Running head: NEURAL NETWORK MODELS OF MUSICAL EMOTIONS

The Use of Spatio-Temporal Connectionist Models in Psychological Studies of Musical Emotions

Eduardo Coutinho and Angelo Cangelosi

University of Plymouth, Plymouth, Devon, United Kingdom

Abstract

This article presents a novel methodology to analyze the dynamics of emotional responses to music. It consists of a computational investigation based on spatiotemporal neural networks, which “mimic” human affective responses to music and predict the responses to novel music sequences. The results provide evidence suggesting that spatiotemporal patterns of sound resonate with affective features underlying judgments of subjective feelings (arousal and valence). A significant part of the listener’s affective response is predicted from a set of six psychoacoustic features of sound – loudness, tempo, texture, mean pitch, pitch variation, and sharpness. A detailed analysis of the network parameters and dynamics also allows us to identify the role of specific psychoacoustic variables (e.g., tempo and loudness) in music emotional appraisal. This work contributes new evidence and insights to the study of musical emotions, with particular relevance to the music perception and cognition research community.

Received February 25, 2008, accepted January 22, 2009.

Key words: emotion, music, arousal and valence, psychoacoustics, neural networks

Ever since antiquity, the relationship between music and emotion has been acknowledged as a fascinating quality of the human experience. Ancient philosophers such as Socrates, Plato, and Aristotle, in their theories of emotion, considered the sound of music and the unique way in which it can reflect "states of the soul." For the Greek philosophers music has the power to alter and drive the collective consciousness of massive groups of people.

Many years have passed and we still haven't found an answer to expose the mechanisms that music uses to interact with emotional systems. Nevertheless, the revival of studies on emotions during the late 19th century, together with the new technological developments in measurement techniques, have contributed with new insights for such an old question: how does music affect emotions?

Cognitivist and Emotivist Views

There are two principle, complementary views regarding the relationships between music and emotions. "Cognitivists" defend that music simply expresses emotions that the listener can identify, while "emotivists" defend that music can elicit affective responses in the listener (see Kivy, 1990; Krumhansl, 1997).

One of the most influential works from a cognitivist perspective was by Meyer (1956). He developed a theory in which musical emotions depend mainly upon expectations about the unfolding events and their meanings, which create patterns of tension and release in the listener (Meyer, 1956). For Meyer, expectation is a necessary condition for emotion and meaning to be conveyed in music. The nature of these expectations derives from the development of psychological schemas of systems of sound relationships. These include the general Gestalt principles for perceptual organization, but mainly psychological schemas derived from the interaction with a given (musical) culture. Without the “stylistic experience” music becomes meaningless and consequently lacks in affect. Empirical support for Meyer’s ideas has come from different formalizations of his theory (e.g., Cuddy & Lunney, 1995; Krumhansl, 1991; Narmour, 1992). Meyer’s cognitivist perspective is especially evident in a passage of his 1956 book: “... when a listener reports that he felt this or that emotion, he is describing the emotion which he believes the passage is supposed to indicate, not anything which he himself has experienced” (Meyer, 1956, p. 8). Although our affective experiences with music are ultimately individual and culturally dependent, emotivists claim that music can itself elicit emotions in listeners. From this perspective, there are certain music dimensions and qualities that induce similar affective experiences in all listeners, crossculturally, and independent of context and personal biases or preferences. Some evidence about the universality of music affect comes from a crosscultural study by Balkwill and Thompson (1999). Western listeners (who had no familiarity with North Indian ragas) listened to Hindustani music and were able to identify emotions of joy, sadness, and peace.

More compelling evidence suggesting that music itself can elicit emotions without the involvement of cognition, favoring the emotivist view on musical emotions, can be found in Peretz, Gagnon, and Bouchard (1998). Peretz et al. (1998) described a patient (I.R.) suffering from severe loss of music recognition and expressive abilities. I.R. showed no evidence of impairment in the auditory system but couldn’t discriminate pitch and temporal deviations in music. Even violations of the scale structure, or judgments of adequacy of a pitch as the ending of a harmonic sequence (tonal closure), were impossible to I.R. Despite all this, I.R. still claimed the capacity to enjoy music. In the experiment, the patient was able to derive the emotional tone of the excerpts, manipulated in terms of tempo and mode, to achieve the intended emotional qualities. Although I.R. was not aware of the music manipulations,[1] she performed as well as the control group on the affective content identification task. This study shows that the perceptual analysis of the music input can be maintained for emotional purposes, even if impaired for cognitive ones. Peretz et al. (1998) suggest the possibility that emotional and nonemotional judgments are the products of distinct neurological pathways. Some of these pathways were found to involve the activation of subcortical emotional circuits (Blood & Zatorre, 2001; Blood, Zatorre, Bermudez, & Evans, 1999), which also are associated with the generation of human affective experiences (e.g., Damasio, 2000; Panksepp, 1998), and can operate even outside an individuals' awareness. Panksepp and Bernatzky (2002) even suggest that a great part of the emotional power derived from music may be generated by lower subcortical regions, where basic affective states are organized (Damasio, 2000; Panksepp, 1998).

Taken together, these findings provide evidence of the universality of music affect and that cognitive mediation is not a required element in music appreciation. But in that case, for the affective experience to happen, it is plausible to think that the listener must derive affective meaning from the nature of the stimulus. This approach follows the view advocated by Langer (1942) on the existence of expressive forms ("iconic symbols") of emotions in all art forms. She believed that the arts and music in particular, are fundamental forms of human physical and mental life.

Music Elements and the Construct of Emotion

One of the major obstacles for experimental studies on the emotional power of music is the subjective nature and multiple components of the affective experience. Nevertheless, by focusing on the time course of emotional responses to music, several experimental studies suggest some generalizations. One of the most important is that different listeners report emotional responses to music, consistent in their quality and intensity. This led some studies to focus on the music features (e.g., tempo, mode, dynamics, among others), attributing to the variables correspondences with particular affective experiences. Some pioneering studies that investigated the influence of music parameters on perceived emotion were published by Hevner (1936). Hevner attempted a systematic explanation of such relationships. Since then a core interest amongst music psychologists has been the isolation and measurement of the perceptible factors in music that may be responsible for the resultant affective value (Gabrielsson & Lindström, 2001). The belief is that the way the sound elements are chosen and organized in time is linked with the listeners' affective experience.

Much of the research in this area has focused on general emotional characterizations of music (e.g., identification of basic emotion, lists of adjectives, or affective labels), by controlling parameters that can show some degree of stability throughout a piece (e.g., tempo, key, timbre, mode). In some studies, sets of specially designed stimuli have been used (e.g., probe tone test), while other studies were based on a systematic manipulation of real music samples (e.g., slow down tempo, changing instruments). More recently, following the claim that music features and structure are characterized by emotionally meaningful changes over time (e.g., Dowling & Harwood, 1986), new frameworks using use real music and continuous measurements of emotion emerged (e.g., Schubert, 2001).

Schubert (2001) proposed the use of continuous measurements of cognitive self-report of emotion; using a dimensional paradigm to represent emotions on a continuous scale. According to Wundt (1896), differences in the affective meaning among stimuli can succinctly be described by three pervasive dimensions (of human judgment): pleasure ("lust"), tension ("spannung"), and inhibition (“beruhigung”). This model has received empirical support from several studies that have shown that a large spectrum of continuous and symbolic stimuli can be represented using these dimensions (see Bradley & Lang, 1994). They can be represented in a three-dimensional space, with each dimension corresponding to a continuous bipolar rating scale: pleasantness-unpleasantness, rest-activation, and tension-relaxation. Other studies have provided evidence that the use of only two dimensions is a good framework to represent affective responses to linguistic (Russell, 1980), pictorial (Bradley & Lang, 1994), and music stimuli (Thayer, 1986). These dimensions are labeled as arousal and valence. Arousal corresponds to a subjective state of feeling activated or deactivated. Valence stands for a subjective feeling of pleasantness or unpleasantness (hedonic value; Russell, 1989).

The use of dimensional models is by itself a limitation on the representation and measurement of music emotions. Principally, the limitation is due to the wide variety of emotions conveyed by music and their limited representation by such a model. Another limitation arises due to the focus placed on a limited characterization of emotion: by asking participants to focus on their feelings, other components of emotion are not controlled for. Nevertheless the model shows important advantages compared with other methods used (generally classified as discrete emotions and eclectic approaches; Scherer, 2004). First, they are suitable to be used with continuous measurement frameworks. In this way, they allow analysis of the time course of emotion in more detail than other methods. Second, because they describe a continuous space not attached to a specific label, they also allow for the representation of a very wide range of emotional states, which is especially important in the context of music. By acknowledging its disadvantages, and by considering the important advantages offered by this method (particularly the simplicity in terms of psychological experiments and good reliability;

Scherer, 2004), dimensional approaches to emotion representation have been consistently used in emotion research.

Models of Continuous Measurements of Emotion in Music

Following a continuous measurement framework, some studies have focused on analyzing temporal patterns in music and emotion. The music stimuli are encoded into time-varying patterns in the form of psychoacoustic features. These correspond to perceptually separable elements (or groups of elements) that, when combined, provide a description of the “perceptual object”. Their division into separable sound dimensions allows for the study of their dynamics individually. The phenomena of music perception can then be described at different levels of detail, by selecting among different combinations of features.

Within this framework, two mathematical models that used time-varying patterns of music and emotion ratings have been proposed. Schubert (1999a) applied an ordinary least squares stepwise linear regression and a first order autoregressive model to his experimental data. He created regression models of emotional ratings, for selected music features, at different time lags for each piece. The relationships between music and emotional ratings were assumed to be linear and mutually independent, not accounting for the interactions among variables. The models also had the disadvantage of being piece specific.

Korhonen (2004) adopted a different modeling paradigm and extended the sound feature space and the music repertoire. He chose System Identification (Ljung, 1987) to model time-varying patterns of psychoacoustic features and emotion ratings. Korhonen’s contributions are the integration of all music features into a single module and the possibility to use the model with unknown pieces. Despite some improvements over Schubert’s work, the performance of this model is irregular. It outperformed Schubert’s models for some pieces, but performed worse in others. Another important disadvantage is that no insights on the processes used by the models to achieve the predictions are made. It is difficult to assess the meaningfulness of the relationships established to model the affective reactions based on sound features.

Although traditional time series analysis techniques allow for an investigation of the relationships between different processes, they often assume too much about the nature of the signals and their underlying behavior (due to assumptions like stationarity; Brockwell & Davis, 1991). The work by Schubert (1999a) and Korhonen (2004) has shown the relevance of auto and crosscorrelations among psychoacoustic variables, and the limitations associated with the use of time series analysis techniques (e.g., pdf's, stationarity, linear correlations). In the two studies described, the model analysis only highlights positive relationships between tempo and loudness gradients and arousal ratings. Other observations derived from the model analysis often lack in generality.

Spatio-temporal Connectionist Networks

In order to overcome such limitations we suggest that spatio-temporal connectionist networks (Kremer, 2001) offer an ideal platform for the investigation of the dynamics of affective responses to music. Specifically we propose the use of recurrent neural networks. The fundamental additional aspect of this neural network (when compared with the traditional feed-forward model) is the use of recurrent connections that endow the network with a dynamic memory.

Various proposals and architectures can be found in literature for time-based neural networks (see Kremer, 2001, for a review), which make use of recurrent connections in different contexts. In our study we have selected the Elman network (Elman, 1990), also called Simple Recurrent Network. An Elman Neural Network (ENN) is based on the standard architecture of a multilayer perception with an additional “context” or “memory” layer. The units in this layer receive a copy of the previous internal state of the hidden layer. They are connected back to the same hidden layer, through adjustable weights. These units endow the network with a dynamic memory, achieved through recursive access to past information of internal representations of input stimuli.

The internal representations of an ENN encode not only the prior event but also relevant aspects of the representation that were constructed in predicting the prior event from its predecessor (that is the effect of having learned weights from the memory to the hidden layer). The basic functional assumption is that the next element in a time-series sequence can be predicted by accessing a compressed representation of previous hidden states of the network and the current inputs. If the process being learned requires that the current output depends somehow on prior inputs, then the network will need to “learn” to develop internal representations that are sensitive to the temporal structure of the inputs. During learning, the hidden units must accomplish an input-output mapping and simultaneously develop representations that systematic encodings of the temporal properties of the sequential input at different levels (Elman, 1990). In this way, the internal representations that drive the outputs are sensitive to the temporal context of the task (even though the effect of time is implicit). The recursive nature of these representations (acting as an input at each time step) endows the network with the capability of detecting time relationships of sequences of features, or combinations of features, at different time lags (Elman, 1991). This is an important feature of this network because the lag between music and affective events has been consistently shown to vary over a range of five seconds (Krumhansl, 1996; Schubert, 2004; Sloboda & Lehmann, 2001).

ENNs use a training phase and a testing phase. Learning algorithms (supervised or unsupervised) define the way the model behaves during training when tuning its parameters for a certain task. The testing phase serves to test the model with novel data, either for prediction or validation of the model. A typical example of neural network training is categorization: given a set of training stimuli, the model is asked to separate them into a predetermined set of categories. An interesting phenomenon arises when we present the system with novel stimuli. These new inputs, after a successful learning process, should ideally be categorized within the learned categories space, reflecting the underlying grammar of the process being modeled. This process is called generalization and allows connectionist models to categorize novel stimuli.

In this article we will use an ENN to model continuous measurements of affective responses to music, based on a set of psychoacoustic components extracted from the music stimuli. Following the modeling stage, we make use of a set of analytical techniques, which allow for a better understanding of the relationships between sound features and affective responses. We then discuss the performance of our model and the implications of our findings for the emotivist and cognitivist perspectives on musical emotions.

Simulation Experiments

Method

The data for the experiments were obtained from a study conducted by Korhonen (2004).[2] The original self-report data include the emotional appraisals of six selections of classical music (see Table 1), obtained from 35 participants (21 male and 14 female). Using a continuous measurement framework, emotion was represented by its valence and arousal dimensions (using the EmotionSpace Lab; Schubert, 1999b). The emotional appraisal data were collected at 1Hz (second-by-second).

– Insert Table 1 –

Encoded features. Korhonen (2004) encoded the music pieces into the psychoacoustic space by extracting low and high level features, using Marsyas (Tzanetakis & Cook, 1999) and PsySound (Cabrera, 1999) software packages. Only Tempo was calculated manually, using Schubert's (1999a) method. The 13 psychoacoustic variables chosen (the 5 sound features representing Harmony variables included in Korhonen's study are not included here in order to exclude higher level features specific to the music culture, and with controversial methods for its quantification) are shown in Table 2 and described below (for convenience we will refer to the input variables with the aliases indicated in this table). Because some of these measures refer to the same psychoacoustic dimension, they were clustered into 6 major groups: Dynamics, Mean Pitch, Pitch Variation, Timbre, Tempo, and Texture.

– Insert Table 2 –

Dynamics: The Loudness Level (D_1) and the Short Term Maximum Loudness (D_2) represent the subjective impression of the intensity of a sound (measured in sones). Both algorithms estimate the same quantity (described in Cabrera, 1999) and output similar values.

Mean Pitch: The Mean Pitch was quantified using two power spectrum calculations (one from PsySound, and another from Marsyas). The Power Spectrum Centroid (P_1) represents the first moment of the power spectral density (PSD; Cabrera, 1999). The Mean STFT Centroid (P_2) is a similar measure and corresponds to the balancing point of the spectrum (Tzanetakis & Cook, 1999).

Pitch Variation: The pitch contour was quantified using 3 measures. The Mean STFT Flux (Pv_1) corresponds to the Euclidian norm of the difference between the magnitudes of the

Short Time Fourier Transform (STFT) spectrum evaluated at two successive sound frames. The standard deviation of P_2 (Pv_2) and of Pv_1 (Pv_3) also were used to quantify the pitch variations[3] (refer to Tzanetakis & Cook, 1999, for furthers details).

Timbre: Timbre was represented using the 4 different measures. Sharpness (Ti_1), a measure of the weighted centroids of the specific loudness, approximates the subjective experience of a sound on a scale from dull to sharp. The unit of sharpness is the acum (one acum is defined as the sharpness of a band of noise centered on 1000 Hz, 1 critical-bandwidth wide, with a sound pressure level of 60 dB); details on the algorithm used in Psysound can be found in Zwicker and Fastl (1990). Timbral Width (Ti_2) is a measure proposed by Malloch (1997) that measures the flatness of the specific loudness function, quantified as the width of the peak of the specific loudness spectrum (see Cabrera, 1999, for further details and slight modifications to that algorithm). The mean and standard deviations of the Spectral Roll-off (the point where a frequency that is below some percentage of the power spectrum resides; refer to Tzanetakis & Cook, 1999, for the details on these measures) are also two measures of spectral shape (Ti_3 and Ti_4). Although they do not directly represent timbre, Korhonen (2004) included these measures because they have been used successfully in music information retrieval.

Tempo: Tempo was estimated from the number of beats per minute. Because the beats were detected manually, a linear interpolation between beats was used to transform the data into second-by-second values (details on the tempo estimation are described in Schubert, 1999a).

Texture: Multiplicity (Tx) is an estimate of the number of tones simultaneously noticed in a sound; this feature was quantified using Parncutt's algorithm (1989), which was included in Psysound.

Modeling procedure. The psychoacoustic features constitute the input for our model. Each of these variables corresponds to a single input node of the network. The output layer consists of 2 nodes representing Arousal and Valence. Three pieces of music (1, 2, and 5), corresponding to 486 s, were used during the training phase. In order to evaluate the response to novel stimuli, we used the remaining 3 pieces: 3, 4, and 6 (632 s of music). Throughout this article we refer to the "Training set" as the collection of stimuli used to train the model, and "Test set" to the novel stimuli, unknown to the system during training, that test its generalization capabilities and performance. The task at each training iteration is to predict the next (t+1) values of Arousal and Valence. The target values (aka "teaching input") are the average Arousal/Valence pairs across all participants in Korhonen's (2004) experiments. In order to adapt the range of values of each variable to be used with the network, all variables were normalized to a range between 0 and 1.

The learning process was implemented using a standard back-propagation technique (Rumelhart, Hintont, & Williams, 1986). During training the same learning rate and momentum were used for each of the 3 connection matrices. The network weights were initialized with different random values. The range of values for each connection in the network (except for the connections from the hidden to the memory layer which are set constant to 1.0) was defined randomly between -0.05 and 0.05.

If the model also is able to respond with low error to novel stimuli, then the training algorithm was able to extract from the training set more general rules that relate music features to emotional ratings. To avoid the overfitting of the training set, we estimated the maximum number of training iterations and learning parameters. After preliminary tests and analysis, we decided upon 20,000 iterations as the duration of training, using a learning rate of .075 and a momentum of 0. The size of the hidden layer (which defines the dimensionality of the internal space of representations) also was optimized by testing the model with different numbers of hidden nodes. The best performance was obtained with a hidden layer of size five.

The root mean square (RMS) error is used here to quantify the differences between values predicted by the model and the values actually observed experimentally. Although this is a common measure to assess the performance of connectionist models, it gives little guaranties about a successful modeling process. We will use this measure only to compare the model performance with alternative sets of inputs to the network (next subsection). To assess the model ability to categorize the stimuli in terms of their affective value (and so the meaningfulness of the modeling process), we will analyze in detail the model categorization process.

Simulation 1: Reduction of the Psychoacoustic (Input) Dimensions

The choice of the input space must consider musical, psychological, and modeling aspects. The psychoacoustic features chosen by Korhonen (2004) include a significant set of perceptually relevant dimensions, although there are some redundancies to address. A recurrent problem in dealing with these types of data are the correlations among the encoded dimensions, especially redundant information and collinearity (as discussed by Schubert; 1999a). Because of that we decided to use only one variable of each of the psychoacoustic dimensions considered.

We started our simulations by training the neural network with different groups of inputs. Tempo, Texture, Dynamics, Mean Pitch, Pitch Variation, and Timbre are all considered to be included in the model as separate dimensions. In the case of Tempo and Texture, because they ere estimated using a single method (algorithm), they are included directly because there is no choice among alternative measures to be made.[4] In order to select one sound feature from the remaining music dimensions (Dynamics, Mean Pitch, Pitch Variation, and Timbre), each set of inputs considered included all unique features for each music dimension as a basic set (T and Tx as explained before), plus one other test variable(s). For instance, in the case of Dynamics we tested T, Tx, D_1, and D_2,[5] but also T, Tx, and D_1, and T, Tx, and D_2. We followed the same procedure for Mean Pitch, Pitch Variation, and Timbre.

For each test case we trained three different neural networks (with different random configuration of initial weights) and averaged their errors. Table 3 showns the RMS errors for each test condition.

– Insert Table 3 –

For the loudness measures, we found that the inclusion of both variables, or only $D_{1,}$ produced the best results. We selected D_1 from this group. Regarding Timbre, the best performance was achieved using only Ti_1, and so this variable also was selected. The variable selected to represent Mean Pitch is P_1, because it performs better than the remaining variables. Finally, Pitch variation shows very similar error values for all test cases. We chose Pv_1 because it yields a lower error than Pv_2 and Pv_3.

We trained another network including all the variables chosen (T, Tx, D_1, P_1, Ti_1, and Pv_1) in order to assess the performance with all variables together. The results are shown at the bottom of Table 3. An inspection of the RMS error shows that combining all the features improved the model performance substantially, suggesting that the interaction among different features conveys relevant information. In the following simulation experiment, we will use the selected 6 input features as the inputs for the model. The model architecture is shown in Figure 1.

Simulation 2: Analysis of Model Performance

We trained 37 neural networks (the same number of participants in Korhonen, 2004, experiments) with the data set comprising the psychoacoustic variables selected in Simulation 1 (see Figure 1). The average error (for both outputs) of the 37 networks was .05 for the Training set, and .076 for the Test set. These values correspond to 20000 iterations of the training algorithm.

– Insert Figure 1 –

In order to compare the model output with the experimental data for each piece, we calculated the Mutual Information (MI) between the model outputs and the respective target values (experimental data). The MI is a quantity that measures the mutual dependence of the two variables or, in other words, how much they vary together, and it detects both linear and nonlinear correlations between data sets. Because its interpretation, in terms of magnitude, is heavily dependent on data sets used (rendering difficulties for comparisons between different variables), we use a standardized measure for the MI (c.f. Dionísio, Menezes, & Mendes, 2006; Granger & Lin, 1994), based on the global correlation coefficient (λ), defined by $\lambda(X,Y) = \sqrt{1 - e^{-2*I(X,Y)}}$.[6]

The following analysis was performed on the network that showed the lowest average RMS error and λ for both data sets (network 24). The RMS errors and λ of each output for all the music pieces are shown in Table 4. Figures 2 and 3 show the Arousal and Valence outputs of the model for Training and Test sets, versus the data obtained experimentally (target values).

– Insert Table 4 –

– Insert Figure 2 –

– Insert Figure 3 –

The model was able to track the general fluctuations in Arousal and Valence for both data sets, although the performance varied from piece to piece. The model performance for Arousal was better for pieces 1, 2, 5, and 6 ($RMS_1 = .05$, $RMS_2 = .04$, $RMS_5 = .04$, and $RMS_6 = .05$), as shown by the low RMS errors (lower than the mean Arousal for all pieces: $RMS_{all} = .06$) and high λ. Pieces 3 and 4 had a higher RMS error than the mean of all the remaining pieces. Nevertheless, only piece 4 shows a λ significantly lower than the remaining pieces). This weaker performance is visible in Figure 3b). Even though the initial 80s (approximately) of the model predictions show the same increasing tendency of the experimental data, they do not follow the same pattern: they are lower during the initial 50 s ("dialogue" between flutes and strings) to which follows a strong increase (only strings playing in bigger number louder) until around 80s of the piece (a transition to a new section in piece).

The best Valence predictions were obtained for pieces 1, 2, 3, and 5 ($RMS_1 = .04$, $RMS_2 = .05$, $RMS_3 = .05$, and $RMS_5 = .05$): all these pieces had a RMS error lower than the average of all pieces: $RMS_{all} = .06$). The worst performances were obtained for pieces 4 and 6, although only piece 4 had a λ coefficient significantly lower than the remaining ones (with the exception of piece 5). In these cases, as for the Arousal predictions, poor performance is particularly evident during the initial 80s of the piece, as seen in Figure 3b).

The successful predictions of the affective dimensions for both known and novel music support the idea that music features contain relevant relationships with emotional appraisals. A visual inspection of the model outputs, confirmed by the RMS and λ measures, also indicates that the model output resembles the experimental data (with the exception of the initial 80s of piece 4). The spatio-temporal relationships learned from the Training set were successfully applied to a new set of stimuli.

These relationships now encoded in the network weights, and the flux of information in the internal (hidden) layer of the neural network represents the dynamics of the internal categorization (or recombination) of the input stimuli, that enables output predictions. One of the advantages of working with an artificial neural network is the ability to explore the internal mechanisms that generate the behavior and indirectly show how the model processes the information. In the following paragraphs we will analyze their spatial representation accordingly to Arousal and Valence levels using a method for dimensionality reduction.

Model internal dynamics: Discriminant functions. Clustering diagrams of hidden unit activation patterns are good for representing the similarity structure of the representational space. In order to analyze the internal dynamics of our model we use Linear Discriminant Analysis (LDA). The LDA is a classic method of classification using categorical target variables (features that somehow relate or describe the objects). Unlike Principle Component Analysis (PCA), in LDA the groups are known or predetermined.[7]

The main purpose of this algorithm is to find the linear combination of features that best separate between classes or object properties. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. Because we are interested in establishing the dynamics of the psychological report, we defined as the classification model the four quadrants of the two-dimentional emotional space (2DES; Q_1, Q_2, Q_3, and Q_4). We hypothesized that the quadrants division of the A/V space represents the underlying internal representations of the model. This method also allows us to identify the hidden units related with each dimension of the categorical space (an important aspects because it will allow for the study of the input-output mapping of the model).

The analysis has shown that two discriminant functions can explain 99.7% of the variance in the data.[8] The canonical correlations of the original data set are .821 for the 1st discriminant function (F_1) and .506 for the 2nd function (F_2). In Figure 4, we show the two discriminant functions. Each point corresponds to the internal state of the model at a particular moment in time. The dot's color identifies the category hypothesized for each internal state of the model, which correspond to the affective space quadrants (indicated by the labels Q_1 to Q_4).

– Insert Figure 4 –

The model shows an internal discrimination of the input stimuli, which is very similar to the affective space quadrants division. This indicates that the input stimuli were successfully categorized accordingly to their affective value, suggesting that the relationships built in the model transform meaningful patterns of sound features into the Arousal and Valence components of emotion.

As the discriminative power of the model is embedded in the hidden unit activations (the ones that connect to the output), we needed to assess the influence of each hidden unit on the pair of canonical variables. This was done by analyzing the factor structure coefficients shown in Table 5. These values correspond to the correlations between the variables in the model and each of the discriminant functions (similar to the factor loadings of the variables on each discriminant function in PCA).

The 1st discriminant function (F_1) receives the highest contributions from H_1, H_3, H_4, and H_5. F_2 receives the strongest contributions from H_2, H_4, and H_5. The next step was to identify how these units relate with the input and output layers. With that information we can estimate the input-output transformations of the model.

– Insert Table 5 –

Input/output transformation: Model production rules. To study the relationships between inputs and model predictions, we analyzed their relationships with the internal states of the model, which we saw to reorganize the sequence of input stimuli into meaningful affective representations (see previous section). One possibility would be to inspect the weight's matrixes in the model to identify the highest weights. Although simple, this methodology only compares weight values (long-term memory) and excludes the level of activity of each unit (including its bias) and implicit time representations (the short-term memory of the model).

In order to account for the temporal dynamics of the model, the correlations between inputs, hidden, and output units were computed using a Canonical Correlation Analysis (CCA) (Hotelling, 1936). A canonical correlation is the correlation of two canonical variables: one representing a set of independent variables, the other a set of dependent variables. The CCA optimizes the linear correlation between the two canonical variables to be maximized in the context of many-to-many relationships. There may be more than one linear correlation relating the two sets of variables, each representing a different dimension of the relationship, which explain the relation between them. For each dimension it is also possible to assess how strongly it relates each variable in its own set (canonical factor loadings). These are the correlations between the canonical variables and each variable in the original data sets.

In this article the CCA is used to assess the relationships between the sequences of input, hidden, and output layer activity. This method permits the analysis of the contribution of each network layer node or (sets of nodes) to the activity of a different layer. Relevant for our analysis are the relationships between input and hidden layers (how the inputs relate to the internal representations of the model), and these with the outputs (which sets of hidden units are more related to the output). In Table 6 we show the details of a CCA for the activity of the neural network layers.

– Insert Table 6 –

Input to hidden: Three canonical variables explain 98.3% of the variance in the data (see left side of Table 6). The first pair of variables loads on P_1, Tx, Ti_1 (inputs set), and H_2 and H_5 (hidden layer). The second loads only on input D_1 but it loads on all nodes of the hidden layer. The third canonical variable loads on Pv_1, H_2, and H_4. These three dimensions encode the general levels of shared activation in the input and hidden layers.

Hidden to output: Two canonical variables explain all the variance in the data (see right side of Table 6). The first root correlates strongly with Arousal and the activity in hidden units H_1 and H_2. The second pair of canonical variables correlates with both Valence (positive) and Arousal (negative), and with the activity in units H_3 to H_5.

Input to output: By taking together these two groups of relationships we can establish qualitative patterns of correlations illustrative of the general model dynamics. Hidden units H_1, H_2, and H_5 have a positive correlation with Arousal. H_5 correlates negatively with Valence and positively with Arousal. H_3 and H_5 correlate negatively with Arousal and positively with Valence. Because Tx, P_1, and Ti_1 relate positively to H_2, they have a positive effect on Arousal. The negative correlation with H_5 indicates that they correlate positively with Valence. D_1 correlated with the activity in all the hidden units. These correlations were consistently positive with Arousal. Finally, Pv_1 shows a negative correlation with Valence (through H_4).

In summary, the general strategies for input-output (sound features - affective dimensions) mapping found are:

Tempo (bpm): Fast tempi are related to high Arousal (quadrants 1 and 2) and positive Valence (quadrants 1 and 4). Slow tempi exhibit the opposite pattern;

Texture (multiplicity): Thicker textures have positive relationships with Valence and Arousal (quadrants 1, 2, and 4);

Dynamics (loudness): Higher loudness relates with positive Arousal;

Mean Pitch (spectrum centroid): The highest pitch passages relate with high Arousal and Valence (quadrants 1, 2, and 4);

Timbre (sharpness): Sharpness showed positive associations with Arousal and Valence (especially the first);

Pitch variation (STFT Flux): The average spectral variations relate negatively with Valence and positively with Arousal, indicating that large pitch changes are accompanied by increased intensity and decreased hedonic value.

Discussion and Conclusions

In this paper we presented a novel methodology to study the affective experience of music. From an emotivist perspective we considered that music can elicit affective experiences in the listener, focusing on sound features as a source of information about this process. Emotions were represented in terms of two pervasive dimensions of affect: Arousal and Valence. By focusing on continuous measurements of emotion we investigated the relationships between perceptual features of sound and reports of subjective feelings of emotion.

Initially we focused on the reduction of psychoacoustic variables used by Korhonen (2004), in order to identify a group of variables relevant for our hypothesis, but also to reduce the redundancy within the set. The initial simulations allowed us to select 6 variables: dynamics (loudness), pitch level (spectral centroid), pitch variations (mean spectral flux), timbre (sharpness), texture (multiplicity), and tempo. Then we conducted a series of simulations to "tune" and test our model. We used 486 seconds of music (three pieces) as the sample set (used to train the neural network to respond as close as possible to the human participants). A further 632 s of music (three pieces) were used as the test set. The model did not have any previous knowledge about these three pieces. We have shown that our model's predictions resemble those obtained from human participants.

In terms of modeling technique our model constitutes an advance in several respects. First, we are able to incorporate all music variables together in a single model, which permits to consider interactions among sound features (overcoming some of the drawbacks from previous models Schubert, 1999a). Second, artificial neural networks, as nonlinear models, enlarge the complexity of the relationships between music structure and emotional response observed since they can operate in higher dimensional spaces (not accessible to linear modeling techniques such as the ones used by Schubert, 1999a, and Korhonen, 2004). Third, the excellent generalization performance (prediction of emotional responses for novel music stimuli) validated the model and supported the hypothesis that psychoacoustic features are good predictors of the subjective experience of emotion in music (at least for the affective dimensions considered). Fourth, another advantage of our model is the possibility to analyze its dynamics; an excellent source of information about the rules underlying input/output transformations. This is a limitation inherent in the previous models we wished to address. It is not only important to create a computational model that represents the studied process, but also to analyze the extent to which the relationships built-in are coherent with empirical research. In our analysis we have identified consistent relationships between music features and the emotional response, which support important empirical findings (e.g., Davidson, Scherer, & Goldsmith, 2003; Gabrielsson & Juslin, 1996; Hevner, 1936; Scherer & Oshinsky, 1977; Thayer, 1986; see Schubert, 1999a, and Gabrielsson & Lindström, 2001, for a review).

Our work presented some evidence supporting the emotivist views on musical emotions. We have shown that a significant part of the listener's affective response can be predicted from the psychoacoustic properties of sound. We found that these sound features (to which Meyer, 1956, referred as "secondary" or "statistical" parameters) encode a large part of the information that allows the approximation of human affective responses to music. Contrary to Meyer's belief, our results suggest that "primary" parameters (derived from the organization of secondary parameters into higher order relationships with syntactic structure) do not seem to be a necessary condition for the process of emotion to arise (at least in some of its components). This also is coherent with Peretz et al.'s (1998) study, in which a patient lacking the cognitive capabilities to process the music structure (including Meyer's "primary" parameters), was able to identify the emotional tone of music.

Our research focuses on the expansion of the model. In an attempt to overcome the limitations of using a dimensional representation of emotion, we conducted an experiment using a similar framework as Schubert (1999a) and Korhonen (2004) but with the additional measurement of physiological activity. We intend to improve the description of music's affective experience by accounting for other components of emotion. Our goal is to assess the relevance of physiological cues for the prediction of the affective experience of music. We also will examine individual features in listeners, such as music training/expertise and personality traits, that may alter affective experience. These are also candidates to be incorporated into the model.

Author Note

The authors would like to acknowledge the courtesy of Mark Korhonen for sharing his experimental data, and the financial support from the Portuguese Foundation for Science and Technology (FCT). We are also grateful to Dr. Leonid Perlovsky for his relevant comments and suggestions, Dr. Stephen Hanson and Dr. Andreia Dionísio for their support and discussion on the analysis methodology, and to Andrew Hennell for his help on the preparation of this manuscript.

Correspondence concerning this article should be addressed to Eduardo Coutinho, University of Plymouth, B110 Portland Square, Plymouth, PL48AA, UK. E-mail: eduardo.coutinho@plymouth.ac.uk

References

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

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

Blood, A. J., Zatorre, R. J., Bermudez, P., & Evans, A. C. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature Neuroscience, 2*, 382-387.

Bradley, M., & Lang, P. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*, 49-59.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer.

Cabrera, D. (1999). PsySound: A computer program for psychoacoustical analysis. *Proceedings of the Annual Conference of the Acoustical Society of Australia* (pp. 47-54). Melbourne: Australian Acoustical Society.

Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception and Psychophysics, 57*, 451-462.

Damasio, A. (2000). *The feeling of what happens: Body, emotion and the making of consciousness.* London: Vintage.

Davidson, R., Scherer, K., & Goldsmith, H. (2003). *Handbook of affective sciences.* Oxford, New York: Oxford University Press.

Dionísio, A., Menezes, R., & Mendes, D. (2006). Entropy-based independence test. *Nonlinear Dynamics, 44*, 351-357.

Dowling, W. J., & Harwood, D. (1986). *Music cognition.* San Diego, CA: Academic Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-225.

Gabrielsson, A., & Juslin, P. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music, 24*, 68-91.

Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression. In P. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 223-248). Oxford, UK: Oxford University Press.

Granger, C., & Lin, J. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis, 15*, 371-384.

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

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika, 28*, 321-377.

Kivy, P. (1990). *Music alone: Philosophical reflections on the purely musical experience.* Ithaca, NY: Cornell University Press.

Korhonen, M. (2004). *Modeling continuous emotional appraisals of music using system identification.* Unpublished master's thesis, University of Waterloo.

Kremer, S. (2001). Spatiotemporal connectionist networks: A taxonomy and review. *Neural Computation, 13*, 249-306.

Krumhansl, C. L. (1991). Melodic structure: Theoretical and empirical descriptions. In J. Sundberg, L. Nord & R. Carlson (Ed.), *Music, language, speech and brain* (pp. 269-283). London, UK: MacMillian.

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception, 13*, 401-432.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology, 51*, 336-353.

Langer, S. K. (1942). *Philosophy in a new key.* Cambridge, MA: Harvard University Press.

Ljung, L. (1987). *System identification: Theory for the user* (2nd ed.). Old Tappan, NJ: Prentice Hall.

Malloch, S. (1997). *Timbre and technology: An analytical partnership. The development of an analytical technique and its application to music by Lutosławski and Ligeti.* Unpublished doctoral dissertation, University of Edinburgh, Edinburgh.

Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago, IL: University Of Chicago Press.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model.* Chicago, IL: University Of Chicago Press.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* New York, NY: Oxford University Press.

Panksepp, J., & Bernatzky, G. (2002). Emotional sounds and the brain: The neuro-affective foundations of musical appreciation. *Behavioural Processes, 60*, 133-155.

Parncutt, R. (1989). *Harmony: A psychoacoustical approach.* Berlin: Springer.

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

Rumelhart, D., Hintont, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

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

Russell, J. A. (1989). Measures of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience* (pp. 83-11). Toronto: Academic.

Scherer, K. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research, 33*, 239-251.

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

Schubert, E. (1999a). *Measurement and time series analysis of emotion in music.* Unpublished doctoral dissertation, University of New South Wales.

Schubert, E. (1999b). Measuring emotion continuously: Validity and reliability of the two dimensional emotion space. *Australian Journal of Psychology, 51*, 154-165.

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 393-414). Oxford, UK: Oxford University Press.

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

Sloboda, J., & Lehmann, A. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception, 19*, 87-120.

Thayer, J. (1986). *Multiple indicators of affective response to music.* Unpublished doctoral dissertation, New York University.

Tzanetakis, G., & Cook, P. (1999). MARSYAS: A framework for audio analysis. *Organised Sound, 4*, 169-175.

Wundt, W. (1896). *Grundriss der Psychologi*e [Outlines of Psychology]. Leipzig: Engelmann.

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics.* New York: Springer.

Notes

[1] In the screening tests, used to test I.R.s ability to process music, I.R. did not give any indication that she could perceive and/or interpret pitch and temporal variations in melodies.

[2] Data available online at http://www.sauna.org/kiulu/emotion.html, courtesy of the author.

[3] Although these algorithms are not specific measures of melodic contour, they have been used successfully as such in music information retrieval applications (Korhonen, 2004). Nevertheless, in this article we refer to this variable as pitch variation because it characterizes better the nature of the encoding. Moreover, the relationships between pitch variations and emotion were the object of some studies (e.g., Scherer & Oshinsky, 1977), as described in Schubert (1999a).

[4] T and Tx were chosen as the variables for the initial features for a few reasons. First is that they are the only variable for the sound features that they represent. A second important factor is that T and Tx are expected to contain important information about changes in the affective experience (Schubert, 1999a).

[5] In the tables we indicate no index when we include all variables from that music feature; in this case D indicates D_1 and D_2.

[6] X and Y are the data sets being compared and $I(X,Y)$ is the MI score.

[7] Both methods are very similar because they look for linear combinations of variables which best explain the data; the essential difference consists of the rules for classification (clustering), which is based on distance measures in PCA while LDA explicitly attempts to model the difference between the classes.

[8] This does not mean that we can reduce the number of units in the model, but instead that some of these units might vary along similar dimensions. As we'll see, all the hidden units have relevant contributions to at least one of the discriminant functions.

Table 1

Pieces Used in Korhonen's (2004) Experiment and their Aliases for Reference in this Paper.

Piece ID	Alias	Title and Composer	Duration	Set
1	Aranjuez	Concierto de Aranjuez - II. Adagio (J. Rodrigo)	165 s	Training
2	Fanfare	Fanfare for the Common Man (A. Copland)	170 s	Training
3	Moonlight	Moonlight Sonata - I. Adagio Sostenuto (L. Beethoven)	153 s	Test
4	Morning	Peer Gynt Suite No 1 - I. Morning mood (E. Grieg)	164 s	Training
5	Pizzicato	Pizzicato Polka (J. Strauss)	151 s	Test
6	Allegro	Piano Concerto no.1 - I. Allegro maestoso (F. Liszt)	315 s	Test

Note: The pieces were taken from the Naxos "Discover the Classics" CD 8.550035-36

Table 2

Psychoacoustic Variables Considered for this Study.

Musical Property	Musical Feature	Alias
Loudness Level	Dynamics	D_1
Short Term Maximum Loudness	Dynamics	D_2
Power Spectrum Centroid	Mean Pitch	P_1
Mean STFT Centroid	Mean Pitch	P_2
Mean STFT Flux	Pitch Variation	Pv_1
Standard Deviation STFT Centroid	Pitch Variation	Pv_2
Standard Deviation STFT Flux	Pitch Variation	Pv_3
Sharpness (Zwicker and Fastl)	Timbre	Ti_1
Timbral Width	Timbre	Ti_2
Mean STFT Rolloff	Timbre	Ti_3
Standard Deviation STFT Rolloff	Timbre	Ti_4
Beats per Minute	Tempo	T
Multiplicity	Texture	Tx

Note: These variables are indicated within the article by their alias.

Table 3

RMS Error for Each Input Data Set Using a Model with 5 Hidden Units.

Input Set	RMS Train		RMS Test		Mean RMS
	Arousal	Valence	Arousal	Valence	
T-Tx-D	.06	.06	.07	.08	.07
T-Tx-D_1	.06	.06	.07	.08	.07
T-Tx-D_2	.07	.06	.09	.08	.07
T-Tx-Ti	.07	.07	.09	.09	.08
T-Tx-Ti_1	.07	.06	.08	.09	.08
T-Tx-Ti_2	.11	.07	.10	.08	.09
T-Tx-Ti_3	.11	.07	.14	.12	.11
T-Tx-Ti_4	.11	.08	.13	.09	.10
T-Tx-P	.08	.07	.11	.10	.09
T-Tx-P_1	.07	.07	.11	.08	.08
T-Tx-P_2	.14	.08	.23	.11	.14
T-Tx-Pv	.10	.06	.12	.08	.09
T-Tx-Pv_1	.10	.06	.13	.09	.10
T-Tx-Pv_2	.11	.07	.13	.08	.10
T-Tx-Pv_3	.10	.07	.13	.09	.10
T-Tx-D_1 -P_1 –Ti_1 -Pv_1	.05	.05	.07	.08	.06

Note: The values shown were averaged across 3 simulations for each test case.

Table 4

Comparison Between the Model Outputs and Experimental Data: Root Mean Square (RMS) Error and Global Correlation Coefficient (λ).

Piece	RMS error		MI (λ)		Set
	Arousal	Valence	Arousal	Valence	
1	.05	.04	.94	.77	Training
2	.04	.05	.76	.87	Training
3	.06	.05	.75	.65	Test
4	.09	.08	.54	.56	Test
5	.04	.05	.90	.49	Training
6	.05	.08	.96	.74	Test

Table 5

Factor Structure Matrix: Correlations Between Discriminant Variables and Each Hidden Unit.

Hidden unit	F_1	F_2
H_1	-.49	-.25
H_2	.37	.90
H_3	-.79	.29
H_4	-.52	.57
H_5	.63	-.60

Table 6

Canonical Correlation Analysis (CCA)

Canonical Loadings (Input/Hidden)			
Variable	var. 1	var. 2	var. 3
H_1	-.40	-.63	-.03
H_2	.48	.66	-.44
H_3	.14	-.89	-.24
H_4	.16	-.65	-.63
H_5	-.64	.65	.02
T	.26	.48	.15
Tx	.61	.28	.22
D_1	.45	.67	.14
P_1	.82	.30	.43
Ti_1	.75	.42	.26
Pv_1	.19	.27	.83
Canon Cor.	.73	.55	.45
Pct.	61.1%	23.4%	13.8%
Wilks' L.	0.26	0.55	0.78
Sig.	.000	.000	.000

Canonical Loadings (Hidden/Output)		
Variable	var. 1	var. 2
H_1	-.50	.48
H_2	.98	-.06
H_3	-.29	.86
H_4	.01	.80
H_5	-.07	-.97
A	.77	-.64
V	.26	.97
Canon Cor.	.99	.98
Pct.	56.0%	44.0%
Wilks' L.	0.00	0.03
Sig.	.001	.000

Note: The canonical correlations (interpreted in the same way as the Pearson's linear correlation coefficient) quantify the strength of the relationships between the extracted canonical variates to assess the significance of the relationship. To assess the relationship between the original variables (inputs and hidden units activity) and the canonical variables, we also include the canonical loadings (the correlations between the canonical variates and the variables in each set),

Figure Captions

Figure 1. Neural network architecture and units identification (model used in simulations).

Figure 2. Training data set (Aranjuez, Fanfare and Pizzicato): Arousal and Valence model outputs compared with experimental data.

Figure 3. Test data set (Moonlight, Morning and Allegro): Arousal and Valence model outputs compared with experimental data.

Figure 4. Canonical Discriminant Functions plot: Each point corresponds to the internal state of the model at a particular moment in time. The dot's color identifies the internal states of the model belonging to each of the categories hypothesized (the affective space quadrants), and the labels (Q_1 to Q_4) indicate the correspondent quadrant in the 2DES to which each color group belongs to.