2 displays the *full feature set* of 39 features that represent these elements of music in a way that uses the collinearity minimization principle, as employed in a previous study by Eerola et al. (2009). In this procedure, features within the theoretical elements were selected by minimizing correlation between the candidate features, resulting in a low median ($r = 0.165$, $p = \text{ns}$) correlation within the elements.

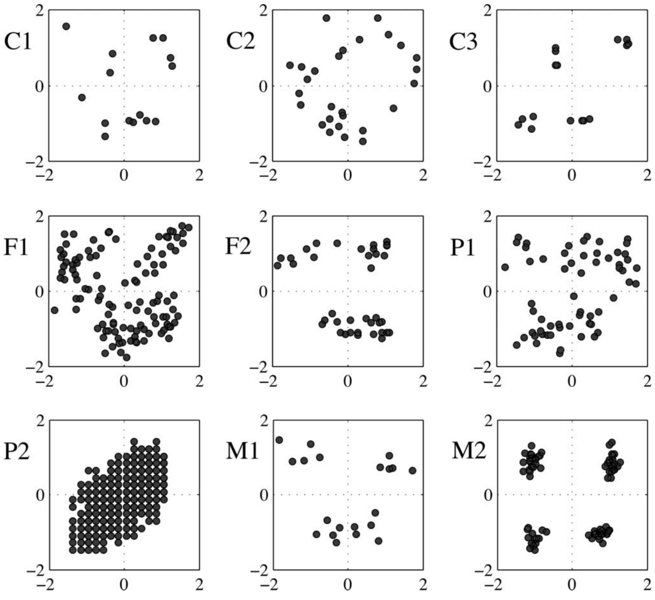

Fig. 1. Behavioural ratings in all datasets (normalized) showing valence (x-axis) and arousal (y-axis).

All features were extracted with the *MIRtoolbox*[2] (Lartillot, Toiviainen, & Eerola, 2008) using a frame-based approach (Tzanetakis & Cook, 2002a). For low-level features, analysis frames of 42 ms were used, a 50% overlap between frames. For high level features [Tempo, Pulse Clarity, Key Clarity, Key Mode and HCDF (Harmonic Change Detection Function by Harte, Sandler, & Gasser, 2006), the centroid of an uncollapsed Chromagram], the frames were 2 s long, with an overlap of 50%, while for structural features (Repetition), frame length was 100 ms and overlap also at 50%. The results from the frames were then summarized by either the mean, the standard deviation, or the slope function. As all the features are available and documented in MIR toolbox, only recently proposed features will be briefly explained here.

Event Density is evaluated by detecting onsets by peaks picked from the amplitude envelope. For each onset, the attack time and slope is also estimated. Rhythmic periodicity is assessed both from a spectral analysis of each band of the spectrogram, leading to a fluctuation pattern (Pampalk, Rauber, & Merkl, 2002), and based on the assessment of autocorrelation in the amplitude envelope extracted from the audio. Pulse Clarity is evaluated by the global characteristic of the autocorrelation function of the amplitude envelope (Lartillot, Eerola, Toiviainen, & Fornari, 2008). In Roughness, the peaks in the spectrogram are utilized to estimate a measure of roughness, based on a model by Sethares (1998). A somewhat related measure, Spectral Entropy, is based on the spectrum collapsed into one single octave from which the entropy is calculated as an indication of simplicity or complexity of the chroma components. Harmonic Change Detection Function (HCDF) indicates the tonal diversity along the time (Harte et al., 2006) and Key Clarity, based also on chromagram, compares the ensuing tonal profile to pre-established key profiles representing major and minor keys (Gomez, 2006). Also there is a new measure of majorness (Eerola et al., 2009), which is the amplitude difference between the best major score and minor score obtained by correlation with the key profiles. Pitch uses an estimation of the centroid and deviation of the unwrapped chromagram, and also in parallel a statistic description of pitch component based on advanced

[2] Version 1.3, which is available from www.jyu.fi/music/coe/material/mirtoolbox

pitch extraction method (Tolonen & Karjalainen, 2000). Finally, a degree of repetition is estimated through the computation of novelty curves (Foote & Cooper, 2003) based on various functions already computed such as the spectrogram, the autocorrelation function, key profiles and the chromagram, each representing a different aspect of the novelty or static temporal nature of the music.

4.3 Feature reduction

In order to improve the model's robustness and make the individual features easier to interpret, further reduction of the 39 features was needed. The most straightforward way to do this is to apply Principal Component Analysis (PCA) to the full feature set, based on the features extracted from an unrelated dataset. For this purpose, a collection of 6372 songs was used to establish a musical sample (Eerola & Ferrer, 2009). It was used for two major reasons. Firstly it provided maximal variety, spanning 15 musical genres: alternative, folk, iskelmä, pop, world, blues, gospel, jazz, rock, classical, heavy metal, soul, electronic, hip-hop, and soundtrack. Secondly, it remained ecologically valid, as there were approximately 400 of the most typical examples for each of these genres. The full feature set was extracted by taking a 15 s excerpt out of each file in the collection (randomly selected from a point between 30% and 50% into the song). Next PCA was applied to all 39 Z-scores of the features thus obtained. Using eigenvalues >1 as the component selection criteria, this yielded a nine-component solution, that explained 81% of the variance for the original matrix. In other words, nine principal component features could be used instead of the full 39 features with a tolerable loss of only 19%.

The next step was to take the individual features as the *reduced set of features*, instead of relying solely on the principal components (PCs) as predictors. This is because the original features are more straightforward to explain than the PCs, as PCs unfortunately require a linear combination of the whole initial matrix to account for themselves. So the selection of optimal individual features (representing the nine PCs) was performed instead using feature selection principles as outlined by Jolliffe (see Al-Kandari & Jolliffe, 2001). In other words, the highest correlating feature for each PC was chosen and highlighted in Table 2. This operation resulted in nine features, which are roughly divisible into four groups: the *dynamic* (RMSl, RMSd), the *rhythmic* (PCd, Td, Tm, FMm), the *timbral* (SFm), and the *tonal* (Mm and Pm). These features are not only representative and compact, but are also, to all intents and purposes, orthogonal since they are based on PCs which are orthogonal. This makes them optimally suited for regression modelling.

Table 2. The full feature set and the reduced feature set (nine highlighted features).

Domain	No.	Feature	Acronym
Dynamics	1–3	RMS energy	RMSm, **RMSd***, **RMSl***
	4	Low-Energy ratio	LE*m*
Articulation	5–6	Attack Time	AT*m*, At*d*
	7	Attack Slope	AS*m*
Rhythm	8	Event Density	ED*m*
	9	Fluctuation Peak	FP*m*
	10–11	Fluctuation Peak (mag.)	**FMm***, FC*m*
	12–13	Tempo	**Tm***, **Td***
	14–15	Pulse Clarity	PCm, **PCd***
Timbre	16–17	Spectral Centroid	SC*m*, SC*d*
	18	Spectral Spread	SS*m*
	19–20	Roughness	R*m*, R*d*
	21–22	Spectral Flux	**SFm***, SF*d*
	23–24	Regularity	RE*m*, RE*d*
	25–26	Inharmonicity	I*m*, I*d*
Pitch	27–28	Pitch	**Pm***, P*d*
	29–30	Chromagram (unwrapped) centr.	C*m*, C*d*
Tonal	31–32	Key Clarity	KC*m*, KC*d*
	33	Key Mode (majorness)	**Mm***
	34	HCDF	H*m*
	35	Spectral Entropy (oct. coll.)	SE*m*
Structure	36	Repetition (Spectrum)	RS*m*
	37	Repetition (Rhythm)	RR*m*
	38	Repetition (Tonal)	RT*m*
	39	Repetition (Register)	RG*m*

Note: In acronyms, $m =$ mean, $d =$ standard deviation, and $l =$ slope.

4.4 Regression model construction and cross-validation

To construct models capable of predicting emotions using either of the feature sets (the full or the reduced), stepwise regression was employed, since it provides an analytically transparent method of model building, and simplifies a model considerably by only including predictors that contribute to the regression equation above a certain threshold. Here the number of predictors was limited to five, as this proved to be the optimal number in Eerola et al. (2009), where several regression models with a different number of predictors and prediction rates had been compared using the *Akaike Information Criteria* (AIC). This previous study was also dealing with related analysis issues (prediction of emotions) and using similar materials. Using such optimization should tend to favour models that are simple but efficient and punish complex, inefficient models (Sawa, 1978). Fixing the number of components in each model should also stabilize comparisons across the datasets that have varying numbers of observations.

Using this analysis scheme, one regression model was constructed for valence and arousal per dataset. More importantly, each model thus obtained, was then applied to all other datasets in order to determine the model's ability to generalize and predict ratings elsewhere. Since the datasets represent four genres (classical, film, pop, mixed), this operation will at the same time be able to assess the genre-specificity of the models. This evaluation strategy will henceforth be called *exhaustive cross-validation,* as is illustrated in Figure 2, since each model is being tested with eight unrelated datasets. A conventional internal cross-validation (leave-one-out, k-fold, etc.) would have been problematic since the N differs across datasets and the appropriate selection of cross-validation principles would not have been easy either with traditional forms of regression, particularly the stepwise selection principle (Shao, 1993). However, a form of internal cross-validation will be applied later when Random Forest regression will be used to estimate the individual feature contributions in detail. The strength of the present cross-validation scheme lies in the fact that it is based on external, heterogeneous datasets, a procedure which is known to provide the most rigorous evaluation of the true generalizability of statistical models (Taylor, Ankerst, & Andridge, 2008).

5. Results

Firstly the success rate is reported for the regression models of each individual dataset. The regression is calculated using the full feature set, first for valence and then for arousal. The possible asymmetries in model performances between, and within, genres are then examined by collapsing the success rates accordingly. After this, the impact of feature sets on these results is briefly examined. In the final section, the possible underlying core features of emotions will be described using the full feature set and a separate regression analysis procedure.

5.1 Prediction of valence across genres

The results of the regression for valence with the full feature set, is summarized in Table 3. The rows refer to datasets used for training the model and the columns to datasets used for prediction. The diagonal therefore displays the R squared value for each model (the same dataset will have been used both for training and testing the model) and the shadowed area denotes the application of the model to datasets of the same genre. Looking at the diagonal, for each dataset first (median $R^2 = 0.73$), decent models could be built with the full feature set, using in each case, the five best features for the particular dataset. The only genre that presents any problems in this respect, is pop music, since the model fits are considerably lower for these two datasets (P1 and P2), and this may well relate to the different sample size of P2. If we look at the results in non-diagonal cells, even those of the same genre (shadowed cells), we get much worse results. In most cases, a model trained with one dataset is not able to predict any significant variance on another dataset, even when these are of the same genre (within genres the median R^2 value is 0.43). One possible reason for the considerably low success rate of both the pop music sets (P1 and P2) could be the effect of lyrics. Even though Ali and Peynircioğlu (2006) demonstrated that musical features have usually more emotional significance than lyrics, nearly all the excerpts in these two

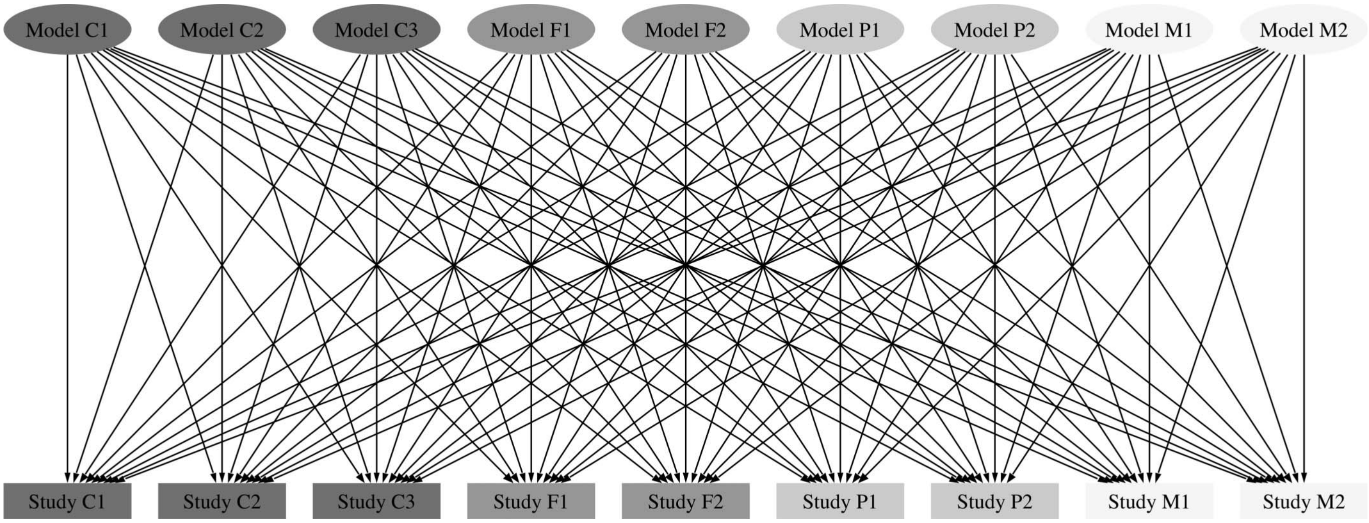

Fig. 2. The exhaustive cross-validation scheme used in comparing models across genres.

Table 3. Model performances (R^2) for valence across the datasets.

		Prediction									
		C1	C2	C3	F1	F2	P1	P2	M1	M2	Mdn
Training	C1	0.74*	0.03	0.01	0.00	0.16	0.00	0.03*	0.15	0.02	0.03
	C2	0.00	0.73*	0.13	0.17*	0.17	0.01	0.01*	0.41*	0.07	0.13
	C3	0.04	0.13	0.97*	0.05	0.41*	0.19*	0.02*	0.04	0.09	0.09
	F1	0.36	0.43*	0.56*	0.70*	0.50*	0.18*	0.00	0.48*	0.35*	0.43
	F2	0.02	0.17	0.26	0.27*	0.85*	0.05	0.00	0.09	0.05	0.10
	P1	0.08	0.26	0.84*	0.29*	0.37*	0.43*	0.05*	0.29	0.22*	0.29
	P2	0.10	0.02	0.27	0.00	0.18	0.01	0.51*	0.07	0.00	0.07
	M1	0.17	0.31*	0.73*	0.26*	0.39*	0.11	0.01*	0.88*	0.13*	0.26
	M2	0.09	0.22	0.62*	0.51*	0.45*	0.16*	0.00	0.49*	0.60*	0.45
	Mdn	0.09	0.22	0.56	0.26	0.39	0.11	0.01	0.30	0.09	

Note: *$p < 0.05$ using Bonferroni correction for multiple testing.

samples contained lyrics, many of which are explicitly related to an expression of emotional valence (e.g. 'I feel good', 'the black coach of sorrow has taken you'). Prediction rates are even more abysmal across genres (the cells with white background), yielding a median R^2 of only 0.16, which suggests that there are no clear underlying principles for predicting valence ratings with acoustic features. If the diagonal is included, only about a third of the models (39 out of 81) are statistically significant, at $p < 0.05$ level when a Bonferroni correction for multiple testing is used. Despite these rather meagre results for valence, some datasets, trained with mixed genres (M2) or film music (F1), seem to be tapping into more general aspects of valence. This can be inferred from the higher median prediction rates for models trained with these two datasets (in the column furthest to the right). On the other hand, the easiest dataset to predict, whatever the training dataset, seems to be C3 (with Haydn and Mozart examples), as it features the highest of the median prediction values in the bottom row of the table. This could be due to the high emotional clarity of these excerpts within the dataset (neatly representing the four quadrants in the affect space) but it is worth noting that the highest prediction of C3 is obtained with datasets that have highly stereotypical symphonic musical excerpts (F1, or M1), which suggest that an underlying textural similarity is part of the reason. Another plausible explanation lies in the clear and unambiguous distinction between major and minor keys in these examples. This musical feature, in the form of Key Mode in particular, will later turn out to be the central cue in identifying the core features for valence across the datasets.

To investigate the genre-specific aspects of the models in more detail, they can be summarized by calculating the mean prediction rates within genres (e.g. the mean R^2 across exclusively C1, C2, and C3) and between genres (e.g. the mean prediction of F1 and F1 with models C1, C2, and C3). The summary of these results is displayed in Table 4. Although the figures might seem somewhat inflated, as training and prediction with the same dataset have been left in, it is interesting to note the higher overall success rates for models trained with the mixed genre dataset (median row $R^2 = 0.38$), while the film music datasets are easiest to predict ($R^2 = 0.31$). The models trained with popular music, however, do not seem to generalize to other genres ($R^2 = 0.23$) and they are also the hardest genre to predict in general ($R62 = 0.07$).

5.2 Prediction of arousal across genres

Turning now to the prediction of arousal ratings, which has previously been an easier concept than valence to predict from acoustic features (Schubert, 2004; Leman et al., 2005; Coutinho & Cangelosi, 2009; Eerola et al., 2009), Table 5 provides the overview of the model performances within and across all datasets. The overall success is here significantly higher within the datasets (the median of the diagonal is 0.87) and also considerably better within genres than was the case with the valence ratings (the shaded cells, ranging from 0.07 to 0.81 with a median of 0.62). The lowest predicted genre is the model trained with P1 predicting P2 ($R^2 = 0.07$). There might be several reasons for the low fit. One could be related to cultural aspects since the ratings of the P1 were carried out by Taiwanese participants while the rest of the datasets originate from American or Central European laboratories. However, this possible explanation is not borne out, since datasets other than P2 are predicted moderately well. A more direct explanation could lie in the lyrics, which, in the case of P1, are quite emotional explicit. There seems to be reasonable success too, in the performance of the models beyond the genres. This is borne out by the decent median prediction rate (0.45) between the genres. For instance, the film soundtrack models (F1 and F2) predict the results in all other datasets except in P2 (the exceptional collection containing nearly 3000 annotated by three experts). Also the datasets containing stimuli from a mixture of genres (M1 and M2) perform well across the datasets. The generalizability index (Median column) also corroborates these interpretations. F1 has the highest generalizability index (0.69) followed by M2 (0.55) and F2 (0.51). Most tellingly, 74 out of 81 models are statistically significant at $p < 0.05$ level, even when a Bonferroni correction for multiple testing is used.

Predicting arousal between the genres, shown in Table 6, reveals few large asymmetries since the performance of the models between the genres is relatively constant (median $R^2 = 0.43$). This is in marked contrast to the results for valence, where the differences between genres were clearly evident.

To summarize these results, valence and arousal differed radically in terms of the overall performance and applicability of the models across genres. Valence remains inherently difficult to predict from acoustic features, as shown in previous studies. This is probably related to the concept of valence in the context of music, which is not straightforward. The dimensions of this bipolar scale are typically dubbed as positive and negative, but even negative affects are sometimes considered to be enjoyable within the context of music (Juslin & Laukka, 2004; Zentner et al., 2008) and therefore positive—which could explain the confusion within this concept. Also, the large variety in actual terms used for labelling the bipolar axes may have contributed to the discrepancies between the studies[3].

It is worth emphasizing that the accuracy of the models is lower between genres than within, which would make sense, and suggests that, in many cases, the same emotion must be perceived through a different constellation of musical features in each genre. However, since at least the arousal ratings could be predicted with models trained in almost any genre, there is an element of overlap in the musical features, as suggested by theoretical accounts (Balkwill & Thompson, 1999).

Table 4. Model performances (R^2) for valence across the four genres.

		Prediction				
		Classical	Film	Pop	Mixed	*Median*
Training	Classical	0.31	0.16	0.04	0.13	*0.15*
	Film	0.30	0.58	0.06	0.25	*0.28*
	Pop	0.26	0.21	0.25	0.15	*0.23*
	Mixed	0.36	0.40	0.07	0.52	*0.38*
	Median	*0.31*	*0.31*	*0.07*	*0.20*	

5.3 Comparison of the feature sets

With respect to the reduced feature set with nine features chosen by PCA, the regression analysis was replicated for both valence and arousal using the same analysis options (the five best predictors for each model using the statistical selection criteria). Since the results do not deviate significantly from the results obtained with the full feature set, a rundown of the model performance within and between the feature sets and concepts is made here in terms of the model performance for within and between genres. The within refers to all models and predictions concerning C1, C2, and C3, i.e. the shaded cells in Tables 3 and 5) and the between genres prediction refers to the median value of the white cells in the previously mentioned tables.

For valence, the median within genre prediction rate was slightly lower with the reduced feature set (0.33) in comparison with the full set (0.43) but the between genre success was somewhat higher (0.20) than with the full feature set (0.16). In the case of arousal, the reduced feature performed less well in the within comparison (0.47, when the full achieved a median of 0.62). In the between genres comparison the situation was similar, 0.23 for reduced set and 0.43 for full feature set. Generally speaking, the prediction rates with the reduced feature set are tolerable for valence, which was more difficult to predict with any set of features. In arousal however, the performance plummets more significantly (15% and 20%) implying that the reduced feature set, although capable of capturing 80% of the variance in the original feature matrix, lacks some of the features (particularly the structural ones) that contribute to arousal. A more precise diagnosis of these critical features will be carried out next.

5.4 Identification of the core features of emotions

To evaluate the individual contribution of the features, Random Forest (RF) regression was employed (Breiman, 2001). RF is designed to overcome the problem of overfitting; bootstrapped samples are drawn to construct

[3] For instance, in Gomez and Danuser (2004), graphical and non-verbal Self-Assessment Manikin (SAM, Bradley & Lang, 2000) was used to collect the stimulus evaluations. In Dibben (2004), a variety of terms such as *agitated-excitement*, *peacefulness*, *happiness*, and *sadness* were used but reduced to valence and arousal by taking the mean across the concepts organized into affect quadrants. In the study of Gosselin et al. (2005), the terms were *relaxing-stimulating* and *pleasant-unpleasant* for valence whereas Eerola and Vuoskoski (2011) used several labels for the extremes of the two bipolar scales—e.g. *positive*, *good*, and *pleasant* for positive valence and *sleepy*, *tired*, and *drowsy* for low arousal.

Table 5. Model performances (R^2) for arousal across the datasets.

		Prediction									
		C1	C2	C3	F1	F2	P1	P2	M1	M2	Mdn
Training	C1	0.96*	0.57*	0.38	0.38*	0.21*	0.45*	0.09*	0.60*	0.25*	0.38
	C2	0.57*	0.88*	0.47*	0.36*	0.16	0.55*	0.12*	0.64*	0.13*	0.47
	C3	0.81*	0.62*	0.94*	0.52*	0.29*	0.13*	0.00	0.44*	0.41*	0.45
	F1	0.75*	0.46*	0.78*	0.83*	0.17	0.75*	0.16*	0.69*	0.36*	0.69
	F2	0.51*	0.15	0.56*	0.47*	0.87*	0.53*	0.05*	0.51*	0.20*	0.51
	P1	0.71*	0.46*	0.10	0.43*	0.26*	0.86*	0.07*	0.81*	0.34*	0.43
	P2	0.70*	0.44*	0.44*	0.40*	0.31*	0.63*	0.29*	0.76*	0.26*	0.44
	M1	0.62*	0.35*	0.65*	0.30*	0.27*	0.73*	0.24*	0.94*	0.23*	0.35
	M2	0.72*	0.68*	0.53*	0.53*	0.45*	0.35*	0.00	0.75*	0.55*	0.55
	Mdn	0.71	0.46	0.53	0.43	0.27	0.55	0.09	0.69	0.26	

Note: *$p < 0.05$ using Bonferroni correction for multiple testing.

Table 6. Model performances (R^2) for arousal across the four genres.

		Prediction				
		Classical	Film	Pop	Mixed	*Median*
Training	Classical	0.69	0.32	0.22	0.41	*0.37*
	Film	0.54	0.59	0.37	0.44	*0.49*
	Pop	0.47	0.35	0.46	0.54	*0.47*
	Mixed	0.59	0.40	0.35	0.62	*0.50*
	Median	*0.57*	*0.38*	*0.36*	*0.49*	

multiple trees (here set to 5000), which have randomized subsets of predictors, hence eliminating the need for separate cross-validation. Another benefit is that RF does not require reduction of the predictor space prior to regression. The main advantage of the method, however, is to be able to reliably measure the importance of variables (Archer & Kimes, 2008) by using out-of-bag samples to estimate the error rate and thereby the variable's importance. Finally, RF provides transparent models since the output is dependent only on one input variable, namely the number of predictors chosen randomly at each node, heuristically set to 10 in the present study. Most applications of RF have demonstrated that this technique has improved accuracy in comparison to other supervised learning methods (e.g. Breiman, 2001; Svetnik et al., 2003).

Table 7 shows increments to the variance explained in the regression model provided by each feature of the full feature set. In this notation, +5 refers to an increase of 0.05 in R^2 value. The results are shown separately for each genre, in which the datasets corresponding to the particular genre have been pooled together before the analysis. The overall model performance, also displayed in the table, shows higher prediction rates across genres than in the previous summary (Table 4), with the exception of classical music (R^2 only 0.21 in comparison with the previous 0.31). The improvement of the results may stem from (a) pooling the datasets together, and (b) using 10 features instead of five as in the previous analysis, and (c) by using a more effective form of regression. These results, nevertheless, further consolidate the observations drawn in the previous analysis, since similar magnitudes of variance explained were obtained. What is new, however, lies in the individual contributions of the features.

For a quick summary, each feature has been given a ranking (last column). The most effective feature, *Key Mode*, has a dominant contribution to valence across musical genres, although in pop music, this effect is diminished (+5) compared with the other genres (+15 to +29). The importance of mode is not surprising, since it has been implicated previously in numerous empirical studies of music and emotions (Hevner, 1936; Juslin, 1997a; Dalla Bella et al., 2001; Vieillard et al., 2008; Fritz et al., 2009; Livingstone et al., 2010). Nevertheless, it has scarcely been so explicitly present in studies which extract features from audio-based representations. In second place is *Pulse Clarity*, a measure of how easy it would be to tap with the beat of the excerpt, which has not previously been suggested as an emotion candidate. Although the overall contribution of pulse clarity seems to be strong (+19%), it is most important in pop music and in mixed genres, suggesting a fairly high genre specificity. The third ranked feature is deviation within the fluctuation magnitude, which measures the periodicity over different spectral bands (Pampalk et al., 2002). It is similarly related to pop music and mixed genres and not so important for classical or film music genres. Two features based on Repetition detection across the excerpt are placed in fourth and fifth positions. These

Table 7. Contribution of individual predictors for valence across pooled excerpts across genre and random forest regression.

Domain	No.	Name	Class.	Film	Pop	Mix	All	Rank
		N	59	150	2658	95	2962	
		R^2	0.21	0.69	0.44	0.56	0.57	
Dynamics	1–3	RMS energy			+7/+7		+11/+9	10
	4	Low–Energy		+6	+6			
Articul.	5–6	Attack Time				−/+5	−/+6	
	7	Attack Slope			+7		+9	
Rhythm	8	Event Dens.			+11		+10	
	9	Fluct. Peak			+15	+6	+13	6
	10–11	Fluct. Mag.	−/+ 5		+9/+28	+13/−	+13/+18	3
	12–13	Tempo			+11/+17		+8/+12	9
	14–15	Pulse Clar.			+ 6/−	+9/−	+19/−	2
Timbre	16–17	Sp. Centroid		−/+6	+5/+6		+11/+12	8
	18	Sp. Spread			+7		+11	
	19–20	Roughness			+6/+7		+9/+8	
	21–22	Sp. Flux			+13/+8		+9/+10	
	23–24	Regularity		+15/+8			+11/+7	
	25–26	Inharmon.		−/+10	+13/+8	+7/−	+7/+10	
Pitch	27–28	Pitch		−/+7	+5/+8		−/+10	
	29–30	Chromagram		+ 7/−			+8/−	
Tonal	31–32	Key Clar.		+ 14/−			+12/−	7
	33	Key Mode	+15	+29	+5	+22	+36	1
	34	HCDF		+8	+7		+10	
	35	Sp. Entropy			+8	+8	+7	
Structure	36	Rep. (Sp.)		+7	+8		+8	
	37	Rep. (Rh.)			+5	+5	+7	
	38	Rep. (Ton.)		+11	+8		+16	4
	39	Rep. (Reg.)		+7	+11		+16	5

Note: Only increments over 5% in variance explained are shown.

both suggest that highly valenced music has static tonal and registral content.

It is noteworthy that four of the top 10 ranked features represent rhythm, while features of pitch and articulation are not present. Although this is at odds with some of the earlier studies (e.g. Ilie & Thompson, 2006), one of the main reasons for this might be due to the difference concerns in musical materials, which in the present case often do not have a clearly defined melody and harmony. Past studies using symbolic manipulation, or production studies with clearly defined melodies have, in contrast, favoured these features (Scherer & Oshinsky, 1977; Juslin, 1997b; Juslin, 2000).

Table 8 summarizes the role of all 39 musical features in modelling arousal across genres. Again, the variance explained by the RF regression models is largely the same as in the previous analysis, but with minor improvements for most genres. The only exception to this is in mixed genre, which was actually better predicted in the stepwise regression (0.62 as opposed to 0.47). Again the reasons for this may lie either in the pooling of data or the different regression technique. Since the variations in prediction rates were indeed minor, the previous results for arousal are simply vindicated by this second analysis.

The ranked list of most important features for arousal is not restricted to a particular domain of feature (like valence was, in the case of rhythmic features), but spread out over all of them and more than five features are able to, single-handedly, deliver an increment of 15% to the regression equation. Pulse clarity and inharmonity are leading the list. Also spectral entropy, which is closely related to inharmonicity, together with register and fluctuation magnitude are among the top five ranked features. Although the beta coefficients of the predictors are not shown, it could be inferred that highly arousing music has high spectral entropy and inharmonicity and particularly clear pulse and rhythms (for model coefficients, see Schubert, 2004; Eerola et al., 2009). In contrast to individual features in the regression models for valence, those in the models for arousal operate more consistently across the genres. Most features contribute to at least three genres, and there are no massive discrepancies between the rates of increment between these. In fact, some of the most consistent features, such

Table 8. Contribution of individual predictors for arousal across pooled excerpts across genre and random forest regression.

Domain	No.	Name	Class.	Film	Pop	Mix	All	Rank
		N	59	150	2658	95	2962	
		R^2	0.69	0.75	0.58	0.47	0.65	
Dynamics	1–3	RMS energy		+6/−	+8/+6		+9/+10	
	4	Low–Energy			+8	+6		
Articul.	5–6	Attack Time		+6/+8			+7/+10	
	7	Attack Slope		+7	+12		+13	
Rhythm	8	Event Dens.	+5		+12		+11	
	9	Fluct. Peak			+7		+9	
	10–11	Fluct. Mag.		+13/−	+23/+17	+5/−	+15/+12	6
	12–13	Tempo	−/+5		−/+14		+6/+10	
	14–15	Pulse Clar.	+7/+5	+6/−	+12/−	+9/−	+21/−	1
Timbre	16–17	Sp. Centroid		−/+6	+18/+6		+13/+13	
	18	Sp. Spread			+12		+10	
	19–20	Roughness		+8/+9	+7/+7	+5/−	+9/+9	
	21–22	Sp. Flux	+6/−	+13/+8	+10/+8	+9/−	+14/+12	7
	23–24	Regularity		+12/+8	−/+5		+12/+9	
	25–26	Inharmon.	+9/+8	+8/+12	+22/+11		+20/+14	2, 9
Pitch	27–28	Pitch		−/+13	+8/+8	−/+7	−/+16	5
	29–30	Chromagram		−/+5			+9/+6	
Tonal	31–32	Key Clar.					+7/−	
	33	Key Mode						
	34	HCDF			+14			
	35	Sp. Entropy	+12	+10	+15	+8	+18	3
Structure	36	Rep. (Sp.)	+8	+12	+9	+5	+13	10
	37	Rep. (Rh.)	+11	+17	+10	+16	+18	4
	38	Rep. (Ton.)		+6	+11		+14	8
	39	Rep. (Reg.)		+5	+7		+11	

Note: Only increments over 5% in variance explained are shown.

as spectral entropy, spectral flux, and pulse clarity and the features based on novelty estimation, operate almost identically across all the genres.

Previous literature (Schubert, 2004; MacDorman, 2007; Yang et al., 2007; Eerola et al., 2009) would have predicted that tempo, articulation and loudness would have been amongst the most important features of arousal. Here they are not insignificant, but due to the difference in musical materials and possibly due to the availability of better features (the top five), they do not have such an important role in explaining the ratings of arousal.

6. Conclusions

A systematic study was presented of the genre-specific aspects of acoustic features that are used in the perception of emotions in music. Nine datasets, coming from laboratories spanning six countries, and having entirely different pools of participants were deployed. At the same time, identical features were extracted from each, enabling a comparison of musical factors which contribute to valence and arousal ratings. Moreover, robust model building principles were employed, in which overfitting was minimized and individual feature contribution was assessed by an advanced regression technique.

The results indicated that the ratings of valence could not be consistently predicted across musical genres using the acoustic features, since only about a third of the regression models could explain valence ratings in other datasets. Arousal, on the other hand, did not demonstrate such a genre-specific property since nearly all of the models could predict a significant portion of variance in other datasets spanning any genre. These results corroborate those findings of previous studies that have observed how valence is more difficult and arousal is considerably easier to model using acoustic features. However, the results also extend our understanding of the genre-specificity of emotions since none of the previous studies have explicitly used several datasets representing different genres to estimate whether the same features may transgress the boundaries set by musical genres. The implication of the results obtained here is that musical genres do matter for valence, since large differences were observed, whereas genres are less important for arousal.

It is unlikely that the heterogeneity of the nine datasets is the reason for the low success rate in modelling valence through acoustic features, since arousal could be predicted across the same materials. Of course, the datasets have entirely different emphases in the rating tasks since the listener backgrounds, listening contexts, and instructions varied across the studies. But, if anything, it is likely that such a variation actually makes the results stronger and more robust, since the generalizability of the models have truly been put to test. The fact that a model is being tested and found accurate across diverse settings, also called *transportability,* is the most valuable way of establishing external validity for the model (Justice et al., 1999), and also the only way to build generalizability into the modelling process (Altman & Royston, 2000).

This study implied that a set of core features seem to operate across musical genres. These features, however, were not the pitch or tonal elements usually mentioned in the literature, but more rhythmic aspects of the music. The reasons for these discrepancies may be related to (a) inaccuracies in feature extraction, (b) differences in musical materials, (c) genre-related differences, (d) differences in the emotion concepts between the genres and studies. Out of these possibilities, the differences related to musical materials is something that is worth considering in more detail. In many of the classic studies about the musical features of emotions, the material has been simplified, either by having instrumentalists perform melodies in different expressions (e.g. Juslin, 1997a, 1997b, 2000), which therefore emphasizes so-called performance features (such as articulation, tempo, and dynamics), or in factorial manipulations of two or three selected features (Dalla Bella et al., 2001; Ilie & Thompson, 2006). Both of these types of study focus on features that are easy to manipulate in a symbolic domain, such as mode, tempo, and register, thus probably over-emphasizing their role in emotions. Therefore it is feasible that real recordings from a variety of genres brings out other types of features, which in this case, pertain to overall timbre (e.g. inharmonicity and spectral entropy) and rhythm (e.g. pulse clarity and fluctuation magnitude). These new musical features are excellent candidates for systematic, factorial studies of emotion features since such parameters have not previously been fully utilized in causal studies of emotional features (e.g. timbre has only once been the subject of such a study; Scherer & Oshinsky, 1977). One way of explaining the relatively large number of features capable of contributing to emotions in music has been provided by the brunswikian lens model (Juslin, 2000), according to which, communication may be accurate using multiple cues although the relative contribution of the cues will depend on the context (for an overview, see Juslin & Scherer, 2005).

Computational models capable of predicting emotions expressed by music are already valuable in content-based retrieval applications of music (e.g. Casey et al., 2008) and in educational contexts (e.g. Juslin, Karlsson, Lindstrom, Friberg, & Schoonderwaldt, 2006). Yet there are other intriguing prospects for such models. For instance, an emotionally accurate description of the stimulus qualities would be indispensable in questions such as explaining the individual differences in emotions (Vuoskoski & Eerola, 2011), for testing the role of arousal in performance of non-musical tasks (e.g. Schellenberg, 2005), for understanding the subjective experience of chills (e.g. Blood & Zatorre, 2001), or for understanding aesthetic experiences in music in general (e.g. Müller, Höfel, Brattico, & Jacobsen, 2010). These are all fundamental issues in music cognition that have been only partially solved with little focus on the qualities of stimuli. Moreover, computational models of emotions may also be helpful in finding unfamiliar for use in the study of emotions, when such a model is applied to very large collections from sources such as *Last.fm* or the *Naxos collection* (Downie & Futrel, 2005; Casey et al., 2008). Alternatively, studying the preferred stimuli of participants could take into account musical preferences, naturally expressed in terms of genres (e.g. Rentfrow & Gosling, 2003). Subsequently, the genre-specificity of emotions could be better taken into account in future studies since different models could be utilized for different genres, or different excerpts may be chosen from different genres depending on the preferences of the participants.

In future studies of the genre-specifics of emotions, it would be crucial to combine the current, multi-genre approach with special populations, such as children, people from different cultures or patients with specific neural pathologies, in order to further isolate the musical factors and their dependence on the learned patterns of the culture. These combinations would allow us to determine more specifically what aspects of expressed emotion are mostly the product of learning, and what aspects belong to our shared understanding of how emotions are expressed in music.

References

Ali, S.O., & Peynircioğlu, Z.F. (2006). Songs and emotions: Are lyrics and melodies equal partners? *Psychology of Music*, *34*(4), 511–534.

Al-Kandari, N., & Jolliffe, I. (2001). Variable selection and interpretation of covariance principal components. *Communications in Statistics – Simulation and Computation*, *30*(2), 339–354.

Altenmüller, E., Schuermann, K., Lim, V.K., & Parlitz, D. (2002). Hits to the left, flops to the right: Different emotions during listening to music are reflected in cortical lateralisation patterns. *Neuropsychologia*, *40*(13), 2242–2256.

Altman, D., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*(4), 453–473.

Archer, K.J., & Kimes, R.V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), 2249–2260.

Aucouturier, J.J., & Pachet, F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research*, *32*(1), 83–93.

Balkwill, L.L., & Thompson, W.F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, *17*(1), 43–64.

Balkwill, L.L., Thompson, W.F., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners 1. *Japanese Psychological Research*, *46*(4), 337–349.

Banse, R., & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kögl, B. (2006). Aggregate features and ADA BOOST for music classification. *Machine Learning*, *65*(2), 473–484.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, *19*(8), 1113–1139.

Blood, A.J., & Zatorre, R.J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceeding of National Academy of Sciences, USA*, *98*(20), 11818–11823.

Bradley, M.M., & Lang, P.J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, *37*(2), 204–215.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bresin, R., & Friberg, A. (2000). Emotional coloring of computer-controlled music performances. *Computer Music Journal*, *24*(4), 44–63.

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

Coutinho, E., & Cangelosi, A. (2009). The use of spatio-temporal connectionist models in psychological studies of musical emotions. *Music Perception*, *27*(1), 1–15.

Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*(3), B1–10.

Dibben, N. (2004). The role of peripheral feedback in emotional experience with music. *Music Perception*, *22*(1), 79–115.

Downie, J.S., & Futrel, J. (2005). Terascale music mining. In *SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing* (p. 71). Washington, DC: IEEE Computer Society.

Eerola, T., & Ferrer, R. (2009). Setting the standards: Normative data on audio-based musical features for musical genres [Poster]. In *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music, ESCOM*, Jyväskylä, Finland.

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR 2009)*, Kobe, Japan, pp. 621–626.

Eerola, T., & Vuoskoski, J.K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49.

Evans, P., & Schubert, E. (2008). Relationships between expressed and felt emotions in music. *Musicae Scientiae*, *12*(1), 75–99.

Feng, Y., Zhuang, Y., & Pan, Y. (2003). Popular music retrieval by detecting mood. In *Proceedings of the 26th Annual International ACM SIGIR Conference*, Toronto, Canada, pp. 375–376.

Foote, J., & Cooper, M. (2003). Media segmentation using self-similarity decomposition. *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, *5021*, 167–175.

Frijda, N.H. (2007). What might an emotion be? Comments on the comments. *Social Science Information*, *46*, 433–443.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., & Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology*, *19*(7), 573–576.

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different. *Musicae Scientiae*, *2001–2002*, 123–147.

Gabrielsson, A., & Lindström, E. (2010). The influence of musical structure on emotional expression. In P.N. Juslin & J.A. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research and Applications* (pp. 223–248). New York: Oxford University Press.

Gagnon, L., & Peretz, I. (2000). Laterality effects in processing tonal and atonal melodies with affective and nonaffective task instructions. *Brain & Cognition*, *43*(1–3), 206–210.

Gerardi, G., & Gerken, L. (1995). The development of affective responses to modality and melodic contour. *Music Perception*, *12*, 279–279.

Gjerdingen, R., & Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, *37*(2), 93–100.

Gomez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, *18*(3), 294–304.

Gomez, P., & Danuser, B. (2004). Affective and physiological responses to environmental noises and music. *International Journal of Psychophysiology*, *53*(2), 91–103.

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, *7*(2), 377–387.

Gosselin, N., Peretz, I., Johnsen, E., & Adolphs, R. (2007). Amygdala damage impairs emotion recognition from music. *Neuropsychologia*, *45*, 236–244.

Gosselin, N., Peretz, I., Noulhiane, M., Hasboun, D., Beckett, C., Baulac, M., & Samson, S. (2005). Impaired recognition of scary music following unilateral temporal lobe excision. *Brain*, *128*(3), 628–640.

Gosselin, N., Samson, S., Adolphs, R., Noulhiane, M., Roy, M., Hasboun, D., Baulac, M., & Peretz, I. (2006). Emotional responses to unpleasant music correlates with damage to the parahippocampal cortex. *Brain*, *129*(10), 2585–2592.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate Data Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, Santa Barbara, CA, pp. 26–31.

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

Hevner, K. (1937). The affective value of pitch and tempo in music. *The American Journal of Psychology*, *49*(4), 621–630.

Hu, X., & Downie, J.S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the International Symposium on Music Information Retrieval*. Vienna: Austrian Computer Society.

Hu, X., Downie, J.S., Laurier, C., Bay, M., & Ehmann, A.F. (2008). The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Philadelphia, PA, pp. 462–467.

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

Ilie, G., & Thompson, W. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319–329.

Izard, C.E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280.

Juslin, P.N. (1997a). Emotional communication in music performance: a functionalist perspective and some data. *Music Perception*, *14*(4), 383–418.

Juslin, P.N. (1997b). Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy. *Musicae Scientiae*, *1*(2), 225–256.

Juslin, P.N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797–1813.

Juslin, P.N., Karlsson, J., Lindstrom, E., Friberg, A., & Schoonderwaldt, E. (2006). Play it again with feeling: Computer feedback in musical communication of emotions. *Journal of Experimental Psychology Applied*, *12*(2), 79–94.

Juslin, P.N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin*, (129), 770–814.

Juslin, P.N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238.

Juslin, P.N., & Scherer, K.R. (2005). Vocal expression of affect. In J.A. Harrigan, R. Rosenthal, & K.R. Scherer (Eds.), *The New Handbook of Methods in Nonverbal Behavior Research* (pp. 65–135). Oxford, MA: Oxford University Press.

Juslin, P.N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559–575.

Justice, A., Covinsky, K., & Berlin, J. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, *130*(6), 515–524.

Kallinen, K. (2005). Emotional ratings of music excerpts in the western art music repertoire and their self-organization in the Kohonen neural network. *Psychology of Music*, *33*(4), 373–393.

Kallinen, K., & Ravaja, N. (2004). Emotion-related effects of speech rate and rising vs. falling background music melody during audio news: The moderating influence of personality. *Personality and Individual Differences*, *37*(2), 275–288.

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation, and optimization. In *ISMIR 2008 International Conference on Music Information Retrieval*, Philadelphia, PA, pp. 521–526.

Lartillot, O., & Toiviainen, P. (2007). MIR in Matlab (II): A toolbox for musical feature extraction from audio. In S. Dixon, D. Bainbridge, & R. Typke (Eds.), *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 237–244). Vienna: Österreichische Computer Gesellschaft.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 261–268). Saksa: Springer.

Leman, M., Lesaffre, M., & Tanghe, K. (2001). *A Toolbox for Perception-based Music Analysis*. Ghent: IPEM – Dept. of Musicology, Ghent University.

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research*, *34*(1), 39–67.

Li, T., & Ogihara, O.M. (2003). Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval* (pp. 239–240). Baltimore: John Hopkins University.

Lindström, E. (2003). The contribution of immanent and performed accents to emotional expression in short tone sequences. *Journal of New Music Research*, *32*(3), 269–280.

Livingstone, S.R., Muhlberger, R., Brown, A.R., & Thompson, W.F. (2010). Changing musical emotion through score and performance with a computational rule system. *Computer Music Journal*, *34*(1), 41–65.

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

MacDorman, K. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, *36*(4), 281–299.

Marques G., Langlois T., Gouyon F., Lopes M., & Sordo M. (2011). Short-term feature space and music genre classification. *Journal of New Music Research*, *40*(2), 127–137.

Müller, M., Höfel, L., Brattico, E., & Jacobsen, T. (2010). Aesthetic judgments of music in experts and laypersons–an ERP study. *International Journal of Psychophysiology*, *76*(1), 40–51.

Nater, U.M., Abbruzzese, E., Krebs, M., & Ehlert, U. (2006). Sex differences in emotional and psychophysiological responses to musical stimuli. *International Journal of Psychophysiology*, *62*(2), 300–308.

Pampalk, E., Rauber, A., & Merkl, D. (2002). Content-based organization and visualization of music archives. In *Proceedings of the 10th ACM International Conference on Multimedia*, Juan les Pins, France, pp. 579–585.

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, *68*(2), 111–141.

Purwins, H., Grachten, M., Herrera, P., Hazan, A., Marxer, R., & Serra, X. (2008). Computational models of music perception and cognition II: Domain-specific music processing. *Physics of Life Reviews*, *5*(3), 169–182.

Purwins, H., Herrera, P., Grachten, M., Hazan, A., Marxer, R., & Serra, X. (2008). Computational models of music perception and cognition I: The perceptual and cognitive processing chain. *Physics of Life Reviews*, *5*(3), 151–168.

Rawlings, D., & Leow, S.H. (2008). Investigating the role of psychoticism and sensation seeking in predicting emotional reactions to music. *Psychology of Music*, *36*(6), 269–287.

Rentfrow, P.J., & Gosling, S.D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*(6), 1236–1256.

Ritossa, D.A., & Rickard, N.S. (2004). The relative utility of 'pleasantness' and 'liking' dimensions in predicting the emotions expressed by music. *Psychology of Music*, *32*(1), 5–22.

Russell, J. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*, 102–102.

Russell, J., Weiss, A., & Mendelsohn, G. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica: Journal of the Econometric Society*, *46*, 1273–1291.

Schellenberg, E. (2005). Music and cognitive abilities. *Current Directions in Psychological Science*, *14*(6), 317–320.

Scherer, K.R., & Oshinsky, J.S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception*, *21*(4), 561–585.

Schuller, B., Dorfner, J., & Rigoll, G. (2010). Determination of nonprototypical valence and arousal in popular music: Features and performances. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 1–20.

Sethares, W. (1998). *Tuning, Timbre, Spectrum, Scale*. Berlin: Springer-Verlag.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Silla Jr, C.N., Kaestner, C.A.A., & Koerich, A.L. (2007). Automatic music genre classification using ensemble of classifiers. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Montréal, Quebec, pp. 1687–1692.

Skowronek, J., McKinney, M., & Par, S. van de. (2007). A demostrator for automatic music mood estimation. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 345–346). Vienna: Österreichische Computer Gesellschaft.

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., & Feuston, B. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, *43*(6), 1947–1958.

Taylor, J.M.G., Ankerst, D.P., & Andridge, R.R. (2008). Validation of biomarker-based risk prediction models. *Clinical Cancer Research*, *14*(19), 5977–5983.

Tolonen, T., & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, *8*(6), 708–716.

Tzanetakis, G., & Cook, P. (2000). Marsyas: A framework for audio analysis. *Organised Sound*, *4*(03), 169–175.

Tzanetakis, G., & Cook, P. (2002a). *Manipulation, Analysis and Retrieval Systems for Audio Signals*. Princeton, NJ: Princeton University.

Tzanetakis, G., & Cook, P. (2002b). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302.

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *22*(4), 720–752.

Vuoskoski, J.K., & Eerola, T. (2011). The role of mood and personality in the perception of emotions represented by music. *Cortex*, *47*(9), 1099–1106.

Wedin, L. (1972). A multidimensional study of perceptual-emotional qualities in music. *Scandinavian Journal of Psychology*, *13*(4), 241–257.

Wu, T.-L., & Jeng, S.-K. (2006). Automatic emotion classification of musical segments. In *Proceedings of the 9th International Conference on Music Perception and Cognition*, Bologna, 2006.

Yang, Y., Lin, Y., Su, Y., & Chen, H. (2007). Music emotion classification: A regression approach. In *Proceedings of the IEEE International Confererence on Multimedia*, Beijing, China, pp. 208–211.

Zentner, M.R., & Eerola, T. (2010). Self-report measures and models. In P.N. Juslin & J.A. Sloboda (Eds.), *Handbook of Music and Emotion* (pp. 187–221). Boston, MA: Oxford University Press.

Zentner, M.R., Grandjean, D., & Scherer, K.R. (2008). Emotions evoked by the sound of music: Differentiation, classification, and measurement. *Emotion*, *8*(4), 494–521.