A Multi-Genre Model for Music Emotion Recognition using Linear Regressors

Darryl Griffiths[a], Stuart Cunningham[b,a], Jonathan Weinel[c], and Richard Picking[a]

[a]Faculty of Arts, Science and Technology, Wrexham Glyndwr University, Wrexham, LL11 2AW UK; [b]Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester M1 5GD, UK;[c]School of Computing and Mathematical Sciences, University of Greenwich, London, SE10 9LS UK.

ARTICLE HISTORY
Compiled February 6, 2023

ABSTRACT
Making the link between human emotion and the properties of a piece of music is challenging, especially given the individual and subjective nature of emotion. Our aim was to produce an efficient system that emotionally labels songs in databases containing multiple genres of music successfully. To achieve this, we employed a series of online self-report studies, utilising Russell's circumplex model of emotion as a framework. The first study (n=44) identified audio features that map to the dimensions of arousal and valence, for 20 songs. We found significant relationships between energy features and user arousal ratings, spread and spectral flatness features, and user valence ratings. From this we constructed a set of linear regressors. The second study (n=158) measured the efficacy of our system, utilising 40 new songs to create a ground-truth and using the coefficient of determination as a measure of goodness-of-fit. By evaluating our system against other work in the field, results show our approach may be effective at labelling music on the arousal and valence dimensions and is able to deal with music from a diverse range of musical genres. Notably, it performs particularly well in terms of valence, which previous studies have found challenging to address.

KEYWORDS
Arousal; emotion; MER; music; Perception; regression; valence

1. Introduction

The ability to carry a large digital library of music is now commonplace, thanks to the advent of psychoacoustic audio compression techniques and portable digital music players, the function of which have increasingly become integrated with ubiquitous devices, such as laptops, tablets and smartphones. The choice and portability of music afforded by the digital revolution has actually made listeners more *active* as opposed to *passive*, in that they seek out particular songs for specific emotional or contextual purposes, rather than being exposed to the playback of music over which they have little or no control in the selection of (Krause, North, & Hewitt, 2015; North, Hargreaves, & Hargreaves, 2004). This has led to a revolution in the way music is used in everyday life, where music playlists can be easily created to reflect a host of situations,

CONTACT Stuart Cunningham. Email: s.cunningham@mmu.ac.uk

intentions and contexts (Juslin & Laukka, 2004; Kamalzadeh, Baur, & Möller, 2012).

Music recommendation systems have become commonplace on platforms that support the organisation and acquisition of digital music. The approaches that are employed by such systems often draw upon collaborative or content-based filtering techniques, which have their digital origins in recommendation systems for online shopping and other commercial applications. The former is a community-based process that typically uses a rating system, which bases its recommendations upon *other users* that have similar tastes (Ricci, Rokach, & Shapira, 2011). The latter also uses a rating system, but focusing upon the *attributes of users* (e.g. demographics, preferences, etc.) and features they are interested in (musical genre, tempo, artist, etc.) (Mobasher, Cooley, & Srivastava, 2000).

While these approaches provide one possible solution for generating music playlists, they have several disadvantages. Collaborative filtering methods require a large amount of historical data from users, which gives rise to issues of privacy. Content-based methods also lack an understanding of users' interests and preferences (Shao, Wang, Li, & Ogihara, 2009). More importantly with respect to affective computing, these systems do not make recommendations based upon an individual's emotional state or incorporate knowledge of the emotional content and perceptual effects of the music. Affective music playlists have recently begun to receive more widespread attention as an alternative approach, which takes these emotional concepts into account. Music is often regarded as a *language of emotions* (Cooke, 1959); it expresses feelings that listeners perceive, recognise, and are moved by. Studies have suggested that the most common motive for listening to music is to influence emotions (Juslin & Sloboda, 2011). In correspondence with this, the use of emotion in creating music playlists is a frequently occurring concept with users (Stumpf & Muscroft, 2011), yet currently there are few effective systems for organising music according to emotion using computational methods. To this end, a scalable solution for designing affective music playlists is proposed in this paper, which has the ability to account for such emotional factors.

The aim of this research is to develop a system that can organise the vast amount of music that is available to the general public in a meaningful, personalised way, by providing a mechanism by which music can be emotionally labelled. In practice, users of the system could self-report their emotional state as coordinates in arousal and valence space. This would allow recommendation of similar music by selecting and sequencing songs according to their Euclidean distance from the user's coordinates, within a defined radius. Alternately, users could emotionally traverse their music collection by defining start and end points in arousal and valence space, plotting the vector between the two points, and selecting songs within a fixed Euclidean distance of intervals along the vector. This work constitutes a core component of a larger project that examines methods to produce affective and contextually-aware music playlists (Griffiths, Cunningham, & Weinel, 2016; **?**), which was carried out in the context of the (anonymised research team name) (Weinel, Cunningham, Griffiths, Roberts, & Picking, 2014).

In what follows, a critical overview of the fields of music recommendation systems and affective playlists is provided. Next, the methodology used for our two research studies is described, before each study is presented in detail. The first study deals with the creation of a set of regressors by mapping audio features to emotionally labelled music. The second study uses a similar approach to obtain a ground-truth of emotionally labelled music, plotted to the circumplex model of affect. Following this, we determine the efficacy of our affective models through a validation study utilising this ground-truth data. Finally, we summarise our findings, compare our approach to alternatives in the literature, and provide suggestions for future work that may build upon these outcomes.

2. Background

The field of Music Emotional Retrieval (MER) has grown in recent years and is a specific subset of the longer established investigation into Music Information Retrieval (MIR). MER deals with the determination of affective information from either symbolic or signal-based musical representation. Typically, research in the field of MER seeks to recognise emotions by analysing the musical features in songs, so that the music can be classified or the data used for recommendation purposes. As such, the field incorporates knowledge and tools from multiple disciplines.

Y.-H. Yang, Lin, Su, and Chen (2008) modelled emotions according to a two-dimensional arousal and valence system. This work employed a regression approach with the intention of alleviating ambiguity, in contrast to an alternate model that classifies music in terms of categorical emotions. Their work made use of a corpus of music, consisting of 195 songs in total. A ground-truth was created using subjective data gathered from 253 human participants with a test-retest methodology subsequently being applied, using 22 participants, alongside statistical measures of the regression model's error rates to validate the approach. The authors used the coefficient of determination R^2 as a measure of the model's performance against the set of results obtained from participants. This is one of the most common indicators of affective classification performance used in MER. This analysis showed that their model was able to account for 58.3% of the variance in the arousal dimension and 28.1% of the variance in the valence dimension.

Lu, Liu, and Zhang (2005) proposed a hierarchical framework to automate mood detection from acoustic music. Their framework comprised three musical feature sets including intensity, timbre, and rhythm, which are used to determine the emotional content of the music. The features were clustered using a Gaussian Mixture Models (GMM) to classify the music using the dimensions of arousal and valence. Specifically, the model was based upon Thayer's concept, which has four quadrants categorised as: *contentment, depression, exuberance,* and *anxious/frantic.* The system itself was validated using a dataset consisting of approximately 250 songs, which were annotated by 3 expert human participants, who listened to short 20 second clips of each song. Validation procedures demonstrated that it was effective with average accuracy, which is another ad-hoc measure of affective classification performance, being reported as 86.3%.

Eerola, Lartillot, and Toivianen (2009) provide a comparison of two common paradigms of emotion representation within music is presented, by contrasting multi-dimensional affective space and a collection of basic emotions. Ground-truth data from 360 different film soundtracks samples were used in order to assess the compatibility of such dimensional and discrete concepts. In order to evaluate these models, an expert panel (n=12) with an extensive musical background rated the musical excerpts with basic emotions (i.e. *happy, sad, tender, scary,* and *angry*) and dimensional concepts using Likert scales. The authors used a three-dimensional hybrid model of emotion based upon Thayer and Russell's theories (i.e. energetic-arousal, tense-arousal, and valence). Model prediction of perceived emotions in music was based upon timbre, harmonicity, register, rhythm, articulation, and structural audio features using the MIRToolbox (Lartillot & Toiviainen, 2007) and a sample of 110 soundtracks, rated by 116 participants. Three methods of regression were contrasted, which were Multiple

Linear Regression, Partial Least Squares, and Principal Component Analysis in conjunction with a Box-Cox transform. In performance evaluation using R^2, they were able to account for a maximum (upper-bound) of 85% of the variance in the activity dimension, 72% in the valence dimension, and 79% in the tension dimension.

Table 1 provides a summary of the aforementioned studies, and a number of others, providing an indication of the variety of approaches used by MER researchers. In MER tasks, decisions must be made regarding the underlying model of emotion adopted, type of emotion being measured, classification approach, and the music to be used. Models of emotion commonly used in the literature include: Russell's circumplex model of affect (Russell C) (Russell, 1980); Thayer's arousal valence model (Thayer AV) or Thayer's Energy-Stress model (Thayer ES) (Thayer, 1990); and an adaption of Thayer and Russell's work creating a three-dimensional energy-tension-valence model (ETV) (Eerola et al., 2009).

It is important to be clear if measurements are being made of the emotion that a listener would *perceive* in a piece of music versus an emotion that might be *induced* in the listener. The key difference here relates to whether or not the music makes the listener *think* upon an emotion or whether they actually *feel* an emotion (Juslin, 2009). The classification approach may vary and typically involves statistical or machine-learning techniques to model the relationship between features extracted from the music to aspects or dimensions of the selected model of emotion. Such approaches typically take the form of: Support Vector Regression (SVR); Support Vector Machines (SVM) regression; Gaussian Mixture Models (GMM); Fuzzy Logic, Principal Component Analysis (PCA), and so on. Finally, the choice of songs used is important to understand, especially how they represent a diversity of musical genres and the culture from which they are drawn.

Other authors have recognised the disparity in approaches of the MER field (Huq, Bello, & Rowe, 2010), particularly the adoption of various models of emotion and techniques for elicitation of responses in participants. The absence of participant or user context within MER scenarios is highlighted as being a factor that needs to be addressed in the field. In response, as well as providing affective musical responses, the Moodo dataset (Pesek, Strle, Kavčič, & Marolt, 2017) also includes context and colour response information. A key driver of the Moodo dataset is to support development of personalised affective recommendation systems, much like the larger project within which our own work is located. The affective music ratings, perceived and induced, of Moodo were constructed by participants rating music by way of a discrete set of emotional labels. A total of 200 music clips, each 15 seconds long, were rated by 741 participants, each of whom listened to a subset of 10 music clips. High-level music features were produced for the songs in the datset by a human annotator (rhythm, tempo, harmonic complexity, and tunefulness) along with automatically extracted musical similarity metrics and analysed using DataFusion by Matrix Factorization (DFMF). Using linear models, the authors of Moodo found that user profile information was a significant predictor of emotional labels, suggesting that knowledge of the user and their content is valuable in producing personalised affective outputs from a MER system. Given the discrete approach adopted by the Moodo creators, the results must be viewed as a task of classification, rather than regression.

The studies in Table 1 have predominantly adopted a dimensional approach to the modelling of emotion, and that all but one of the existing studies use models by Russell (Russell, 1980), Thayer (Thayer, 1990), or variations thereof. Most of the measures of emotion presented in Table 1 are based upon the *perceived* affect (emotions the listener may associate with the music), as opposed to the impact the music may *induce* on

Table 1. Recent Work in Music Emotion Recognition (MER)

Source	Model	Emotion	Method	Materials
Lu et al. (2005)	Thayer AV	Pereived	GMM	Western classical music
Y.-H. Yang, Lin, Su, and Chen (2008)	Thayer AV	Pereived	SVR and AdaBoost.RT	Western, Chinese, and Japanese popular music
Y. H. Yang, Lin, Cheng, and Chen (2008)	Russell	Perceived	Regression	Western popular music
Jun, Rho, Han, and Hwang (2008)	Thayer AV	Induced Fuzzy logic	Western popular music from Allmusic.com	
Sun and Tang (2009)	Thayer ES	Perceived	ICMM (GMM Variant)	Western classical music
Eerola et al. (2009)	Thayer and Russell ETV	Perceived	Multiple Linear Regression, PCA, Partial Least Squares Regression, and Box-Cox	360 soundtrack excerpts
Schmidt, Turnbull, and Kim (2010)	Thayer AV	Induced	Regression	USPOP2002(Berenzweig, Logan, Ellis, & Whitman, 2004): Western popular music
Chang, Lo, Wang, and Chung (2010)	Thayer AV	Induced SVM	Not specified	
Myint and Pwint (2010)	Thayer ES	Perceived	SVM and Fuzzy SVM	Western popular music
Deng and Leung (2012)	Russell C	Perceived	Graph embedding, SVR and PCA	Western classical music
Hu and Yang (2014)	Russell C	Perceived	SVR with Radial Basis Function	CH496 dataset Y.-H. Yang and Hu (2012): Chinese popular music; MER60 dataset (Y. Yang & Chen, 2011): Western popular music; DEAP120 (Koelstra et al., 2011) dataset: video clips from Europe and North America
Roda, Canazza, and De Poli (2014)	Russell C	Perceived	Regression	Western classical music
Baume, Fazekas, Barthet, Marston, and Sandler (2014)	Semantic Mood Model (5-D MDS) (Barthet, Marston, Baume, Fazekas, & Sandler, 2013)	Perceived (Saari, Barthet, Fazekas, Eerola, & Sandler, 2013)	SVR	Production music
Wang, Yang, Wang, and Jeng (2015)	Russell C	Perceived	AEG Extension of GMM	DEAP and MER 60 datasets
Saari et al. (2015)	Valence, arousal, tension	Perceived	Semantic tagging and semantic tagging/audio feature	TR100k TR10k TE600
Mo and Niu (2017)	Thayer ES and Eerola and Vuoskoski categorical model (Eerola & Vuoskoski, 2011)	Not stated	OMPGW (Orthogonal Matching Pursuit, Gabor functions, Wigner distribution function) SVM	Soundtracks; MIREX dataset; MTV database; MediaEval 2015
Hu and Yang (2017)	Russell C	Perceived	Regression	MER60; CH818; AMG1608
Malheiro, Panda, Gomes, and Paiva (2016)	Russell C	Not stated	Regression	AllMusic

one's affect. However, it has been argued that the two are not easy to disentangle, since one is likely to influence the other (Roda et al., 2014). The majority of the studies in Table 1 used varying types of regression or machine learning, which were generally formed from high-level audio features, such as musical tempo or key, and used to predict music in terms of dimensional concepts. With some exceptions where Chinese music or soundtrack music is employed, the majority of studies are limited to music drawn from Western popular music and classical music. It is notable that authors have tended to select datasets from distinct genres of music, rather than to analyse a wide cross-section of musical genres from within a specific musical culture. This presents a notable opportunity, since it seems reasonable to assume that an individual's digital music library is likely to contain music from a range of genres. Previous investigation into MER approaches for multiple genres have showed indication that arousal features may be feasibly and accurately predicted across genre, but that valence prediction across genres may be more challenging (Eerola, 2011).

3. Method

The research carried out was the result of two distinct studies. Both share a common foundation in terms of the model of emotion being employed and the method behind the elicitation of participatory ground-truth data. This was coupled with a signal-level analysis of extracting a range of audio features, with the intention of identifying links between the participants' emotional ratings and the audio features to produce regression models.

Linear regression was selected on the basis that the best MER results currently reported in the literature utilize regression (Eerola et al., 2009), as do other highly performing MER systems (Hu & Yang, 2017; Malheiro et al., 2016; Roda et al., 2014). Given the new sets of data being created in our work this was a natural starting point for investigation. Regression offers benefits of being computationally efficient, stable in the presence of small variations, and less prone to over-fitting, when compared to other machine learning techniques. Since this research sought to work with specific audio features, and their prominence in MER tasks, a key advantage of linear regression is that its performance is easily explainable (Hagras, 2018).

A procedural overview of the research stages we report upon in this work is shown in Figure 1.

The first study sought to quantify emotional responses to a cross-section of music using a group of participants, affording the generation of a model of music and affect that links a series of audio features to dimensions of arousal and valence. The second study demonstrates the efficacy of this model by acquiring a ground-truth from another participant group, using a separate selection of songs, and evaluating the emotional prediction of our system to that of the ground-truth. For ease of interpretation, we deal with the particulars of each study in turn in later sections.

3.1. Model of Emotion

In order to standardise the method of representing emotion in the system that has been developed, it is necessary to adopt an appropriate model of emotion. There is a plethora of models in existence that describe affect. Many of the most popular methods in the MER domain fall into one of two branches: categorical (Ekman, 1992, 1999; Panksepp, 1992) or dimensional (Russell, 1980; Thayer, 1990; Watson & Tellegen, 1985).

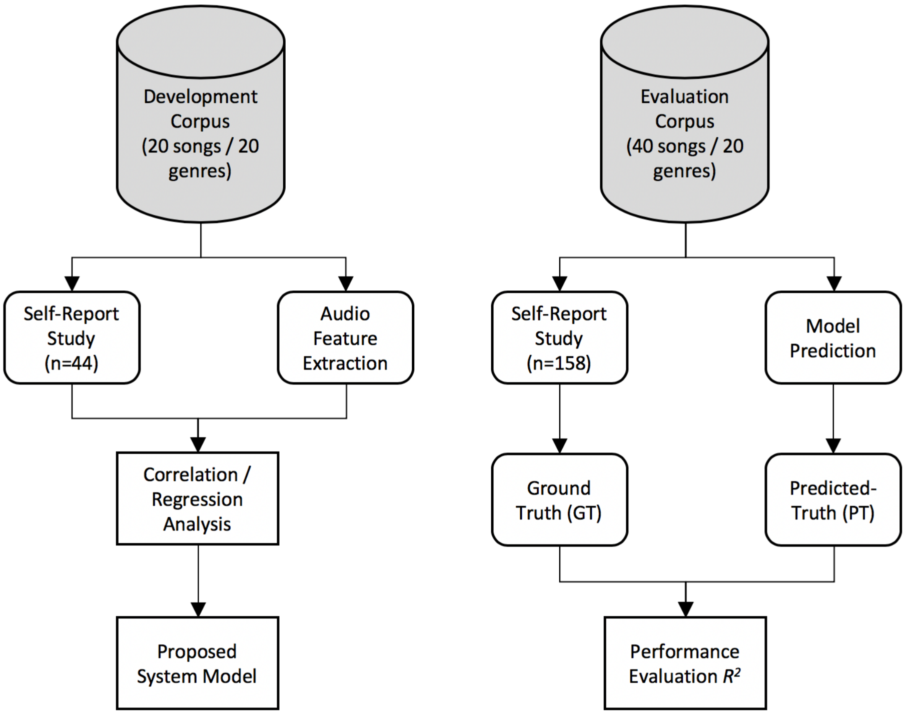

Development Corpus (20 songs / 20 genres)
Self-Report Study (n=44)
Audio Feature Extraction
Correlation / Regression Analysis
Proposed System Model
Evaluation Corpus (40 songs / 20 genres)
Self-Report Study (n=158)
Model Prediction
Ground Truth (GT)
Predicted-Truth (PT)
Performance Evaluation R^2

Figure 1. Overview of the Research Process.

In terms of the categorical paradigm, the disagreement concerning the exact number and labelling of distinct emotions suggests that these may be based upon linguistic and cultural taxonomies, rather than on actual defined sensations (Zentner & Eerola, 2010) and is a disadvantage. Conversely, a possible limitation of dimensional models is that the alignment of each axis does not necessarily correspond with the underlying physiological system that underpins the affective experience, although additional dimensions can be added to attempt to address this deficiency (Thayer, 1990). There is already a large body of work in existence describing both models (Barthet, Fazekas, & Sandler, 2012; Eerola & Vuoskoski, 2011, 2012; Scherer et al., 2000) although there is no agreed standard in affective music research mer-49, mer-49a. To this extent, both models are frequently found in the existing literature on music and emotion (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Dibben, 2004; Juslin, 2000; Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Resnicow, Salovey, & Repp, 2004; Wedin, 1972).

This work adopted an approach that conforms to Russell's circumplex model of affect (Russell, 1980), which is oriented around the use of two-dimensions corresponding to arousal and valence. It is argued that the simplicity of the model is one of the reasons for its popularity (Juslin, 2013) alongside the relative ease with which it can be integrated with variables from user studies and audio features. In particular, the choice of a two-dimensional model means that users being asked to express emotional ratings should find it fast and easy to engage with. However, a possible disadvantage of this approach is that some studies have concluded that arousal and valence dimensions alone are not able to account for all the variance in the emotions expressed by music (Collier, 2007; Ilie & Thompson, 2006) and that higher dimensional approaches may be more appropriate, accounting for subtle and nuanced aspects of affective response. Despite these criticisms, the simplicity of the circumplex model, coupled with evidence suggesting that there are no significant correlations between musical features beyond arousal and valence (Bigand et al., 2005; Eerola et al., 2009; Leman et al., 2005) support this decision.

3.2. Online Self-Report Studies

Data in our studies were collected using post-performance ratings, as opposed to a continuous self-report methodology (Schubert, 2011). Our work with participants used a self-report methodology, facilitated with a series of online questionnaires involving listening to several music selections and then providing affective ratings for each. This approach is simpler and can alleviate cognitive strain when compared to asking participants to produce real-time or continuous ratings (Zentner & Eerola, 2010). To make the mechanism intuitive, participants were required to express affective ratings using a mixture of categorical and ordinal variables, which were then mapped onto the circumplex model, adhering to the *forced choice* paradigm (Kallinen, 2005). According to Scherer, Shuman, Fontaine, and Soriano Salinas (2013), when individuals are permitted to choose their own words, the differences in vocabulary and culture make it difficult to compare data across individuals and studies, hence the compulsion of predefined constructs avoids this obstacle. Self-report approaches are commonly encountered in music and emotion studies (Drossos, Floros, Giannakoulopoulos, & Kanellopoulos, 2015; Hadjidimitriou & Hadjileontiadis, 2013; Kreutz, Ott, Teichmann, Osawa, & Vaitl, 2008; Ritossa & Rickard, 2004; Vuoskoski & Eerola, 2011).

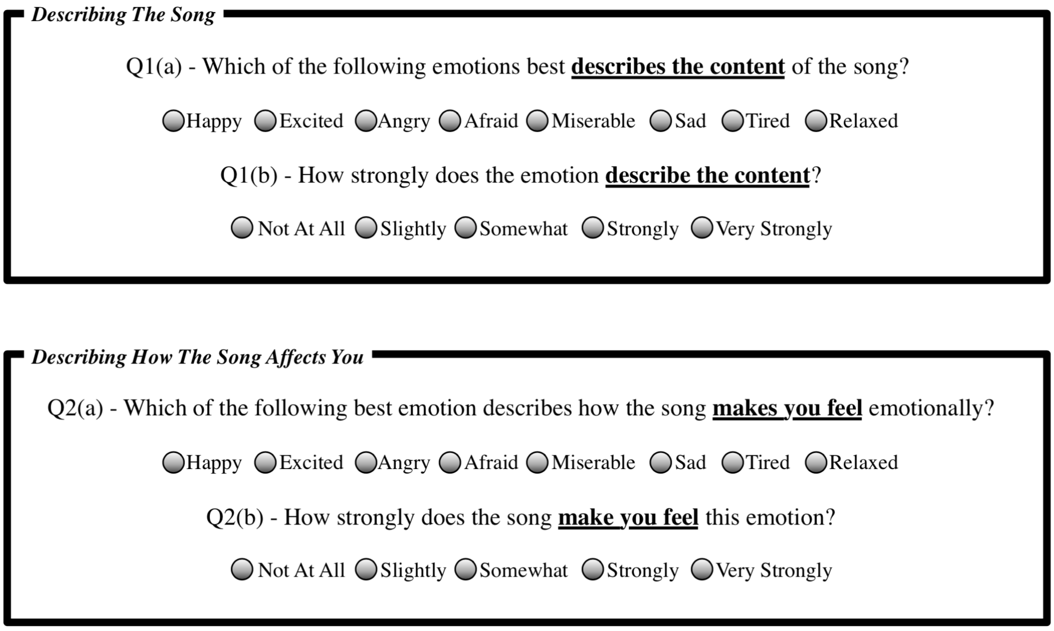
In selecting the number of songs in the study, we wanted to focus on reliability of ratings for each song sample rather than a using larger number of songs with fewer ratings. We took a pragmatic view to the time and number of participants we could realistically recruit in obtaining this number of ratings per song.

3.3. Approach to Mapping Self-Report Data to the Circumplex Model

This subsection recapitulates the salient aspects of an earlier work into mapping self-report data to the circumplex model (Griffiths, Cunningham, & Weinel, 2015). Since our method was to use an online tool to obtain a high number of responses, there was a concern that this could make training participants difficult, especially in the use of expressing emotions on the circumplex model or via an intermediary interface, such as the SAM scale (Bradley & Lang, 1994). To facilitate fluid interaction, we therefore selected a set of eight named emotions, covering all quadrants of the circumplex model: happy; excited; angry; afraid; miserable; sad; tired; and relaxed. After selecting one of these labels, participants were asked to indicate the strength of that emotion using an ordinal scale: not at all; slightly; somewhat; strongly; and very strongly. For each song used in the self-report studies, participants were then asked the following questions, the first two relating to perceived emotion, the third and fourth to induced emotion:

- *Which of the following emotions best describes the content of the song?*
- *How strongly does the emotion describe the content?*
- *Which of the following emotions best describes how the song makes you feel emotionally?*
- *How strongly does the song make you feel this emotion?*

Participants were prompted with these questions as illustrated in Figure 2.



Describing The Song

Q1(a) - Which of the following emotions best **describes the content** of the song?

○Happy ○Excited ○Angry ○Afraid ○Miserable ○Sad ○Tired ○Relaxed

Q1(b) - How strongly does the emotion **describe the content**?

○Not At All ○Slightly ○Somewhat ○Strongly ○Very Strongly

Describing How The Song Affects You

Q2(a) - Which of the following best emotion describes how the song **makes you feel** emotionally?

○Happy ○Excited ○Angry ○Afraid ○Miserable ○Sad ○Tired ○Relaxed

Q2(b) - How strongly does the song **make you feel** this emotion?

○Not At All ○Slightly ○Somewhat ○Strongly ○Very Strongly

Figure 2. A Graphical Representation of the Affective Online Self-Report Question format.

These eight emotions were chosen from a total of twenty-eight emotion vocabularies used in Russell's *category-sort* study (Russell, 1980). and are related to the eight emotion categories (pleasure, excitement, arousal, distress, misery, depression, sleepiness, and contentment) that were formed in a circular ordering in affective space. These conceptually indicate that emotions differ in their level of similarity and that some are usually perceived as contrasting emotions. The eight emotion labels chosen for this study were selected because of their musical relevance and total frequency in Russell's work. In further support of this choice, it can be noted that these terms appear in categories A, C, and F, respectively, in Schubert's work (Schubert, 2003). The emotions relaxed and sad are also used in the Geneva Emotional Music Scales (GEMS) (Zentner, Grandjean, & Scherer, 2008), whilst the term afraid was established in the study of Strong Experiences of Music (SEM) (Gabrielsson, 2001), which sought basic affect labels from users that described the strongest impact music had on their emotions. Since only one term was chosen from each category, a minor reformulation to the wording of the eight labels around the perimeter of the circumplex was undertaken, as shown in Figure 3. The placement of each of the eight emotions in this circular form means that each can be assigned a respective angle at 45°increments in a counter clockwise direction (Griffiths et al., 2015), where the emotion "Happy" represents 0°. The equal spacing of the labels on the model is supported by the majority of studies described in Section 2, that utilise either a Russell or Thayer type model of emotion.

These approaches permit the resultant data from participants to be placed in affective space, afforded by the circumplex model, using polar coordinates, where valence pertains to an angle θ and arousal refers to the radius ρ. As such, the ratings obtained are unipolar in nature, which is more applicable in this case, since it is argued that the opposite ends of the dimensions could be independent of one another (Watson & Tellegen, 1999). This approach allows a translation of the information from participants to the circumplex model. These points can then be converted to Cartesian co-ordinates for convenience and to allow for each emotional response to be mapped

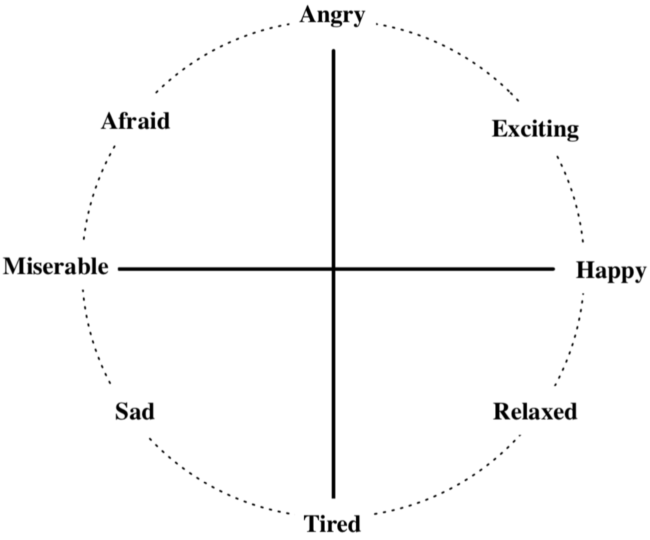


Figure 3. Russell's Eight Categories Placed in a Circular Order on the Circumplex Model.

to the dimensions of valence v and arousal a using the standard form

$$v = \rho \cdot \cos \theta. \tag{1}$$

$$a = \rho \cdot \sin \theta. \tag{2}$$

For example, consider a song receiving a participant rating of 'sad' and a strength of emotion rating of 4. This is represented as an angle $\theta = 225°$ and radius $\rho = 4$. It then follows that valence is calculated as $4 \cdot \cos(225)$ and arousal as $4 \cdot \sin(225)$, producing the Cartesian co-ordinates (x, y) of $(-1.035, -3.864)$.

It has been shown that discrete emotion terms, with underlying arousal and valence dimensions when conducting self-reports, are just as effective as a dimensional approach, and that both of these approaches are largely applicable in this field (Eerola et al., 2009).

4. Study One: Creation of the Model

4.1. Materials

To address the lack of research examining music across multiple genres, we decided to select twenty songs, one from each of twenty identifiable music genres. The songs were selected to provide pieces of music that were available as CD quality WAV PCM samples, and give a typical representation indicative of, and popular within, the given genre. The songs selected are described in Table 2. Using a cross-section of music tracks affords the system the capability to generalise, and thus be broadly applicable to a person's digital music collection, which is likely to consist of multiple styles and genres. In considering the range of broad genres included, it is accepted that these are located in the authors' own culture of Western popular music. Each song was

Table 2. Songs Selected for Model Creation

Genre	Song	Artist
Avant Garde	Passage De Recherche	Scanner
Blues	Reconsider Baby	Jools Holland and his Rhythm and Blues Orchestra
Classical	Adagio for Strings	London Philharmonic Orchestra
Country	9 to 5	Dolly Parton
Dance	Pjanoo	Eric Pridz
Disco	Relight my Fire	Vertigo – Dan Hartmann
Easy Listening	(They Long to Be) Close to You	The Carpenters
Electronic	Oxygene Part 4	Jean Michel Jarre
Folk	Gosport Nancy	Bellowhead
Funk/Soul	Across 110th Street	Bobby Womack
Gospel/Choral	Oh! It is Jesus	Soweto Gospel Choir
Heavy Metal	For Whom the Bell Tolls	Metallica
Hip Hop	Girls, Girls, Girls	Jay-Z
Indie	I Wanna be Adored	The Stone Roses
Jazz	Take Five	Dave Brubeck
Pop	Feels Like Heaven	Fiction Factory
Reggae	Stir it Up	Bob Marley
RnB	Try Sleeping with a Broken Heart	Alicia Keys
Rock	Back in Black	AC/DC
World Music	Chan Chan	Buena Vista Social Club

Table 3. Age Distribution of Participants in Development Study

15 - 24	25 - 54	55 - 64	65 and over
8	31	4	1

represented at CD quality, as uncompressed PCM in the WAV format with a sample rate of 44100 Hz, 16-bit depth, two-channel stereo. For ease of reference, we refer to each song in this study by its respective genre, rather than full name.

4.2. Participants and Approach

Following ethical approval from the University, users were invited to participate in the affective online self-report study principally through invitations distributed via the AUDITORY list (http://www.auditory.org/), an email list that specialises in the discourse of auditory perception, and a convenience snowball sample from the researchers' social media presence. We recruited participants over a period of 2 months with the intention of capturing at least 25 participants responses per song for questions 1(a), 1(b), 2(a), and 2(b), explained in Figure 2. A total of 44 participants took part, 26 female and 18 males, the majority of whom were in the 25 to 54 age group. The full age distribution of participants is shown in Table 3. Each participant rated all 20 songs.

The study was conducted online, using a website specially developed by the researchers to elicit audio ratings with respect to emotion. Participants were advised that the study would take up to 20 minutes to complete, that they would be listening to 20 song excerpts from 20 musical genres and be asked to provide emotional ratings. Emphasis was placed on the participants being instructed to use a good pair of headphones or speakers.

After confirming their consent to engage with the study, participants were presented with a one-minute sample from each of the selected songs, taken from the mid-point in its duration. The songs were presented to participants in a fixed sequence. Each

Table 4. Summary of Emotional Ratings from Participants

	Perceived Emotion				Induced Emotion			
Genre	*Emotion*	θ	ρ	$\chi^2(7)$	*Emotion*	θ	ρ	$\chi^2(7)$
Avant Garde	Sad	225	4	62.55	Sad	225	2	36.73
Blues	Sad	225	2	66.55	Relaxed	315	2	76.36
Classical	Sad	225	3	162.91	Sad	225	4	69.82
Country	Happy	0	3	49.45	Happy	0	3	94.18
Dance	Excited	45	3	188.00	Excited	45	4	77.09
Disco	Excited	45	3	78.18	Happy	0	3	87.27
Easy Listening	Happy	0	3	66.55	Relaxed	315	3	43.27
Electronic	Relaxed	315	3	40.73	Relaxed	315	2	53.82
Folk	Happy	0	3	122.91	Happy	0	3	139.27
Funk/Soul	Sad	225	2	17.82	Relaxed	315	2	30.55
Gospel/Choral	Happy	0	3	141.82	Happy	0	3	56.73
Heavy Metal	Angry	90	3	184.00	Excited	45	4	52.00
Hip Hop	Happy	0	3	21.45	Relaxed	315	2	31.27
Indie	Miserable	180	2	25.09	Happy	0	4	31.27
Jazz	Relaxed	315	3	125.45	Relaxed	315	4	72.36
Pop	Happy	0	3	50.55	Relaxed	315	3	46.55
Reggae	Relaxed	315	4	75.64	Relaxed	315	4	118.18
RnB	Sad	225	4	36.73	Happy	0	1	12.73
Rock	Excited	45	3	165.82	Excited	45	3	94.18
World Music	Relaxed	315	3	98.18	Relaxed	315	3	119.27

song appeared on a separate page of the website so as to avoid confusion. Participants were then asked a series of short questions, asking them to evaluate the perceived and induced emotional impact of each song. Subsequently, demographic information was gathered pertaining to the age range, gender, and listening device (speakers or headphones) used by each participant. Subsequently, 26 participants reported using speakers and the remaining 18 used headphones.

4.3. Results

Due to the nominal nature of the questions relating specifically to emotion and the ordinal characteristic of the emotional strength question, the mode was selected as the measure of central tendency to label each of the 20 songs. The emotional label and strength value for each song was determined by selecting the highest frequently occurring selection. User ratings for each song were verified for independence of the eight named emotion categories using Pearson's Chi-Square test. Ratings for each song were statistically significant ($p < 0.05$), with the exception of the music from the RnB genre in terms of its induced emotion rating. We hypothesise that this may have been due to participant fatigue, since this was the last song to be rated during the study. This process was applied to determine *perceived* and *induced* emotion for each song. The resulting data for each song, alongside significance statistics, are shown in Table 4.

5. Study Two: Development of the Model

With the affective rating of the songs by participants complete, the next stage was to identify significant correlations between the ratings and features extracted from the audio signal representation of each song. A total of 45 audio features were extracted from each of the songs using the MIR Toolbox (Lartillot & Toiviainen, 2007) feature library and the Matlab Audio Analysis Library (Giannakopoulos & Pikrakis, 2014). A set of audio features were selected for the modelling of arousal and valence, based upon the MER studies described in Section 2. The complete set of features used was: mean loudness; regularity loudness; tempo; low-energy rate; RMS energy; mean energy; mean energy entropy; standard deviation RMS energy; standard deviation energy; standard deviation energy entropy; median RMS energy; median energy; median energy entropy; mean zero-crossing rate; mean spectral centroid; mean spectral spread; mean spectral entropy; mean spectral flux; mean spectral rolloff; standard deviation zero-crossing rate; standard deviation spectral centroid; standard deviation spectral spread; standard deviation spectral entropy; standard deviation spectral flux; standard deviation spectral rolloff; median zero-crossing rate; median spectral centroid; median spectral spread; median spectral entropy; median spectral flux; median spectral rolloff; spectral rolloff (Hz); brightness; spectral skewness; spectral kurtosis; spectral flatness; mean spectral roughness; standard deviation spectral roughness; median spectral roughness; mode; mean onset peaks (rhythm strength); mean onset peaks - normal window (rhythm strength); mean auto correlation peaks (rhythm regularity); mean auto correlation peaks - compressed (rhythm regularity); and mean tempo.

The loudness and regularity of loudness features were computed from the overall music track (i.e. globally), but the rest were analysed on a frame-by-frame basis using low, mid and high-level features. Low-level features were extracted using a Hamming window with a size of 20 ms, with a hop size of 10 ms. Mid-level features were obtained using a window of 3 seconds in length.

Initial analysis examined the *perceived* emotion ratings received from the participants. Spearman's rank correlation coefficient was selected to determine the degree of association between the ratings of arousal and valence and each the feature vectors (Hu & Kando, 2012). Only statistically significant audio features obtaining Spearman's $\rho \geq .85$ and $p < 0.05$ were selected for inclusion in the model.

The energy feature correlation coefficients were calculated having removed three outliers. The removal of these outliers was deemed to be justified given our intention to strive for reliability of the dataset and the fact that each song had received 44 ratings. The songs *Try Sleeping with a Broken Heart* and *For Whom the Bell Tolls*, from the RnB and heavy metal genres respectively, were identified as being outliers during the analysis process by calculating Cook's distance using a 4/n cut-off. The rock genre song *Back in Black* was also removed since it had a Cook's distance close to the cut-off ($D = 0.151$) and was stylistically similar to the heavy metal song, with their loudness levels being within 1.01 LUFS of one another. In terms of the rock and heavy metal songs, it may be the case that the loudness levels did not meet the expectations of their corresponding arousal ratings because of the time they were produced (1980 and 1984, respectively). In the 1980s, when the digital CD became the primary medium and engineers were cautious with respect to digital clipping, the average level of a rock song in the 1980s was -16.8 dBFS (Southall, 2006). This led music production at the time to move towards more compressed forms of audio mastering. The resultant features identified to represent *perceived* valence and arousal are described in Table 5.

These features in Table 5 show a strong link between the energy (time domain) values in each song and the arousal determination, with a similar trend between spectral (frequency domain) features and the valence dimension. This family of acoustic features and characteristics have also been studied and found effective in predicting affect by other researchers in the MER field (Hu & Yang, 2017; Juslin, 2000; Y.-H. Yang, Lin, Su, & Chen, 2008).

Table 5. Audio Features Correlated with Arousal and Valence

Arousal		**Valence**	
Feature	*Spearman's* ρ	*Feature*	*Spearman's* ρ
Energy	0.95	Spectral Spread	0.86
Standard Deviation Energy	0.95	Median Spectral Spread	0.88
Median Energy	0.96	Spectral Flatness	0.88

When analysing the ratings of *induced* emotions, none of the audio features yielded significant results for valence or arousal, that met our threshold requirement of significance. This outcome is not surprising, given the highly subjective nature of individual emotion as well as the potential difficulty that participants may have had in identifying an experienced emotion in such a study. It is also notable that, as explained in section 2, the majority of existing studies in MER have focused upon perceived emotion, rather than induced. This finding results in the model that is devised being functional specifically for perceived emotion in music.

The linear regression equations that model the relationship of perceived arousal and valence to each of the significant audio features are shown in Table 6, alongside the goodness-of-fit, measured using the coefficient of determination (R^2), which includes the Sum of Squared Errors (SSE), and the Root Mean Squared Error (RMSE) respectively. In this MER research, we seek to evaluate the relationship between the set of emotion ratings from participants X and the predicted set of emotion ratings from the regression models Y.

The coefficient of determination (3) is a measure of goodness-of-fit calculated based upon set of dependent variables Y output from a regression model and calculated knowing the set of independent input variables X. This fit expressed as the proportion of variance that can be explained by the model and has a range between 0 and 1, where larger values indicate a greater amount of variability being explained by the model and is. The coefficient of determination is calculated by

$$R^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \right)^2 \tag{3}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the sets X and Y and respectively.

SSE is a measure of difference between the regression trend and the dependent variables. Therefore, it is a measure of the residual error between a prediction made by the regression model and mean participants' rating for a set of songs. Values closer to 0 indicate a better fitting model.

RMSE is the standard error of the regression model. It is the square root of the ratio between the SSE and total number of predictions n evaluated by the model. Values closer to 0 indicate a better fitting model.

Since each of the two emotional dimensions have been mapped to three audio features, it is necessary to reduce the output of each regressors to a single value. Initially, we considered weighting each of the three features relative to Spearman's ρ from the correlation analysis. However, the differences between these were miniscule, and so it was decided to use the mean for the initial model and its evaluation. A series of Matlab functions were created that take a new song as input and would produce a set of Cartesian coordinates as output, according to the linear regression equations in Table 6, one representing arousal and the other valence, allowing them to be mapped to the circumplex model.

Table 6. Linear Regression Models for Perceived Arousal and Valence

Arousal				
Audio Feature	*Regression Equation*	*SSE*	R^2	*RMSE*
Energy	$f(x) = 36.83x - 2.809$	2.956	0.921	0.444
Standard Deviation Energy	$f(x) = 166.3x - 2.875$	2.905	0.923	0.440
Median Energy	$f(x) = 37.23x - 2.791$	2.941	0.922	0.443
Valence				
Spectral Spread	$f(x) = 60x - 8.326$	3.942	0.959	0.468
Medial Spectral Spread	$f(x) = 59.94x - 8.302$	3.942	0.959	0.468
Spectral Flatness	$f(x) = 17.56x - 3.963$	3.59	0.962	0.447

6. Ground-Truth and Model Validation

In order to validate our model of perceived affect in music it was necessary to produce a second set of emotionally rated songs. To achieve this, a new set of music, drawn from a range of genres, was rated by a range of human participants and compared to the values generated by our system. As such, we performed a statistical comparison of a *ground-truth (GT)*, the results from a second online self-report study, with respect to a *predicted-truth (PT)*, which are the outputs from the affective regressor described in the previous section.

6.1. Materials

A total of 40 songs were selected for inclusion in this evaluation. These were again drawn from the twenty genres outlined in the first study, but this time two songs were chosen from each. As before, these were chosen by the researchers as being popular songs that typified their particular genre and were readily available as CD quality WAV PCM digital audio. The songs are described in Table 7.

Each song was represented as CD quality uncompressed PCM in the WAV format with a sample rate of 44100 Hz, 16-bit depth, and two channel stereo. For ease of future reference in the remainder of this section, we refer to songs by their respective genre, rather than full name.

6.2. Participants and Approach

Following further ethical approval from the University, users were invited to participate in another affective online self-report study, principally again through the use of the AUDITORY list and a convenience snowball sample from the researchers' social media presence. A total of 158 participants took part, 75 female and 83 males, the majority of whom were in the 25 to 54 age group. The age distribution of participants is shown in Table 8.

Unlike the self-report study used to create the model, this time the music tracks were randomly selected and presented to the users in order to mitigate extraneous factors, such as fatigue or training effects. Since there was a total of 40 tracks in this dataset, a subset of 10 tracks was chosen at random, without replacement, for each participant in the study. The recruitment of participants was continued until each of the 40 song samples had received a minimum of 25 valid responses.

Table 7. Songs Selected for Model Evaluation

ID	Genre	Song	Artist
1	RnB	7 Days	Craig David
2	Jazz	Alone Together	Paul Desmond
3	RnB	Blow	Beyoncé
4	Reggae	Bush Doctor	Peter Tosh
5	World Music	Coumba	Orchestra Baobab
6	Hip Hop	Ghetto Gospel	2Pac
7	Rock	Good Times Bad Times	Led Zeppelin
8	Classical	Grieg: Peer Gynt, Op. 23 - 2	Budapest Philharmonic Orchestra
9	Funk/Soul	Hold on I'm coming	Sam and Dave
10	Rock	I Want It All	Queen
11	World Music	Ibanuje Mon Iwon	King Sunny Ade
12	Reggae	King Tubby Meets Rockers Uptown	Augustus Pablo
13	Folk	Lady Percy	King Charles
14	Electronic	Love on a Real Train	Tangerine Dream
15	Easy Listening	Make It with You	Bread
16	Easy Listening	Mandy	Barry Manilow
17	Heavy Metal	Mouth for War	Pantera
18	Blues	Never Goina Change	Buddy Guy
19	Funk/Soul	Never Never Goina Give Ya Up	Barry White
20	Gospel/Choral	Ntate Mandela	Vusi Mahlasela
21	Hip Hop	Phenomenon	LL Cool J
22	Gospel/Choral	Rain, Rain Beautiful Rain	Ladysmith Black Mambazo
23	Avant Garde	Re-Platform	Yoshihiro Hanno
24	Disco	Ride Like the Wind	Christopher Cross
25	Dance	Right In	Skrillex
26	Country	Rose Garden	Lynn Anderson
27	Dance	Some Chords	Deadmau5
28	Pop	Somewhere in my Heart	Aztec Camera
29	Blues	Stack-A-Lee	Dr. John
30	Indie	Street Spirit	Radiohead
31	Jazz	Summertime	Louis Armstrong
32	Avant Garde	O Superman (for Massenet)	Laurie Anderson
33	Country	Tell Me Why	Taylor Swift
34	Classical	The Death of Falstaff	Neville Marriner
35	Electronic	The Model	Kraftwerk
36	Disco	This Time Baby	Jackie Moore
37	Folk	Truth Is	Levellers
38	Heavy Metal	Wicker Man	Iron Maiden
39	Pop	Wouldn't It Be Good	Nik Kershaw
40	Indie	You Do Something to Me	Paul Weller

Table 8. Age Distribution of Participants in Evaluation Study

15 - 24	25 - 54	55 - 64	65 and over
31	116	3	8

6.3. Results

Since each participant was asked to provide ratings for 10 randomly chosen songs from the set of 40, each song did not receive an identical number of ratings. The mean number of ratings per song was 39.5 (S.D. = 5.25) with the minimum being 25 and the maximum being 54. Ratings from participants were processed in the same manner as explained in section 3, turning the nominal and ordinal indicators into polar coordinates and then producing arousal and valence values as Cartesian coordinates. To add granularity to this evaluation, the GT data for each song is expressed as the mean of the participants' responses, expressed as Cartesian coordinates for arousal and valence. Ratings in the GT are presented along with the PT values, in arousal and valence space, in Figure 4.

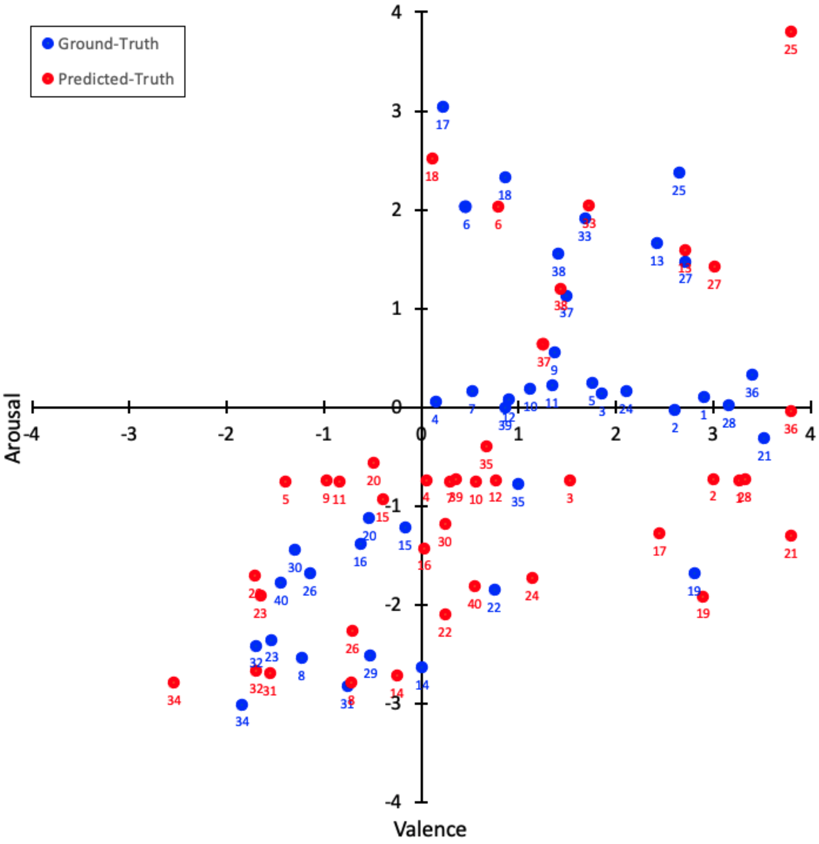


Figure 4. Ground-Truth versus Predicted-Truth Mean Arousal and Valence Values.

Notably, there is an absence of values in Quadrant II of Figure 4. Other MER datasets (Chen, Yang, Wang, & Chen, 2015; Zentner et al., 2008) tend to exhibit a large number of responses in quadrant I and the much lower responses in quadrant II. Given the number of songs examined, it is likely that these songs follow this trend, although it is recognised that this is likely to limit the ability to predict real emotional values that exist in quadrant II. Upon analysing the data in more detail, it was identified that a small number of outliers existed, which were identified by calculating Cook's distance using a $4/n$ cut-off. In the case of arousal, shown in Figure 5, two songs were defined as outliers: *Mouth for War* and *Right In* from the genres of heavy metal and dance respectively. When investigating valence, shown in Figure 6, two outliers were

Table 9. Performance Metrics for Affective Model

Emotion Dimension	Outliers Removed	R^2	SSE	RMSE	R^2 Lower-bound 95%	R^2 Upper-bound 95%
Arousal	N	0.708	29.700	0.884	0.691	1.059
Valence	N	0.674	30.581	0.897	0.562	0.894
Arousal	Y	0.850	12.341	0.585	0.841	1.118
Valence	Y	0.776	19.700	0.740	0.630	0.910

identified: *You Do Something to Me* and *Coumba* from the genres of indie and world music.

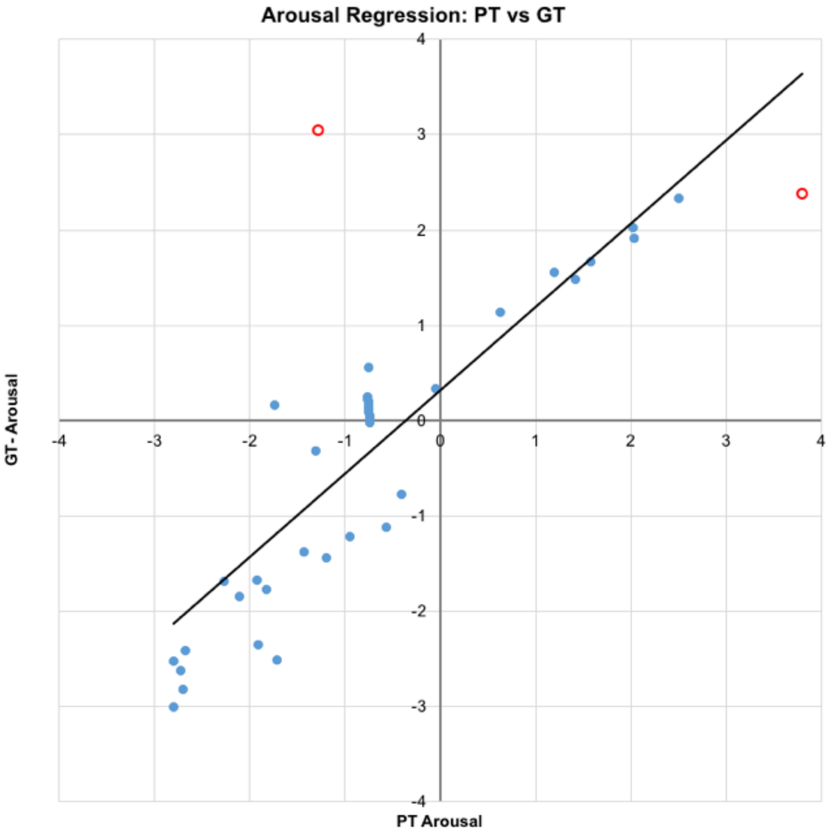


Figure 5. GT versus PT for Arousal Dimension Showing Outliers.

As such, it is useful to examine the model with and without the outliers included. A full description of the performance data of the model, with and without outliers included, is given in Table 9.

These findings show that our system, when adjusted for outliers, explains 85.0% of the variance in the arousal dimension and 77.6% in the valence dimension. Where the outliers are included these values reduce to 70.8% and 67.4% respectively. The imbalance between these two values is not uncommon in the field of MER, where it is often reported that prediction of arousal is more effective than prediction of valence (Eerola et al., 2009; Hu & Yang, 2014; Kim et al., 2010; Y.-H. Yang, Lin, Su, & Chen, 2008).

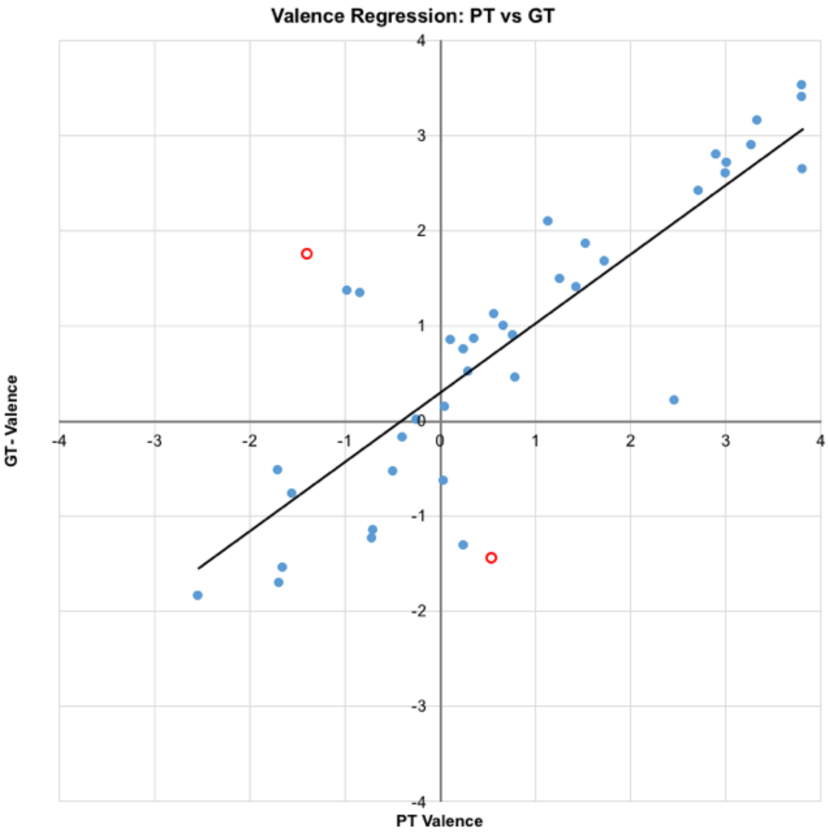


Figure 6. GT versus PT for Valence Dimension Showing Outliers.

7. Analysis and Comparison to Existing Methods

The data obtained in the previous section indicate very good levels of prediction in the system presented, especially since it has been developed using a modest corpus of songs that span multiple genres of music. To contextualise the performance of our approach, and evaluate its efficacy, it is helpful to compare it to existing work in the literature, as explored in section 2. To this extent, Table 10 details previous studies where it is possible to compare the prediction accuracy of the system using the coefficient of determination R^2 for arousal and valence dimensions. The values we have included in Table 10 from these other works represents the strongest results reported, if more than one variation of an approach or dataset was presented in a publication.

Whilst not included in the table due to the volume of individual results, it is worth also noting the MER performance metrics reported in an analysis of approaches to affective prediction in the MediaEval Database for Emotional Analysis in Music (DEAM) dataset (Aljanaki, Yang, & Soleymani, 2017). This dataset was produced partially to address the lack of publicly available data, particularly due to copyright reasons and consists of affective ratings of 1802 songs. It documents various studies in MER that report performance metrics of R^2 with the largest values of 0.83 for arousal and 0.48 for valence. The authors indicate that the use of recurrent artificial neural networks yield the best performance.

In comparing our system to others in Table 10, it is shown that our system is capable of producing extremely good results, in terms of how well these musical features

Table 10. Evaluation of Proposed System against Existing Work: (* = 1st; ** = 2nd, *** = 3rd Place Ranking)

Source	Arousal R^2	Valence R^2
Y.-H. Yang, Lin, Su, and Chen (2008)	0.58	0.28
Eerola et al. (2009)	0.85*	0.72**
Hu and Yang (2014)	0.80***	0.26
Roda et al. (2014)	0.75	0.34
Baume et al. (2014)	0.11	0.22
Saari et al. (2015)	0.74	0.46
Mo and Niu (2017)	0.84**	0.29
Hu and Yang (2017)	0.84**	0.53
Malheiro et al. (2016)	0.59	0.61
Our System	0.71	0.67***
Our System (outliers removed)	0.85*	0.78*

and models are predicting arousal and valence. Our approach outperforms those alternatives with which it is possible to draw meaningful comparison, where a small number of outliers are removed. Where outliers are not removed in the performance analysis, it still obtains better valence prediction than the majority of other methods, which has already been identified as being the harder of the two dimensions to reliably model (Eerola et al., 2009; Hu & Yang, 2014; Kim et al., 2010; Y.-H. Yang, Lin, Su, & Chen, 2008).

In addition to this, the system with outliers included maintains respectable performance in terms of arousal. These performance metrics must also be contextualised against the fact that our model has been developed and evaluated using a corpus of music that represents twenty different genres of music and which uses simple, computationally efficient, aggregated linear regressors and a small number of audio features (six), making it a lightweight, effective, and scalable tool for affective prediction of music.

8. Conclusions and Future Work

The results presented in this paper provide an initial indication that a functional affective playlist generator has been created and evaluated, which is able to function using music from a broad spread of genres. Using multiple linear regressors results in a computationally efficient and easily implementable approach that can label songs on the emotional arousal and valence dimensions. In particular, the approach we have described appears to be efficient in the prediction of songs on the valence dimension, where existing studies consistently report difficulties over the relative ease with which arousal can be dealt with. It is hypothesised that the spread of musical genres and use of regressors in this research might be the reason for this particular finding, since these are features of existing work in the literature where higher valence outcomes are reported (Eerola et al., 2009; Hu & Yang, 2017; Malheiro et al., 2016).

In terms of limiting factors for this study, there are several things to highlight. First, the number of songs used in the study is relatively small compared to other works in the MER domain, which was a deliberate and pragmatic choice to provide initial assessment about the feasibility of the multi-genre approach. Future work is being considered using online crowdsourcing platforms, such as Amazon Mechanical Turk, to obtain ratings for a larger corpus of songs. Second, participant ratings in study one were gathered by presenting songs in a fixed sequence, which may have given rise to an order effect. Similarly, the questions about perceived and induced emotion appeared in the same sequence in both data collection activities. Third, we did not consciously make an effort to direct study one participants to focus upon, or ignore, either the music or lyrics of each song, but rather that they listen to each excerpt as they would in everyday life. However, some participants from study one highlighted that they weren't sure if they should consider the lyrics or not. Consequently, in study two, we updated the instructions to explain that participants could consider lyrics or music as they wished when performing the rating. There is potential for this to have influenced the consistency of ratings in study one in particular. On reflection, providing consistency in this guidance in both studies or asking participants to extent to which they used music and lyrics in the process, would have been a way to control for any effect this may have had and is a feature that future rating studies should employ.

Even stronger validation of the models that we have presented could be achieved by further testing that makes use of existing, larger-scale datasets that contain human annotation of arousal and valence values. Selection of such datasets would need to be carefully considered, particularly with respect to cultural norms and of diversity of the material represented. The majority of existing datasets that focus on Western music tend to be narrow in genre. It may be that a process needs to take place in order to construct a new, multi-genre dataset from a range of those that currently exist. Doing this under expert selection, from musicologists, for example, would be a valid way to approach this type of task. There is also a need to account for the cultural background, preference and familiarity with songs in the dataset for anyone providing a rating. Sourcing affective ratings in controlled listening environments, using standardised reproduction equipment, rather than online methods, is another option to consider for creating highly reliable datasets in future. Such efforts would be a valuable future work.

One avenue that may lead to better models would be to use the original 'stems' from the multitrack music recording process, rather than the final stereo mix master. Whilst this does not reflect the typical way that ordinary users store their music library, and will likely be more complex to process, it may offer more accurate and subtle insights into the emotion of the music. A successful model would only need to be applied once in this manner and the results could then be sent to end users across their usual music download or streaming service. By evaluating the stems, as opposed to all of the content, affective nuances such as articulation (e.g. legato or staccato), what key the music is in, and whether its mode is mostly major or minor could be more noticeable.

Future work would benefit from looking at as wide a range of audio features as possible to determine those that are the most useful predictors of affect, especially since this is an aspect of MER that is continually evolving (Panda, Malheiro, & Paiva, 2018).

As outlined in the introduction, the system presented here forms part of an affective and contextual audio playlist generator (Griffiths et al., 2016; **?**). As such, other expansions of the work include evaluating the affective features in tandem with contextual ones and to determine an appropriate balance or relationship between the two. In the longer term, since it is possible to produce a two-dimensional prediction of each piece of music in a library, we intend to examine the effect of users being able to plot emotional vectors through this space. This, for example, would allow users to indicate their current emotional state and to specify a destination state and a time duration over which the transition is to be made, theoretically allowing them to move from one emotional state to another. We foresee particular applications of this type of system in healthcare and exercise applications, particularly where mental health issues may be present and where regulation of mood may help to reduce stress and agitation, or to prevent users from transitioning from low-mood to a state of depression. This could be especially poignant in the care of people living with dementia, autistic spectrum disorder, stroke, or acquired brain injury, for example.

References

Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS one*, *12*(3), e0173392.

Barthet, M., Fazekas, G., & Sandler, M. (2012). Music emotion recognition: From content-to context-based models. In *International symposium on computer music modeling and retrieval* (pp. 228–252).

Barthet, M., Marston, D., Baume, C., Fazekas, G., & Sandler, M. (2013). Design and evaluation of semantic mood models for music recommendation. In *Proc. international society for music information retrieval conference.*

Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014). Selection of audio features for music emotion recognition using production music. In *Audio engineering society conference: 53rd international conference: Semantic audio.*

Berenzweig, A., Logan, B., Ellis, D. P., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, *28*(2), 63–76.

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.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59.

Chang, C.-Y., Lo, C.-Y., Wang, C.-J., & Chung, P.-C. (2010). A music recommendation system with consideration of personal emotion. In *2010 international computer symposium (ics2010)* (pp. 18–23).

Chen, Y.-A., Yang, Y.-H., Wang, J.-C., & Chen, H. (2015). The amg1608 dataset for music emotion recognition. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 693–697).

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

Cooke, D. (1959). The language of music.

Deng, J. J., & Leung, C. (2012). Emotion-based music recommendation using audio features and user playlist. In *2012 6th international conference on new trends in information science, service science and data mining (issdm2012)* (pp. 796–801).

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

Drossos, K., Floros, A., Giannakoulopoulos, A., & Kanellopoulos, N. (2015). Investigating the impact of sound angular position on the listener affective state. *IEEE Transactions on Affective Computing*, *6*(1), 27–42.

Eerola, T. (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.

Eerola, T., Lartillot, O., & Toivianen, P. (2009). Prediction of multi-dimensional emotion ratings in music from audio using multivariate regression models. In *Proc. of the 10th int. society for music information retrieval conf. (ismir 2009)* (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.

Eerola, T., & Vuoskoski, J. K. (2012). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, *30*(3), 307–340.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Gabrielsson, A. (2001). Emotions in strong experiences with music. *Music and emotion: Theory and research*, 431–449.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to audio analysis: a matlab® approach*. Academic Press.

Griffiths, D., Cunningham, S., & Weinel, J. (2015). A self-report study that gauges perceived and induced emotion with music. In *2015 internet technologies and applications (ita)* (pp. 239–244).

Griffiths, D., Cunningham, S., & Weinel, J. (2016). An interactive music playlist generator that responds to user emotion and context. *Electronic Visualisation and the Arts*, 275–276.

Hadjidimitriou, S. K., & Hadjileontiadis, L. J. (2013). Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings. *IEEE Transactions on Affective Computing*, *4*(2), 161–172.

Hagras, H. (2018). Toward human-understandable, explainable ai. *Computer*, *51*(9), 28–36.

Hu, X., & Kando, N. (2012). User-centered measures vs. system effectiveness in finding similar songs. In *Ismir* (pp. 331–336).

Hu, X., & Yang, Y.-H. (2014). A study on cross-cultural and cross-dataset generalizability of music mood regression models. In *Proceedings of the 40th international computer music conference (icmc 2014)* (pp. 1149–1155).

Hu, X., & Yang, Y.-H. (2017). Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs. *IEEE Transactions on Affective Computing*, *8*(2), 228–240.

Huq, A., Bello, J. P., & Rowe, R. (2010). Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, *39*(3), 227–244.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319–330.

Jun, S., Rho, S., Han, B.-j., & Hwang, E. (2008). A fuzzy inference-based music emotion recognition system.

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.

Juslin, P. N. (2009). Emotional responses to music. In *Oxford handbook of music psychology.*

Juslin, P. N. (2013). What does music express? basic emotions and beyond. *Frontiers in psychology*, *4*, 596.

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., & Sloboda, J. (2011). *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.

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.

Kamalzadeh, M., Baur, D., & Möller, T. (2012). A survey on music listening and management behaviours.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proc. ismir* (Vol. 86, pp. 937–952).

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., ... Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*(1), 18–31.

Krause, A. E., North, A. C., & Hewitt, L. Y. (2015). Music-listening in everyday life: Devices and choice. *Psychology of music*, *43*(2), 155–170.

Kreutz, G., Ott, U., Teichmann, D., Osawa, P., & Vaitl, D. (2008). Using music to induce emotions: Influences of musical preference and absorption. *Psychology of music*, *36*(1), 101–126.

Lartillot, O., & Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects* (Vol. 237, p. 244).

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.

Lu, L., Liu, D., & Zhang, H.-J. (2005). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, *14*(1), 5–18.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9*(2), 240–254.

Mo, S., & Niu, J. (2017). A novel method based on ompgw method for feature extraction in automatic music mood classification. *IEEE Transactions on Affective Computing*, *10*(3), 313–324.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142–151.

Myint, E. E. P., & Pwint, M. (2010). An approach for mulit-label music mood classification. In *2010 2nd international conference on signal processing systems* (Vol. 1, pp. V1–290).

North, A. C., Hargreaves, D. J., & Hargreaves, J. J. (2004). Uses of music in everyday life. *Music perception*, *22*(1), 41–77.

Panda, R., Malheiro, R., & Paiva, R. P. (2018). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), 614–626.

Panksepp, J. (1992). A critical role for" affective neuroscience" in resolving what is basic about basic emotions.

Pesek, M., Strle, G., Kavčič, A., & Marolt, M. (2017). The moodo dataset: Integrating user context with emotional and color perception of music for affective music information retrieval. *Journal of New Music Research*, *46*(3), 246–260.

Resnicow, J. E., Salovey, P., & Repp, B. H. (2004). Is recognition of emotion in music performance an aspect of emotional intelligence? *Music Perception*, *22*(1), 145–158.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1–35). Springer.

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.

Roda, A., Canazza, S., & De Poli, G. (2014). Clustering affective qualities of classical music: Beyond the valence-arousal plane. *IEEE Transactions on Affective Computing*, *5*(4), 364–376.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Saari, P., Barthet, M., Fazekas, G., Eerola, T., & Sandler, M. (2013). Semantic models of musical mood: Comparison between crowd-sourced and curated editorial tags. In *2013 ieee international conference on multimedia and expo workshops (icmew)* (pp. 1–6).

Saari, P., Fazekas, G., Eerola, T., Barthet, M., Lartillot, O., & Sandler, M. (2015). Genre-adaptive semantic computing and audio-based modelling for music mood annotation. *IEEE Transactions on Affective Computing*, *7*(2), 122–135.

Scherer, K. R., et al. (2000). Psychological models of emotion. *The neuropsychology of emotion*, *137*(3), 137–162.

Scherer, K. R., Shuman, V., Fontaine, J., & Soriano Salinas, C. (2013). The grid meets the wheel: Assessing emotional feeling via self-report. *Components of emotional meaning: A sourcebook*.

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on multimedia information retrieval* (pp. 267–274).

Schubert, E. (2003). Update of the hevner adjective checklist. *Perceptual and Motor Skills*, *96*(3_suppl), 1117-1122. Retrieved from `https://doi.org/10.2466/pms.2003.96.3c.1117` (PMID: 12929763)

Schubert, E. (2011). Continuous self-report methods. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications.* Oxford University Press.

Shao, B., Wang, D., Li, T., & Ogihara, M. (2009). Music recommendation based on acoustic features and user access patterns. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(8), 1602–1611.

Southall, N. (2006). Imperfect sound forever. *Stylus Magazine*, *1*.

Stumpf, S., & Muscroft, S. (2011). When users generate music playlists: When words leave off, music begins? In *2011 ieee international conference on multimedia and expo* (pp. 1–6).

Sun, X., & Tang, Y. (2009). Automatic music emotion classification using a new classification algorithm. In *2009 second international symposium on computational intelligence and design* (Vol. 2, pp. 540–542).

Thayer, R. E. (1990). *The biopsychology of mood and arousal.* Oxford University Press.

Vuoskoski, J. K., & Eerola, T. (2011). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, *15*(2), 159–173.

Wang, J.-C., Yang, Y.-H., Wang, H.-M., & Jeng, S.-K. (2015). Modeling the affective content of music with a gaussian mixture model. *IEEE Transactions on Affective Computing*, *6*(1), 56–68.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, *98*(2), 219.

Watson, D., & Tellegen, A. (1999). Issues in dimensional structure of affect—effects of descriptors, measurement error, and response formats: Comment on russell and carroll (1999).

Wedin, L. (1972). *Evaluation of a three-dimensional model of emotional expression in music.* Psycological Laboratories. University of Stockholm.

Weinel, J., Cunningham, S., Griffiths, D., Roberts, S., & Picking, R. (2014). Affective audio. *Leonardo music journal*, 17–20.

Yang, Y., & Chen, H. (2011). Predicting the distribution of perceived emotions of a music signal for content retrieval. In *Ieee trans. audio, speech and lang. proc* (Vol. 19, pp. 2184–2196).

Yang, Y.-H., & Hu, X. (2012). Cross-cultural music mood classification: A comparison on english and chinese songs. In *Ismir* (pp. 19–24).

Yang, Y. H., Lin, Y. C., Cheng, H. T., & Chen, H. H. (2008). Mr. emo: Music retrieval in the emotion plane. In *Proceedings of the 16th acm international conference on multimedia* (pp. 1003–1004).

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, *16*(2), 448–457.

Zentner, M., & Eerola, T. (2010). Self-report measures and models. *Handbook of Music and Emotion: Theory, Research, Applications*, 187.

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