TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 4 (2012) 607–624
Copyright © 2012 Cognitive Science Society, Inc. All rights reserved.
ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2012.01188.x

Modeling Listeners' Emotional Response to Music

Tuomas Eerola

Department of Music, Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä

Received 18 October 2010; received in revised form 21 May 2011; accepted 26 August 2011

Abstract

An overview of the computational prediction of emotional responses to music is presented. Communication of emotions by music has received a great deal of attention during the last years and a large number of empirical studies have described the role of individual features (tempo, mode, articulation, timbre) in predicting the emotions suggested or invoked by the music. However, unlike the present work, relatively few studies have attempted to model continua of expressed emotions using a variety of musical features from audio-based representations in a correlation design. The construction of the computational model is divided into four separate phases, with a different focus for evaluation. These phases include the theoretical selection of relevant features, empirical assessment of feature validity, actual feature selection, and overall evaluation of the model. Existing research on music and emotions and extraction of musical features is reviewed in terms of these criteria. Examples drawn from recent studies of emotions within the context of film soundtracks are used to demonstrate each phase in the construction of the model. These models are able to explain the dominant part of the listeners' self-reports of the emotions expressed by music and the models show potential to generalize over different genres within Western music. Possible applications of the computational models of emotions are discussed.

Keywords: Computational model; Affect; Emotion; Dimensions; Acoustic analysis; Feature extraction

One of the most intriguing aspects of music is its power to arouse emotional experiences in listeners. This process is undoubtedly complex, as it is related to aspects in the communication of emotions such as different emotion induction mechanisms (Juslin & Västfjäll, 2008), models of emotions (Zentner & Eerola, 2010), differences in individuals (Kallinen & Ravaja, 2004), and musical expectations (Huron, 2006). Emotions may also either be either perceived or felt by the listener. These two different processes are similar in many aspects (Evans & Schubert, 2008; Vieillard et al., 2008), but in terms of constructing useful

Correspondence should be sent to Tuomas Eerola, Department of Music, Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, 40014 Jyväskylä, Finland. E-mail: tuomas.eerola@jyu.fi computational models, it is better to focus on emotions communicated by music. The emotions that are readily perceived in music rely more on musical features than felt emotional experiences, which are more prone to individual associations and contextual effects (Gabrielsson, 2002; Juslin & Västfjäll, 2008).

What are the emotions typically expressed in music? Three types of emotion models have been employed in the field: discrete, dimensional, and music-specific (reviewed in Eerola & Vuoskoski, in press). According to the discrete emotion model, commonly used in non-musical contexts, all emotions can be derived from a limited number of universal and innate basic emotions such as fear, anger, disgust, sadness, and happiness. In music-related studies, however, although many of these have been found to be appropriate (Juslin & Laukka, 2004), certain emotions have often been replaced by more appropriate ones. For example, disgust is replaced by something more likely to be expressed in music, such as tenderness or peacefulness. The two-dimensional circumplex model (Russell, Weiss, & Mendelsohn, 1989), which is the most representative of the dimensional models, proposes that all affective states arise from two independent neurophysiological systems: one related to valence and the other to arousal. This particular model has received a great deal of attention in music and emotion studies despite a number of drawbacks. For instance, it is unable to represent mixed emotions, and so several alternative, presumably better, dimensional models have been proposed (e.g., Schimmack & Reisenzein, 2002; Thayer, 1989). One such novel idea for a model has recently been proposed by Zentner and his colleagues, called GEMS (2008), which has nine factors. This model emphasizes positive emotions but since it deals mainly with felt emotions, and since there is not yet enough data to begin comprehensive modelling of it, the emphasis in the present work will be on two-dimensional models of emotions.

Over the past 80 years, there have been many approaches to the problem of teasing apart the elements of music which endow it with emotional meaning (e.g., Gabrielsson & Lindström, 2010). This work, reviewed later in more detail, has distilled these elements in such a way that it is replicable, data-rich, and pertinent to both to expert and non-expert listeners.

The intention of this article is to draw attention to contemporary accounts of these endeavors by covering issues related to the computational modelling of emotions which use audio as underlying representation of music. Attention will be given to the different processes involved in attempting to pinpoint the essential acoustic, perceptual, and musical aspects that underlie the communication of emotions in music. Examples drawn from empirical studies based on film soundtracks will be used to highlight how musical features can be extracted, validated, and connected to emotions. Finally, an evaluation of the constructed emotion model is carried out across several unrelated datasets, which shows a promising degree of success.

1.The role of computational models in understanding music processing

The epistemological concerns of attempting to understand which musical features contribute to emotions in music are related to the three cornerstones of cognitive science, and they are hallmarks of the approach in general: (a) empirical observation, (b) formalization (e.g., computational modeling), and (c) cognitive relevancy (Fodor, 2001). These issues have been discussed in detail within the context of music cognition by Marc Leman (2003), Clarke and Cook (2004), and particularly eloquently by Henkjan Honing (2004, 2006b). The first cornerstone reflects the desire to have some ground theories for systematic observations. Such observations have not always been possible in music research but the advent of contemporary computers and standard protocols for music representation (such as MIDI) have facilitated this approach in music research (e.g., Rink, 1995). Also the adoption of methods from experimental psychology and standards used in the testing of statistical hypothesis have added methodological rigor to conclusions drawn from such observations.

The second cornerstone concerns the falsifiability of scientific models. This formalization favors models that are transparent, such as most computational models, since their very nature makes them explicit, replicable, and testable (e.g., Eck, 2002; Temperley, 2007). Such models are also scalable, that is, they can be easily used in large-scale comparisons, which also enables them to be used to predict novel observations. As an added bonus, it is often relatively easy to develop a version of the computational model that can be applied in commercial, artistic, and educational contexts. This leads to developments in the field becoming more widely known, and more important, allows the field to directly contribute to social and musical practices (Leman, 2008a).

The third cornerstone is related to the cognitive revolution, in which computational modeling has played a central role (Fodor, 2001). During the 1990s, computational modeling established itself as a viable option within the music research community and was applied, for instance, to problems of rhythm (summarized in Clarke, 1999), music performance (e.g., Gabrielsson, 1999), and the perception of tonality (e.g., Leman, 2000). Although the last decade has consolidated this development, common paradigms and model architectures have nevertheless not yet appeared since their evaluation may often lack guiding principles (Honing, 2006a). The central point, however, is that cognitive approach requires cognitive plausibility for any of the features or processes involved in the model. This is in sharp contrast with the practical concerns over computational efficiency, which is often prioritized in an engineering approach (Leman, 2008b; Posner, 1993), although other, more fundamental concerns for this approach have been raised within the engineering field as well (see Wiggins, 2009). In cognitive approach, for example, features that are known to be perceptually relevant for listeners, such as brightness (distribution of energy across the frequency range) or clarity of pulse (regularity of energy along the time line) are over something that is difficult to defend in cognitive terms (e.g., set-theoretic notions about chords or zero-crossings of the audio signal). In essence, certain features of music are more suitable and cognitively plausible for modeling certain behavior, although the level of abstraction depends on the task.

In the following sections, these principles will be applied to the modeling of emotions in music by organizing the model construction into distinct phases related to theoretical selection, feature validation, feature selection, and model evaluation.

2. Principles of model construction

When constructing a computational model that is able to predict emotions expressed by music, it makes sense to construct the model in four phases, namely (a) theoretically select plausible features, (b) validate the chosen features, (c) select and optimize the chosen features, and (d) evaluate the predictive capacity of the model. In the theoretical selection, I will briefly summarize the features, according to earlier research, which contribute to expressed emotions in music. In feature validation, they will be placed under critical scrutiny. In feature selection, empirical observations are then brought to bear on the choice of features to construct a model. Finally, the proposed model is evaluated using several unrelated datasets. To illustrate this process in more detail, as schematically outlined in Fig. 1, I will review each phase separately and connect them to emotions perceived by listeners in music using examples from my own studies.

The proposed method for constructing a computational model is also valuable in assessing existing computational models of emotions in music. If we focus on those which rely on audio-based features and a *correlational* approach, only a handful exist which are relevant. The majority of the computational models adopt an engineering approach, where a large number of any features (signal descriptors) are used to develop an optimal classifier of mood or emotion (Lu, Liu, & Zhang, 2006; MacDorman, 2007; Skowronek, McKinney, & Par, 2007; Yang, Lin, Su, & Chen, 2007) using sophisticated algorithms. Although mainly intended as applications, these studies are not necessarily helping us to understand the cognitive processes involved in the recognition of emotions since the cognitive relevance of many features is uncertain, although the evaluation part of the studies is usually conducted with rigor (i.e., using cross-validation). On the other hand, studies which concentrate on the psychological aspects of emotions (e.g., Coutinho & Cangelosi, 2009; Eerola, Lartillot, & Toiviainen, 2009; Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Schubert, 2004) often lack a proper evaluation of the constructed model, although there are expectations that this will be resolved (Coutinho & Cangelosi, 2009; Eerola et al., 2009; Rutherford & Wiggins, 2002).

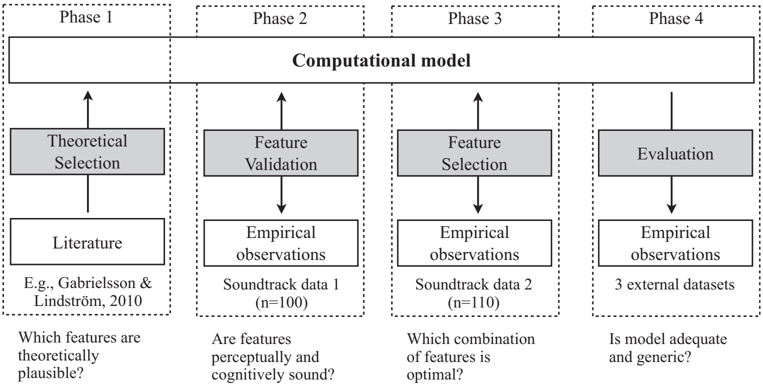
Phase 1
Phase 2
Phase 3
Phase 4
Computational model
Theoretical Selection
Feature Validation
Feature Selection
Evaluation
Literature
Empirical observations
Empirical observations
Empirical observations
E.g., Gabrielsson & Lindström, 2010
Soundtrack data 1 (n=100)
Soundtrack data 2 (n=110)
3 external datasets
Which features are theoretically plausible?
Are features perceptually and cognitively sound?
Which combination of features is optimal?
Is model adequate and generic?

Fig. 1. Schematic outline of the model construction and evaluation.

2.1. Theoretical selection of musical features which relate to the expression of emotions

The starting point of any computational model is an idea, most probably partially expressed in the prior literature and theory. In music and emotions, musical features contributing to emotions have been studied systematically since Kate Hevner's pioneering work (Hevner, 1935, 1936). Contemporary summaries of the musical features connected to particular emotions in music have been offered by Gabrielsson and Lindström (2010) and Juslin and Laukka (2003). These summaries are based on a large number of studies using both experimental manipulations of musical features as well as production studies. Manipulation studies have often focussed on relatively few features at a time, such as tempo, mode, dynamics, and pitch height (Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Ilie & Thompson, 2006). These latter attempts are causal in the sense that the structural features of music are systematically varied to see their effect on the ratings of emotions. While this is a powerful method, it is also limited in terms how subtle the manipulations can be and what musical material can be used. Another way to disambiguate emotional features of music is to approach the question from a correlational point of view. In this approach, existing music is analyzed in various ways and the musical contents are mapped to the emotions via statistical models (e.g., Leman et al., 2005; Schubert, 2004). While this is an informative and useful approach in itself and does not have serious restrictions on the musical materials that can be used, it is also limited by the problems of finding an appropriate self-report method for both experts and laypeople to describe emotions in music (Zentner & Eerola, 2010).

Summaries combining both causal and correlational studies of emotional features in music (Gabrielsson & Lindström, 2010; Juslin & Laukka, 2003) have produced fairly consistent sets of features for particular emotions, and a similar pattern of findings applies to temporal fluctuations in the emotional responses as well (e.g., Coutinho & Cangelosi, 2009; Schubert, 2004). For instance, *Anger* (or high arousal, negative valence in the dimensional affect space) is characterized by a high sound level, fast tempo, high pitch variability, high-frequency energy content, and fast tone attacks. *Happiness* (high arousal, positive valence) is also marked by fast tempo, but typically has a medium sound level, high pitch level, fast tone attacks, major mode, and bright timbre. *Sadness* (low arousal, negative valence) has the opposite trends in musical parameters: slow tempo, dark timbre, low pitch level, slow attacks, and minor mode. Finally, *Tenderness* (low arousal, positive valence) is similar to Sadness but differs in respect to mode, favoring major key. From such summaries, we can draw a list of candidate features that will be featured in a computational model of emotions. Since timbre is prominently present in these theoretical notions, the representation should be rich enough to encompass such aspects of music and therefore audio is taken here as the selected representation. However, this presents its own challenges requiring that we examine what kinds of features may be reliably estimated from audio.

2.2. Validation of the features

Out of the many musical features of emotions offered by the theoretical summaries, we should incorporate those which are relevant in human information processing, which can be demonstrated to have relevance in the processing such information, and which may be reasonably accurately measured in an audio representation of music.[1] These features should be qualified by demonstrating that the extracted features are meaningful for listeners and lend themselves to meaningful annotations of the music.

The computational extraction of musical features has expanded massively with the mushrooming of research in the field of Music Information Research (MIR). A number of efficient tools have been created that allow rich processing of both symbolic and audio data (Lartillot & Toiviainen, 2007; Leman, Lesaffre, & Tanghe, 2001; Tzanetakis & Cook, 2000). In the extraction of musical content from audio, several challenges exist. First, a reliable music transcription (i.e., score) is challenging, particularly for rich textures and polyphonic sources (Klapuri, 2004). However, for a computational model of expressed emotions in music and for many other MIR-related tasks (genre recognition, artist recognition, score follower, etc.), an exact transcription of the music is not a necessary starting point for the analysis (Tzanetakis & Cook, 2002).

With the cognitive approach, an attempt is made to model the human perceptual process by emulating the constraints of the human auditory system (e.g., auditory periphery, transformation of the signal within the cochlea, etc.) and by incorporating aspects of attention and memory processes known to be involved in the processing of music (see recent reviews in Purwins, Grachten, Herrera, Hazan, Marxer, & Serra, 2008; Purwins, Herrera, Grachten, Hazan, Marxer, & Serra, 2008). A summary of these processes and the categories involved in feature extraction is shown in Fig. 2.

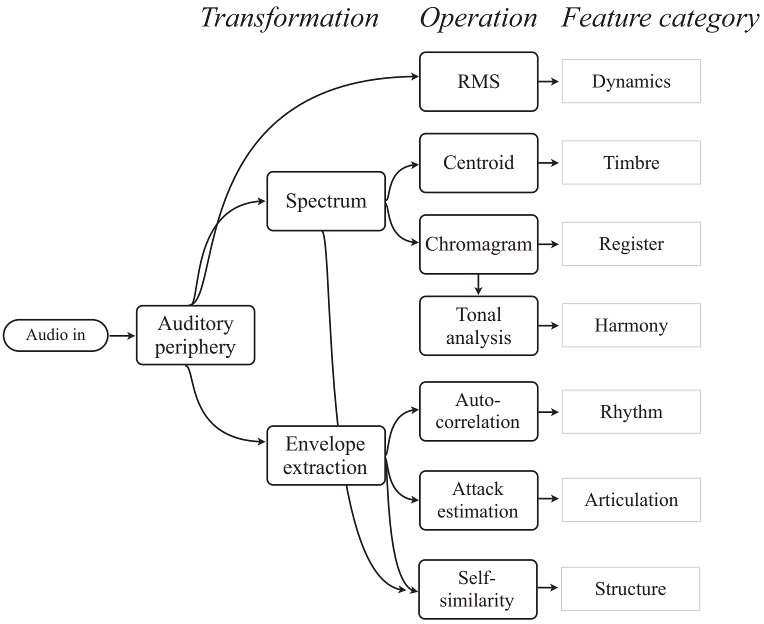

Fig. 2. Schematic representation of extraction of features from audio.

Typically, a variety of features may be extracted from the signal and most of these have been studied in a number of empirical studies. To give the reader an overview of some of the features that have been the subject of perceptual studies, a table of them with brief explanations is given (Table 1). The organization of the table follows loosely music-theoretic division into different elements (dynamics, timbre, register, rhythm, harmony, articulation, form). However, not all of the features may be easily grounded in perceptual and cognitive processes. Some of them, such as mel-frequency cepstral coefficients, do not have a direct counterpart in music theory. These have been successfully used in speech recognition (e.g., Chen, Paliwal, & Nakamura, 2003) and are known to provide a useful form of reduction that is adequately representative of the original signal.

Keeping in mind the cornerstones of cognitive science, it is vital to establish correlates between each of the extracted musical features and listeners' assessment of the music. Besides epistemological concerns about what is being represented, a pragmatic way of qualifying the features is to compare them to human annotations of music. An eminent analysis of music can be considered as one form of annotation. For instance, Robert Gjerdingen's characterization of changing-note patterns in the turn of the 19th century (Gjerdingen, 1988), or Fred Lerdahl's analyses of Tonal Pitch Spaces in famous works (Lerdahl, 2004). Another, more commonly used form of annotation within the MIR community is a detailed description of the sound events within an audio track. Events described might include the onset of each instrument, the type of instrument, the fundamental frequency, or the type of chord (e.g., Bello et al., 2005; Essid, Richard, & David, 2005). This form of annotation can then be directly used to develop a method for accurately detecting these characteristics from audio, although the drawback is that creating these detailed annotations is extremely time-consuming. Whereas the two previously described annotations were more or less expert dri-

Table 1
Examples of audio-based features

Category	Feature and Explanation
Dynamics	*RMS*, the root mean square energy of the signal, the amplitude of the sound volume
Timbre	*Spectral centroid*, the geometric centre of the spectral frequency; high value indicates prevalence of high-frequencies, which makes the sound more sharp and less soft (Juslin & Laukka, 2003)
Harmony	*Key clarity*, the maximum correlation of with so-called key profiles (Krumhansl & Kessler, 1982), which represent the stability of the pitch-classes in a given key (Purwins, Blankertz, & Obermayer, 2000)
Register	*Salient pitch*, the typical pitch in Hz or semitones established by chromagram-based methods (Bartsch & Wakefield, 2005)
Rhythm	*Pulse clarity*, how clear and stable the pulse or beat in music is, also called beat strength (Lartillot et al., 2008)
Articulation	*Attack time*, the time duration between the initial note onset time and its peak time, that is, how percussive or soft the sound attack is
Structure	*Novelty*, the degree of temporal repetition of any particular feature such as spectrum or chromagram across time based on detection of edges within the diagonal of the self-similarity matrix (Foote & Cooper, 2003)

ven, another strategy is to rely on laypeople who evaluate the amount of a certain concept in music using unipolar or Likert scales. The advantage of these type of annotations is that they use a higher conceptual level, because the raters are forced to rely on overall impressions and intuition; however, the very verbalization of these concepts brings its own limitations with it (Lesaffre, Leman, De Baets, & Martens, 2004).

An example of one such annotation, that uses participant ratings, is demonstrated next. To explore the validity of various extracted features, a selection of examples drawn from a database of 360 film soundtracks (Eerola & Vuoskoski, 2011) was taken. From this database, one hundred 5-s excerpts were selected, so that the chosen samples qualitatively covered a large range of musical attributes. Questions were then constructed about clarity of the pulse, the key, the beat structure, the articulation, and the brightness of sound qualities. Twenty-five musically highly trained participants then rated these concepts by answering the questions for each excerpt using 9-point Likert scales in a randomized experimental paradigm.

A considerable consensus among the participants existed for each concept ($\alpha > .90$). Next, the same set of musical stimuli was analyzed using computational means by extracting from each excerpt a measure of pulse clarity (Lartillot, Eerola, Toiviainen, & Fornari, 2008), key clarity, articulation (mean attack time, see Juslin, 1997), and brightness (as measured by the spectral centroid, see Juslin, 2000). Fig. 3 illustrates the match between the ratings and the models, indicating a moderately good correspondence for most concepts. Articulation received the lowest correspondence between the listeners' ratings and model prediction and this is believed to stem from the fact the participants may have been combining several types of information about the articulation character. In sum, an empirical experiment is helpful in checking whether the measures believed to represent certain musical concepts are in fact adequate.

Finally, it should be noted that usually conceptual evaluations by means of rating scales assume that the music is a static entity. This means the rater is forced to construct an impression of the overall qualities of the segment to be rated. If the segments are long (say, above 15 s), it is often more reasonable to have them rate the concepts continuously, by having them indicate the amount of each concept while simultaneously listening. Although in music and emotion research, it has been common to rely on static ratings of musical concepts to validate the features extracted, continuous paradigms have also been successfully employed (Coutinho & Cangelosi, 2009; Korhonen, Clausi, & Jernigan, 2006; McAdams, Vines, Vieillard, Smith, & Reynolds, 2004; Schubert, 2004).

2.3. Selection and weighting of the features

After having selected both theoretically and cognitively plausible features for the prospective computer model, a suitably large dataset needs to be amassed for the selection of prominent features and the weeding out of irrelevant or highly collinear ones. Usually the stimulus sets in music and emotion studies have been fairly small (e.g., < 20 examples) or contained well-known examples of classical music — for example, Adagios by Tomaso Albinoni/Giazotti and Barber — which are both drawbacks for serious modeling. For these

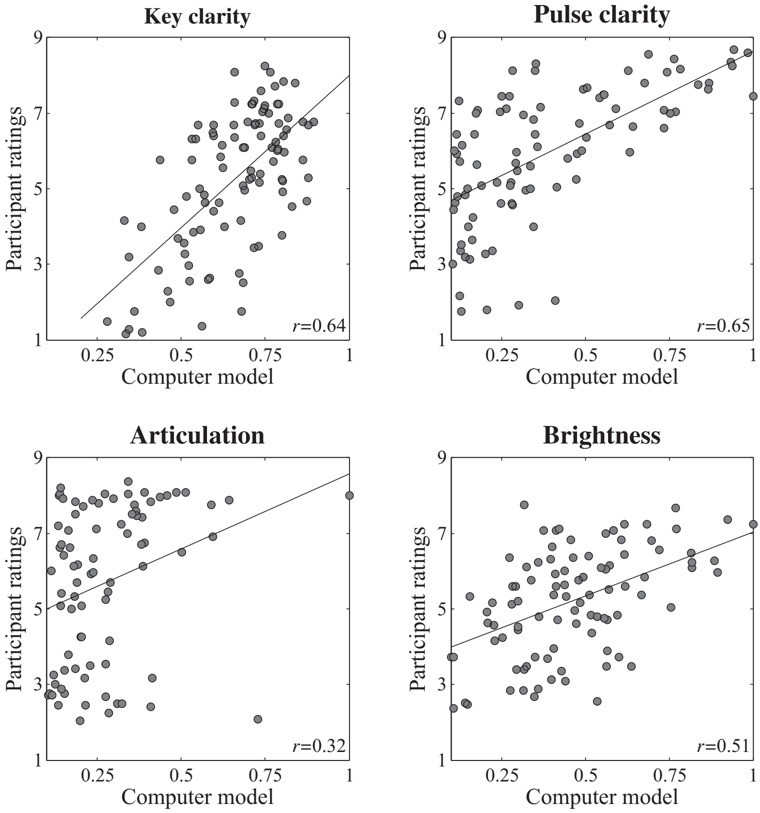

Fig. 3. Comparison of features and participants' ratings for the selected concepts (key clarity, pulse clarity, articulation, brightness) across 100 film soundtrack excerpts.

reasons, we have chosen film soundtracks because of their relevance to emotions in general, and to focus on such a genre which is relatively familiar to most people in Western culture. Film music also avoids the confounding effect that lyrics might have on emotions. In a series of experiments (Eerola & Vuoskoski, 2011), 110 unfamiliar film music excerpts were chosen from a larger pool of examples to represent five categories of basic emotions. In addition, the three-dimensional model of emotions was also utilized to explore whether the dimensional model, in this case valence, energy arousal, and tension (Schimmack & Reisenzein, 2002), could discriminate the examples more effectively. The excerpts were not solely meant to represent the best examples of each emotion but also ambiguous, moderate examples.[2] The excerpts were rated in terms of the five basic emotion concepts and then the three-dimensional emotion model by 116 musically untrained participants (more details in Eerola & Vuoskoski, 2011). Here, we will mainly focus on the dimensions of valence and arousal, since it was shown to be the most non-redundant way of representing the emotions (Eerola & Vuoskoski, 2011) and they also feature in the majority of music and emotions studies (Zentner & Eerola, 2010).

A decision should be made whether to model the continua or the emotion categories (in other words discrete emotions, or quadrants in the valence arousal affect space) that underlie the emotion. Whereas modeling 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 in fact captured more realistically by modelling the dimensions of an affect space, since participants have made subtle distinctions on Likert scales and not simplified the data into categories or quadrants (e.g., Quadrant 1 would be in the positive valence and high arousal category). Hence, linear models are preferred here since they may better capture the original, richly subtle data about perceived emotions (Eerola et al., 2009; MacDorman, 2007; Schubert, 2004).

Fig. 4 (panel A) displays the mean valence and arousal ratings across all 110 music excerpts. Four examples are highlighted — one from each quadrant of the valence–arousal

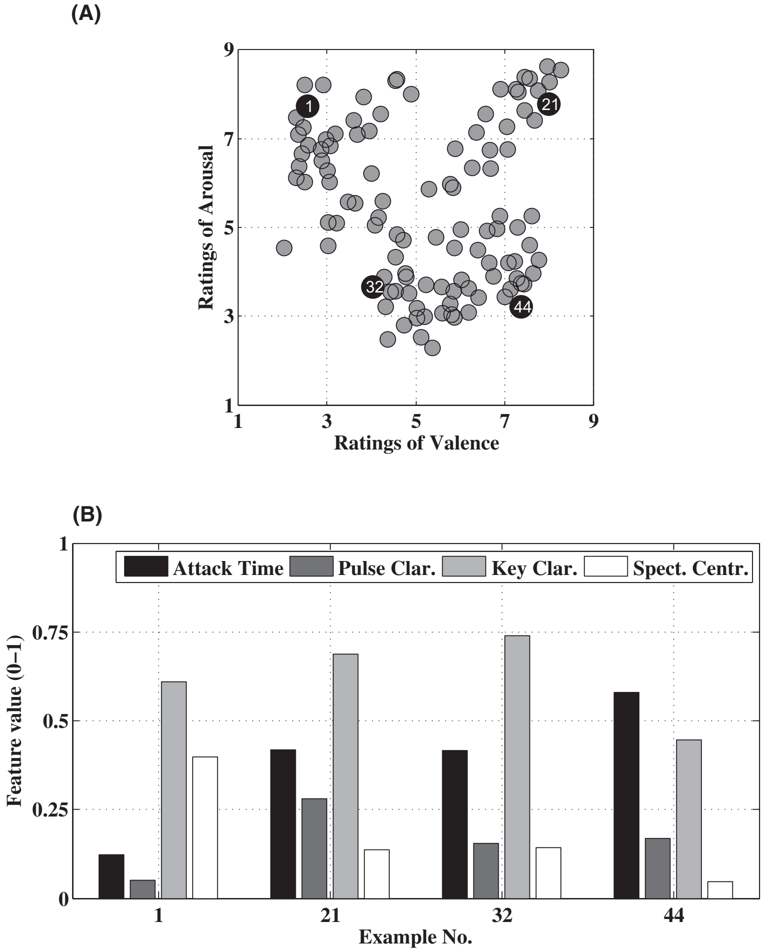

Fig. 4. (A) Visualization of valence and arousal for film soundtrack excerpts with four excerpts highlighted. (B) Four features for the highlighted excerpts demonstrating the musical characteristics of the selected excerpts.

affect space.[3] In panel B, the values for the four previously described musical features are shown for each chosen excerpt (all normalized to the range of 0–1). As anticipated intuitively, and by previous literature on music and emotions, the attack time shows a continuum: with a short attack time for the high arousal/negative valence example (no. 1), and a long one for the low arousal/positive valence (no. 44) example. This leaves the other two examples in the middle with respect to onset time. Contrastingly, the pulse clarity is highest for the high arousal/positive valence example (no. 21) and lowest for the high arousal/negative valence example. Key clarity differs again from these patterns, as the low arousal/positive valence example shows the least clearly defined key center within excerpt 21, while example no. 1 has second place in this ranking. Finally, the timbral characteristics of the examples (and the underlying affect space) corroborate differences in brightness (as measured by the spectral centroid) consistent with literature on speech expression (for instance, see Juslin & Laukka, 2003).

A full description of such correlations requires a form of regression analysis between the features and ratings. For instance, this author and collaborators (Eerola et al., 2009) explored various regression techniques (Robust Regression, Partial Least Squares Regression, and Principal Component Regression) to distill the critical features out of many candidate features that contribute to emotions in an optimal fashion. This operation is tricky if there are many candidate features, since there has to be at least 10 or 20 times more cases than predictors within regression analysis (Hair, Black, Babin, Anderson, & Tatham, 2006). This stresses the importance of the theoretical selection and empirical validation of the features in order to avoid overburdening the model with unnecessary features that will lead to unstable, unrealistic and material-specific models. An operation which guards against such overfitting is also an internal cross-validation, where only a portion of the data is used to develop the model and another part is used for evaluation (i.e., *k*-fold or leave-one-out cross-validation). Using 10-fold cross-validation and stepwise linear regression, 66% and 74% of variance in valence and arousal of the 110 soundtrack excerpts could be explained in terms of linear combinations of five extracted features (Eerola et al., 2009). A step beyond this measure of the model fit is to apply the model to a new dataset, described next.

2.4. Evaluation of the predictive capacity of the model

As shown above, it is possible to construct a viable model of emotions using a suitable set of data and an array of audio-based musical features. Although this approach does not guarantee good predictive accuracy for other musical materials, it provides a way to explore a large number of musical elements simultaneously and comprehensively in relation to emotions. To properly evaluate any such model (cf. Honing, 2006a), the predictive capacity of the model should be tested with other, unrelated datasets (Phase 4 in the model construction scheme). Several such moderately sized collections exist (e.g., Altenmueller, Schuermann, Lim, & Parlitz, 2002; Gomez & Danuser, 2007; Kallinen, 2005), although the concepts rated within these may be incompatible or the examples overly familiar to most listeners. Here three compatible, albeit small datasets are taken for the purposes of assessing the model's capacity to generalize (i.e., Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Dibben, 2004; Gomez & Danuser, 2004). They use real musical examples in a similar style, in this case classical, and have ratings of valence and arousal. The model initially developed for film soundtracks using a regression is applied to each of these datasets by first extracting the features from the musical examples and using the existing model to predict the emotion ratings in each study. The results of the evaluation are shown in Table 2, where the variance explained (R^2) is shown for each dataset. For the comparison, the prediction rates for two applicable, previous studies (two first rows) and the film soundtrack study (third row) are also shown in the Table 2.

Summary Table 2 reveals that the current model constructed with soundtrack dataset is fairly reliable for arousal, having sufficiently high prediction rates for independent datasets (mean $R^2 > .70$), which is considerably higher than the prediction rates obtained in previous studies (Leman et al., 2005; Schubert, 2004) that did not use cross-validation nor testing with unrelated datasets. In predicting valence, the current model is not able to predict the ratings quite as well as it does in arousal, although in comparison with prior studies, the rate of prediction is nevertheless on a par ($R^2 > .45$) with studies optimized with a single dataset. Space does not permit in-depth exploration of the drawbacks with proposed model, but the overall success suggests that such a systematic way of constructing a computational model is capable of producing reliable information about the expressed emotions. It has to be acknowledged that the independent datasets were chosen to be musically compatible in terms of the texture and style, and one might ask how dependent these features are on musical style in general. However, that is another research question altogether.

3. Conclusions

This article has provided a possible outline of the computational prediction of emotional responses to music. It relies on four phases for the model construction, where the features of the model are chosen according to theory, cognitive validity, and empirical observations

Table 2
Comparison of prediction rates for different models

Study	N	Genre	Arousal (R^2)	Valence (R^2)
Previous studies				
Schubert (2004)	4	Classical	.43	.58
Leman et al. (2005)	60	Mixture	.30	.46
Current model				
Eerola and Vuoskoski (2011)	110	Film	.81	.66
Independent datasets				
Gomez and Danuser (2004)	16	Classical	.79	.28
Bigand et al. (2005)	27	Classical	.56	.38
Dibben (2004)	16	Classical	.86	.68

Note. ‘‘Previous studies’’ provide examples of the past model performances (not modeled here). ‘‘Independent datasets’’ refers to the performance of the ‘‘Current model’’ with new materials.

within the context of emotions in music. Finally, a robust model should be able to generalize from this to other materials, which in this case was briefly demonstrated with three external datasets. The starting point of the model was recorded sound, since it provides the richest source of information for emotions expressed in music. This source of information is also particularly challenging, since the sensory, perceptual, and cognitive processes needed to explain high-level concepts are, at the moment, not fully understood. However, as further developments are made in the conceptual and perceptual aspects of music processing by listeners, better features will be incorporated into such models (e.g., Lesaffre et al., 2008; Turnbull, Barrington, Torres, & Lanckriet, 2008). One has to bear in mind that the whole context of this study and the models provided are steeped in Western cultural practices. The current research on cross-cultural issues of emotional experiences of music and the specific musical correlates of these is a relatively little studied area (Fritz et al., 2009). The extant studies seem to imply an existence of a core set of psychophysical features, which are based on underlying physiological changes related to emotions, and features that are unique to each culture. Some aspects of the current model address the core set of features (e.g., *RMS*, *spectral centroid*, and *attack time*), since these are important, cross-cultural cues of emotional expression in speech (Scherer, Banse, & Wallbott, 2001). Moreover, the current approach for modeling the emotional responses to music would offer a viable extension into the music of other cultures since the approach itself is basically a learning system.

Although computational models capable of predicting emotions expressed by music are already in high demand in musical content-based retrieval (e.g., Casey et al., 2008) and education contexts (e.g., Juslin, Karlsson, Lindstrom, Friberg, & Schoonderwaldt, 2006), there are yet more interesting possibilities for basic research if reliable ways of recognizing emotions in music are developed. Computational models of emotions may be helpful in finding unfamiliar but controlled musical stimuli (of a particular genre, e.g., or with lyrics) for studies of emotions where the model is applied to large collections of music from sources such as Last.fm or Naxos (Casey et al., 2008; Downie & Futrel, 2005). A better control of the qualities of musical stimuli in a variety of research contexts would also be highly desirable. Also music needs to be chosen by the participants, particularly when exploring questions such as the differences in emotions for individual music listeners (Vuoskoski & Eerola, 2011); chills and strong emotions when listening to music (e.g., Blood & Zatorre, 2001); aesthetic experiences in music (e.g., Müller, Höfel, Brattico, & Jacobsen, 2010); or arousal, pleasure, and other abilities relating to music (e.g., Schellenberg, 2005). In questions such as these, all of which relate to fundamental issues of music cognition, being able to describe and predict the emotional qualities of music chosen by participants is a resource that has not yet been fully exploited.

Notes

[1] This claim is made liberally here, since any truly accurate measurement of any concept by listeners is difficult; see, for example, roughness (Sottek & Genuit, 2005).

2. All sound examples available in http://www.jyu.fi/music/coe/materials/emotion/sound tracks/
3. The four excerpts used as examples are as follows. No 1 — high arousal, negative valence (*Lethal weapon 3*, track 8: 04:15–04:29); No 21 — high arousal, positive valence (*The Rainmaker*, track 3: 02:55–03:13); No 32 — low arousal, negative valence (*Running Scared*, track 15: 02:06–02:27), and No 44 — low arousal, positive valence (*Pride & Prejudice*, track 12: 00:01–00:15).

Acknowledgment

This research was supported by the Academy of Finland (Finnish Centre of Excellence in Interdisciplinary Music Research).

References

Altenmueller, 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.

Bartsch, M., & Wakefield, G. (2005). Audio thumbnailing of popular music using chroma-based representations. *Multimedia, IEEE Transactions on Multimedia*, *7*(1), 96–104.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5 Part 2), 1035–1047.

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.

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.

Chen, J., Paliwal, K., & Nakamura, S. (2003). Cepstrum derived from differentiated power spectrum for robust speech recognition. *Speech Communication*, *41*(2–3), 469–484.

Clarke, E. F. (1999). Rhythm and timing in music. *The Psychology of Music*, *2*, 473–500.

Clarke, E. F., & Cook, N. (2004). *Empirical musicology: Aims, methods, prospects*. Oxford, England: Oxford University Press.

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–B10.

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.

Eck, D. (2002). Finding downbeats with a relaxation oscillator. *Psychological Research*, *66*(1), 18–25.

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.

Eerola, T., & Vuoskoski, J. K. (in press). A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception*.

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

Essid, S., Richard, G., & David, B. (2005). Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 68–80.

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

Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology*. Boston: The MIT Press.

Foote, J., & Cooper, M. (2003). Media segmentation using self-similarity decomposition. In *Proceedings of SPIE Storage and Retrieval for Multimedia Databases* (Vol. 5021, pp. 167–175). Santa Clara, CA: SPIE-The International Society for Optical Engineering.

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. (1999). The Performance of Music. *The Psychology of Music*, *2*, 501–602.

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

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* pp. (367–400). Oxford, England: Oxford University Press.

Gjerdingen, R. (1988). *A classic turn of phrase: Music and the psychology of convention*. Philadelphia, PA: University of Pennsylvania Press.

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.

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

Hevner, K. (1935). Expression in music: A discussion of experimental studies and theories. *Psychological Review*, *42*(2), 186–204.

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

Honing, H. (2004). The comeback of systematic musicology: New empiricism and the cognitive revolution. *Tijdschrift voor Muziektheorie*, *9*(3), 241.

Honing, H. (2006a). Computational modeling of music cognition: A case study on model selection. *Music Perception*, *23*(5), 365–376.

Honing, H. (2006b). On the growing role of observation, formalization and experimental method in musicology. *Empirical Musicology Review*, *1*(1), 2–6.

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.

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

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., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *5*(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., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559–575.

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.

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.

Klapuri, A. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, *33*(3), 269–282.

Korhonen, M. D., Clausi, D. A., & Jernigan, M. E. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, *36*(3), 588–599.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368.

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, Austria: Österreichische Computer Gesellschaft.

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

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, *4*, 481–509.

Leman, M. (2003). Foundations of musicology as content processing science. *Journal of Music and Meaning*, *1*(FAL), section 3.

Leman, M. (2008a). *Embodied music cognition and mediation technology*. Cambridge, MA: The MIT Press.

Leman, M. (2008b). Systematic musicology at the crossroads of modern music research. In A. Schneider (Ed.), (*Systematic and comparative musicology: Concepts, methods, findings* (Vol. 24, pp. 89–119). Berlin: Peter Lang.

Leman, M., Lesaffre, M., & Tanghe, K. (2001). Introduction to the IPEM toolbox for perception-based music analysis. *Mikropolyphonie – The Online Contemporary Music Journal*. Retrieved February 2012, from http://www.ipem.ugent.be/toolbox/IT_PaperMeeting.pdf.

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.

Lerdahl, F. (2004). *Tonal pitch space*. Oxford, England: Oxford University Press.

Lesaffre, M., Leman, M., De Baets, B., & Martens, J. (2004). Methodological considerations concerning manual annotation of musical audio in function of algorithm development. In X. Serra (Ed.), *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2004)* (pp. 64–71). Philadelphia, PA: International Society for Music Information Retrieval.

Lesaffre, M., Voogdt, L., Leman, M., Baets, B., Meyer, H., & Martens, J. (2008). How potential users of music search and retrieval systems describe the semantic quality of music. *Journal of the American Society for Information Science and Technology*, *59*(5), 695–707.

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.

McAdams, S., Vines, B., Vieillard, S., Smith, B., & Reynolds, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, *22*(2), 297–350.

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.

Posner, M. I. (1993). *Foundations of cognitive science*. Cambridge, MA: The MIT Press.

Purwins, H., Blankertz, B., & Obermayer, K. (2000). A new method for tracking modulations in tonal music in audio data format. In *Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks (IJCNN'00)* (Vol. 6, p. 6270).

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.

Rink, J. (1995). *The practice of performance: Studies in musical interpretation*. Cambridge, UK: Cambridge University Press.

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.

Rutherford, J., & Wiggins, G. (2002). An experiment in the automatic creation of music which has specific emotional content. In C. Stevens, D. Burnham, G. McPherson, E. Schubert & J. Renwick (Eds.), *7th International Conference on Music Perception and Cognition*, (pp. 35–40). Adelaide, Australia: Casual Productions.

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

Scherer, K., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, *2*(4), 412–417.

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

Skowronek, J., McKinney, M., & Par, S. van de. (2007). A demostrator for automatic music mood estimation. In S. Dixon, D. Bainbridge & R. Typke (Eds.), *Proceedings of the International Conference on Music Information Retrieval* (pp. 345–346). Vienna, Australia: Austrian Computer Society.

Sottek, R., & Genuit, K. (2005). Models of signal processing in human hearing. *AEU-International Journal of Electronics and Communications*, *59*(3), 157–165.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford, England: Oxford University Press.

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(2), 467–476.

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

Tzanetakis, G., & Cook, P. (2002). 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.

Wiggins, G. (2009). Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music. In *11th IEEE International Symposium on Multimedia* (pp. 477–482). San Diego, CA: IEEE Computer Society.

Yang, Y., Lin, Y., Su, Y., & Chen, H. (2007). Music emotion classification: A regression approach. In *2007 IEEE International Conference on Multimedia and Expo* (pp. 208–211). San Diego, CA: IEEE Computer Society.

Zentner, M. R., & Eerola, T. (2010). Self-report measures and models. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications*. (pp. 187–221). Oxford, England: 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.