Running head: Emotion in speech prosody and music.

Psychoacoustic cues to emotion in speech prosody and music

Eduardo Coutinho[1, 2] and Nicola Dibben[3]

[1]Swiss Center for Affective Sciences, Geneva, Switzerland

[2]University of Liverpool, Liverpool, United Kingdom

[3]University of Sheffield, Sheffield, United Kingdom

Words: 11939

Tables: 7

Figures: 3

Correspondence address:

Eduardo Coutinho

Swiss Center for Affective Sciences, University of Geneva

7 Rue des Battoirs, Geneva CH-1205, Switzerland

E-mail: eduardo.coutinho@unige.ch

Date: 6 September 2012

Abstract

There is strong evidence of shared acoustic profiles common to the expression of emotions in music and speech, yet relatively limited understanding of the specific psychoacoustic features involved. This study combines a controlled experiment and computational modelling to investigate the perceptual codes associated with the expression of emotion in the acoustic domain. The empirical stage of the study provides continuous human ratings of emotions perceived in excerpts of film music and natural speech samples. The computational stage creates a computer model that retrieves the relevant information from the acoustic stimuli and makes predictions about the emotional expressiveness of speech and music close to the responses of human subjects. We show that a significant part of the listeners' second-by-second reported emotions to music and speech prosody can be predicted from a set of seven psychoacoustic features: loudness, tempo/speech rate, melody/prosody contour, spectral centroid, spectral flux, sharpness, and roughness. The implications of these results are discussed in the context of cross-modal similarities in the communication of emotion in the acoustic domain.

Key words: Emotion, Arousal and Valence, Music, Speech prosody, Psychoacoustics, Neural Networks

Emotional communication in speech prosody and music

In the acoustic domain, two types of stimuli are commonly regarded as especially effective means of conveying emotional meaning in everyday contexts: speech prosody (Scherer, 1986) and music (Gabrielsson & Juslin, 2003). As such they offer an opportunity to compare the means by which emotion is communicated through the auditory domain.

Speech prosody is the pattern of acoustic changes within spoken utterances that communicate meaning independently of verbal comprehension. Prosodic forms are a fundamental means by which speakers convey, and listeners understand, speaker emotions and intentions (Frick, 1985; Juslin & Laukka, 2003). The acoustic changes occur as modulations of tempo and continuity, accentuation, pitch and range, timbre and dynamics of speech and vocalizations. Similarly to facial expressions (Ekman, 1992), certain aspects of emotional expression in speech prosody appear to be universal (e.g., Scherer, Banse & Walbott, 2001) and convincing evidence points to similar prosodic codes used across cultures to convey similar emotions (e.g., Thompson & Balkwill, 2006). This is apparent, for instance, in our capacity to decode emotional meaning even in unfamiliar languages.

Music, like speech, has the capacity to communicate emotions to listeners through the organization of acoustic signals (see, for instance, Juslin & Sloboda, 2010). Although researchers differ in their conceptions of musical emotions, they agree when it comes to asserting music's capacity for emotional expression. As in the case of speech prosody, listeners construe emotional meaning by attending to structural aspects of the acoustic signal and there is evidence of specific acoustic cues and patterns communicating similar emotions to listeners (see Gabrielsson & Lindström, 2010). Although frequently regarded as subjective and culturally grounded, there is convincing evidence that music can express emotions that are recognized universally (e.g., Fritz, Jentschk, Gosselin, Sammler, Peretz, Turner, Friederici & Koelsch, 2009; Balkwill, Thompson & Matsunaga, 2004), a phenomenon that is associated with acoustic profiles which transcend cultural boundaries.

Only recently have researchers begun to compare directly acoustic cues to emotion in music and vocal prosody (Juslin & Laukka, 2003; Patel, Peretz, Tramo & Labreque, 1998; Ilie & Thompson, 2006, 2011). These studies provide evidence of the existence of acoustic profiles common to the expression of emotion in both speech and music, with particular acoustic codes consistently associated with particular emotions. In a review of 104 studies of vocal expression and 41 studies of music performance, Juslin and Laukka (2003) highlight similarities in the emotion-specific patterns of acoustic cues used to express discrete emotions, and conclude that this provides evidence for similar codes and shared neural resources. Subsequent studies directly comparing perception of emotion in music and speech prosody provide further evidence supporting these ideas and highlight the complexity of the relationships. Using a three-dimensional model of emotions to study three particular acoustic cues, Ilie and Thompson (2006) found a variety of effects and interactions: intensity was found to influence both prosody and music in the same way (greater intensity was associated with higher ratings of valence, tension and energy), rate had varying effects on both domains (faster music and speech was associated with higher energy, but whereas fast speech was judged as less pleasant than slow speech, fast music was judged as more tense than slow music), and pitch height influenced the two domains in opposite directions (high pitched speech and low pitched music were both attributed higher ratings of valence). Evidence that prosody and music share processing resources at an intermediate level in the auditory pathway comes from a study of perceptual discrimination in music and speech by two amusic subjects which showed that performance was similar across domains although the participants had different perceptual deficits (Patel et al, 1998).

This previous research into the induction of emotion by music and speech prosody provides a strong basis on which to purport the existence of a general mechanism for the expression and recognition of emotions in the acoustic domain and indicates the importance of shared acoustic features. However, we now need to understand the exact character of this general mechanism, and in our study we will focus on three main aspects.

First, given that music processing comprises a number of different systems, rather than a single "module", it may be that some of these are specific to music, while others apply to all types of auditory information (Ilie & Thompson, 2011). Therefore, direct comparison of the affective influence of music and speech prosody is necessary to determine the extent of overlap.

Second, it remains to be seen the extent to which particular acoustic features relate to a fuller range of emotional states, and to continuous as opposed to discrete emotion judgments. Virtually all studies have focused on the communication of full-blown, discrete emotional states such as happiness, sadness, fear or anger, and very few have considered the wide range of strong and nuanced emotional states expressed by speech and music in daily life – one exception being research using the Geneva Emotional Music Scales (Zentner, Grandjean & Scherer, 2008).

Third, derivation of acoustic profiles from the study of discrete emotional states leads to comparative descriptions of different emotions and the analysis of acoustic factors in terms of extreme levels (e.g., high/low, slow/fast, as seen in Juslin & Laukka, 2003, p. 802, Table 1). The outcome is a set of qualitative classifiers particular to certain emotions, which tend to be generalized to other states by assuming that intermediate states can be extrapolated from these extremes and that the various acoustic parameters do not interact meaningfully. Although those procedures can be useful in certain contexts, such as the recognition of basic emotions, they can result in misleading and contradictory results if applied within a broader context, impairing identification of the prosodic and musical forms associated with the expression of emotion.

Motivated by these gaps in current understanding, and by recent findings in the music domain (Coutinho & Cangelosi, 2009; Coutinho & Cangelosi, 2011), this research project adopts a transdisciplinary method that combines empirical experiments and computational modelling to investigate the psychoacoustic codes associated with the expression of emotion in the auditory domain.

Our theoretical framework conceptualizes the communication of emotion via music as part of the broader study of human and animal vocal and emotional communication, in which the primary function of acoustic communication is to exploit listeners' sensitivity to acoustic information to act upon the hearer and influence them socially, as evidenced in studies of vocal acoustics (Bachorowsky & Owren, 2003), infant directed speech (Schachner & Hannon, 2011), laughter (Bachorowsky & Owren, 2001), and primate studies (Owren & Rendall, 1997). Our focus on stimulus properties is consistent with this approach. Our premise is that musical structure elicits in the listener responses in neurological mechanisms, and bodily changes associated with motivation, cognition and emotion, and that these use the same mechanisms as are recruited by other emotionally significant external stimuli. Evidence supporting this framework comes from the large number of empirical studies which reveal systematic relationships between musical structures and emotional responses (Gabrielsson & Lindström, 2010), and neuroscientific studies suggesting that music can elicit emotions in listeners without the need for cognitive attributions (e.g., Peretz, 2001) or mediation (e.g., Blood & Zatorre, 2001). Specifically, we argue that music evokes emotion by creating dynamic temporal patterns to which our evolved socio-emotional brain is particularly sensitive.

In previous work (Coutinho & Cangelosi, 2009, 2011), Coutinho proposed that spatio-temporal low-level acoustic patterns convey fundamental information about listeners' perception of emotion in music. In order to test that hypothesis, Coutinho and Cangelosi devised a computational method sensitive to the temporal structure of sound that could predict a significant part of listeners' affective response from a set of six psychoacoustic features (basic variables of human audition that are perceived similarly across individuals and shared across acoustic modalities): loudness, pitch level, pitch variation (contour), tempo, texture and sharpness. The strength of this approach, compared to a behavioural study in which the influence of various psychoacoustic features are assessed by means of multiple regression, is that it provides a systematic means of extracting complex relationships from continuous data as is found in ecologically valid materials. The computational model can handle relationships among simultaneous features (thus considering interactions between various psychoacoustic dimensions; see Webster & Weir, 2005), and can incorporate memory of the past states of features, thus reflecting the dynamic and temporal character of emotional experience and associated auditory attributes. In this article we extend that work beyond the musical domain to speech, allowing investigation of a general mechanism for the expression and recognition of emotions in the acoustic domain.

Overview of the present study

Our research combines a behavioural study and computational model in order to determine the extent to which low level acoustic features convey emotion in music and speech.

In the behavioural phase, participants listened to excerpts of music and speech and rated the emotions perceived. Rather than manipulate a small number of acoustic attributes we elicited emotion judgments on unaltered stimuli to allow a fuller investigation of a range of possible features and states. This is the first time, to our knowledge, that any direct comparison of emotion communicated by music and speech prosody has been investigated using continuous evaluations of emotion, and therefore capturing the dynamic, temporal character of emotional experience.

Emotional responses to the stimuli were recorded along two axes: valence and arousal. This two–dimensional model of affective space is adopted from Russell's (1980) circumplex model of emotion, which represents specific emotions as points within a two dimensional space, located in terms of their relative valence (ranging from positive to negative affect) and arousal (ranging from high to low neurophysiological alertness). The model has been shown to account for a large proportion of variance in the emotional labelling of linguistic (Russell, 1980), pictorial (Bradley & Lang, 1994), and musical stimuli (e.g., Schubert, 1999b). Furthermore, there is evidence that arousal and valence are subserved by distinct neural systems during the experience of induced emotions (Colibazzi, Posner, Wang, Gorman, Gerber, Yu, Zhu, Kangarlu, Duan & Russell, 2010). Notably, dimensional models are less able than discrete models to capture mixed emotions of simultaneously positive and negative valence (Larsen, 2011). Nonetheless, the two are largely compatible (Eerola & Vuoskoski, 2011). A two-dimensional model was adopted in this study because, as explained above, it facilitates representation of a wide range of mild and full-blown emotions, allows direct comparison of music and speech emotional ratings, and because it is reliable and economical and can be implemented for the collection, analysis and modelling of continuous data.

This study focuses on emotion portrayed by music and speech stimuli, as distinct from the emotion felt by the listeners. This is an important distinction since emotion recognized and emotion felt are not necessarily identical (Gabrielsson, 2002). Indeed, there is preliminary evidence that perception of emotion in music and speech differs from the emotion induced, although with no clear pattern of responses (Ilie & Thompson, 2011).

In the second part of our study a computational model was used to predict the behavioural data. Excerpts were represented in terms of a number of acoustic features deemed to be relevant by previous studies (see Juslin & Laukka, 2003, p. 802, Table 11). The computational model then predicted listeners' responses to excerpts based on the most successful combination of these features. This revealed which acoustic features in music and/or speech was best able to model listeners' emotion judgments. As discussed above, the advantage of this approach is that it can capture a fuller spectrum of emotional states as they change over the time course of any music and speech.

This study therefore has two main aims. First, to determine the shared acoustic forms communicating emotions across music and speech. Second, to make detailed predictions of the emotional expression of natural speech and music. The purpose of the experiment is to collect continuous ratings of emotions perceived in music and speech stimuli. The goal of the computational study is to create a model capable of predicting participants' subjective feelings of emotion (the data collected in the empirical experiment), using low-level psychoacoustic features. Using these means we determine the extent to which the dynamics of emotional responses to music and speech are the result of the perception of emotional meaning via shared psychoacoustic patterns.

Method

Empirical phase

The experiment collected continuous ratings of emotions perceived in film music and natural speech. A continuous response method was used to obtain fine-grained temporal variations in reported emotional experience. This allows participants to report changes in their emotional state at any moment, instead of doing so only at the end of the piece, and has previously been used successfully in studies of emotional responses to music (e.g. Grewe, Nagel, Kopiez & Altenmüller, 2007b). The output from this self-report method is a time series depicting the dynamics of participants' ratings of emotion at every moment in the music. This continuous method of data collection is central to our study, which hypothesizes that the temporal structure of acoustic features communicates emotional information.

Participants

Sixty volunteers participated in the experiment. Two listeners were excluded from the analysis due to measurement errors, and a further six participants were removed, whose native language was not English. The purpose of the latter selection was to minimize variability in responses to the language samples. The final dataset used for the analysis consists of 52 participants (mean age = 31, SD = 13, range = 18-61 years; 26 females and 26 males). Participants had a range of musical education and experience (median = 5-10 years of formal music training) 15 participants had none or less than one year; 17 participants had ten years or more of music education or practice. Participants also reported enjoying film music (the mean rating was 3.9 on a 5-point Lickert scale where 1 corresponds to "I hate film music" and 5 corresponds to "I love film music"), and all but one reported being exposed to film music at least "occasionally".

Materials

The stimulus materials consisted of eight extracts of music, and nine samples of speech. In order to achieve comparable ecological validity, the music stimuli were excerpted from late twentieth century Hollywood film scores and the speech stimuli were excerpted from commercially available and online film, dramatic performances, poetry recitations and interviews; these genres of music and speech are intended to affect listeners emotionally and are widely experienced by the general public. The music and speech excerpts were selected from a larger set of twenty speech and twenty music pieces, displaying a wide emotional range, and in the case of the music stimuli, a variety of instrumentation. Sub-selection was made from this set via pre-testing with fifteen student and staff participants from the University of Sheffield using a paper-based self-report two-dimensional affect space. Selection of the final set of stimuli was determined by three criteria: highest consistency of emotion rating among respondents; widest coverage of the two-dimensional emotional space (2DES); a diversity of psychoacoustic dimensions represented by the set as a whole (e.g. instrumentation, loudness, tempo). The stimuli were up to two and half minutes in length, in order to allow measurement of dynamic changes in affective experience, and to keep the total experiment less than thirty minutes in duration. A dimensional reduction from ratings performed on a large set of stimuli is often ideal in stimuli selection (Eerola & Vuoskoski, 2011) but would not have been able to discriminate stimuli according to the criteria needed for this experiment.

The music used is shown in Table 1. The emotion communicated by each piece as determined by pre-testing is indicated by the labels Q_1 to Q_4, which represent the four main areas resulting from a division of the 2DES arousal/valence diagram into quadrants: Quadrant 1 (Q_1) – positive arousal and positive valence, Quadrant 2 (Q_2) – positive arousal and negative valence, Quadrant 3 (Q_3) – negative arousal and negative valence, and Quadrant 4 (Q_4) – negative arousal and positive valence.

The speech samples were all chosen from a single language not understood by participants. This was necessary in order to avoid any confounds due to the necessarily different semantic content of ecological speech samples. German was selected due to evidence in previous research that native English speakers not conversant in German are able to decode the emotional nuances of German prosody (Scherer et al, 2001; Thompson & Balkwill, 2006). The speech samples used are shown in Table 2. As above, the emotions communicated by each excerpt, as determined by the pre-test, is indicated by the labels Q_1 to Q_4.

-- Insert Table 1 here --

-- Insert Table 2 here --

Participants also completed a questionnaire collecting information on demographics. This included 5-point Lickert scales for musical training, musical exposure, and musical enjoyment.

Procedure

Each participant sat comfortably in a chair inside a quiet room. The goal of the experiment was explained through written instructions that described the quantification of emotion and the self-report framework to be used during the listening task. Participants reported their emotional state by using software constructed by the first author, which consists of a computer representation of a two-dimensional emotional space (2DES). Physiological data (blood volume pulse, electrocardiography, skin conductance, and respiration rate) was collected using the ProComp5 Infiniti encoder: participants had sensors attached to their left-hand (if right hand – if not, sensors were attached to the left hand), their chest, and wore a strap around the chest. The physiological data is not reported here since it pertains to a related but separate study.

Participants were given the opportunity to practice with the self-report framework using ten pictures taken from the International Affective Picture System manual (Lang, Bradley & Cuthbert, 2005). The selected pictures represented emotions covering all four quadrants of the 2DES (two per quadrant), and the neutral affective state (centre of the axes). The pictures were shown in a nonrandomized order, in order to avoid starting or finishing the picture slideshow with a scene of violence. Each picture was shown for 30 seconds, with a ten seconds delay between presentations. The only aim of this exercise was to familiarize participants with the use of the self-report framework.

After the practice period, participants were asked about their understanding of the experiment, and whether they felt comfortable in reporting the intended affective states with the software provided. Participants were then reminded to rate the emotions thought to be expressed by the music and speech stimuli, and not the ones felt. When the participant was ready, the main experiment started and the first stimuli was played. The stimuli were presented in a randomized order, with a break of 75 seconds between each excerpt (unless the participant needed more time). Each experimental session lasted for about 60 minutes, including debrief, preparation and training periods.

Data Processing

The arousal and valence reported by each participant was recorded from the mouse movements. These values were normalized to a continuous scale ranging from -1 to 1, with 0 as neutral. The central tendency of the individual values of arousal and valence was estimated by calculating the arithmetic mean across all participants, on a second by second basis, for each sound stimulus.

Computational phase

In this part of the study, we employ a computational framework to model the emotion reported by the group of participants that listened to the music and speech stimuli in the empirical study. The first aim is to create a model that is able to learn (from a subset of stimuli) how the dynamics of arousal and valence relate to the psychoacoustic structure of the stimuli heard. The second is to use that "knowledge" to predict as accurately as possible the emotional responses to novel stimuli. The third is to compare music and speech models and to establish whether similar psychoacoustic cues are involved in the communication of emotion for both mediums.

Our first hypothesis is that low-level psychoacoustic features strongly relate to emotion reported by listeners. Our past research revealed that a large percentage of the emotions perceived in music could be consistently inferred from low-level psychoacoustic dimensions (loudness, pitch level and contour, tempo, and timbre; Coutinho & Cangelosi, 2009, 2011). Consequently, we want to know which features are most relevant for the communication of emotions in music and in speech, and to what extent emotionally congruent qualities of these features are reliable predictors of the emotions perceived by human listeners. Our second hypothesis is that a similar set of psychoacoustic cues will be involved in the expression of emotion in both mediums, as per the previous research discussed above.

Computational Framework: Elman Neural Network

Due to their adaptability to deal with patterns distributed across space (relationships among simultaneous features) and time (memory of the past states of the features), Recurrent Neural Networks (RNNs) were used in previous work by the first author to model emotional responses to music (Coutinho & Cangelosi, 2009, 2011). Specifically, this previous work used a type of RNN called an Elman Neural Network (ENN) (Elman, 1990). This model consists of the traditional feed-forward multi-layered perceptron (Rumelhart, Hinton & Williams, 1986) with added recurrent connections on the hidden layer that endow the network with a dynamic memory. While the basic feed-forward network can be thought of as a function that maps from input to output vectors, parameterized by the connection weights, and capable of instantiating many different functions, the ENN can map from the history of previous inputs to predict future states in the output. The key point is that the recurrent connections allow the sequence of internal states of an ENN to hold not only information about the prior event but also relevant aspects of the representation that were constructed in predicting the prior event from its predecessor. 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 which are also sensitive to the temporal structure of the inputs. An overview of artificial neural networks theory, a detailed description of the model, as well as its application to the prediction of emotional responses to music can be found in Coutinho and Cangelosi (2010).

In the context of the present study, we use the ENN as a nonlinear regression model. The dependent variables are the self-reported emotion dimensions (psychological arousal and valence) that correspond to the subjective feelings of emotions perceived by listeners while listening to each stimulus. The independent variables are the sensory and perceptual quantities (psychoacoustic features) extracted from the audio signals (music and speech stimuli), which describe the low-level psychoacoustic structure of the auditory object.

Procedure

Model architecture

The ENN architecture consists of a three-layered simple recurrent neural network comprising: 1) an "internal state" or "hidden" layer; 2) a "memory" or "context" layer; and 3) an output layer (which yields the arousal and valence outputs). Additionally the network has an extra layer holding the input units, which receives and processes the functional data of the psychoacoustic features (see Figure 1)[i].

-- Insert Figure 1 --

The size of the input layer depends on the number of features used in each modelling experiment. We test different groups of inputs, therefore the size of this layer is variable. The size of the hidden and memory layers is identical (the memory layer holds a copy of the hidden layer activations at the previous time step) and it was set to ten, a value optimized in a preliminary set of simulations. The output layer size is two: one unit for arousal and another for valence.

Inputs

Our main hypothesis for this experiment is that low level music structural features show causal relationships with listeners' reports of emotion. To extract such information from the music and speech signals we quantified different psychoacoustic dimensions describing each auditory object. We focused on five main classes of features: loudness, duration (tempo/speech rate), pitch (melody/prosody contour and pitch variation), timbre ("brightness" and sharpness), and roughness (including spectral dissonance). These basic variables of human audition show consistent relationships with emotional arousal and valence (Gabrielsson & Lindström, 2010; see also Coutinho & Cangelosi, 2009, 2011). A summary of the features selected for this experiment is given in Table 3. The following paragraphs provide a brief description of each feature and algorithms for their estimation.

-- Insert Table 3 here --

Loudness Loudness is the perceptual correlate of sound intensity (or physical strength) which we quantified using Chalupper and Fastl's (2002) dynamic loudness model (measured in sones).

Duration The measures of duration consist of the rate of speech and musical tempo. The former was estimated using De Jong and Wempe's (2009) algorithm, which detects syllable nuclei and quantifies speech rate as the number of syllables per minute (SPM). The latter was estimated from the inter-beat intervals obtained for each piece using BeatRoot (Dixon, 2006), and quantified as the number of beats per minute (BPM).

Pitch The perceived pitch level and pitch contour were calculated separately for music and speech. The prosodic contour was calculated using Prosogram (Mertens, 2004), a prosody transcription tool that estimates the intonation contour (the perceptual correlate of the fundamental frequency, F0), as human listeners perceive it. The melodic contour was estimated using Dittmar, Dressler, and Rosenbauer's (2007) toolbox for automatic transcription of polyphonic music. The contour curve yielded by this algorithm estimates a salient stream of audible pitches from the full harmonic structure of the polyphonic signal. In addition to these measures we also calculated the spectrum flux for all stimuli in order to quantify how much the power spectrum of the signal changes in time[ii].

Timbre Timbre is a multi-dimensional attribute and has been associated with many different psychoacoustic attributes. Two of the most commonly used features are “brightness” and “sharpness”. We quantify timbre using: 1) The power spectrum centroid which is calculated by the weighted mean of the frequencies present in the signal (weights being the Fourier transform magnitudes for each frequency band), a quantity strongly associated with the impression of sound “brightness”); and 2) Two sharpness measures: one proposed by Zwicker and Fastl (1999) and another by Aures (1985); Aures’ sharpness formula is a revision of Zwicker and Fastl’s, which considers the positive influence that loudness has on sharpness (both approximate the subjective experience of sharpness on a scale ranging from dull to sharp, and are measured in acum).

Roughness The term auditory roughness describes the perceptual quality of buzz, raspiness or harshness associated with narrow harmonic intervals. In a complex sound (a sound comprising several partials or pure tone components), any two or more partials less than a critical distance apart can lead to the auditory experience of “beating” or “roughness”. This effect is associated with the inability of the basilar membrane to separate the sounds clearly. Roughness is also a perceptual correlate of dissonance, a concept that has acoustic and physiological bases, as well as cognitive and cultural ones (Vassilakis, 2001). The psychoacoustic dimension of dissonance more closely associated with roughness is known (among other terms) as auditory dissonance (Hutchinson & Knopoff, 1978). We use one measure of psychoacoustic roughness (Daniel & Weber, 1997) and two algorithms to measure auditory (spectral) dissonance (Hutchinson & Knopoff, 1978 and Sethares, 1985).

Outputs

The model outputs (the dependent variables) are the arousal and valence time series averaged across participants (see "Empirical experiment" section). The values are solely the result of the network processing.

Training and test data sets

One of the interesting aspects of learning systems, such as the ENN, is their ability to generalize from a subset of the data. This means that if the ENN can learn a set of dynamic rules from a sample data set (the "training set") that constitute a possible solution to the problem being modelled, then the model can subsequently be tested to predict the output data vector to another subset of unseen stimuli (the "test set"). When this is successful the model solution is likely to be generalizable beyond the data set.

The model performance with unseen data sets at each moment of the learning stage is fundamental for testing its generalizability. It is necessary to monitor the performance error for both sets in order to prevent "over-fitting" the model that occurs when a model begins to memorize the training data rather than finding a general solution to the problem. It is therefore vital to ascertain that the error value for the test data set is as low as possible, and that the test data set is large enough (at least comparable to the size of the training set) for the generalization to be robust.

Music and speech stimuli were processed separately. We divided each set of stimuli (8 music pieces and 9 speech samples) into two groups balanced in terms of coverage of the 2DES and total duration. This division ensures maximum coverage of the 2DES by the training set thereby providing a good sample of the solution space, and a test set which is long and varied enough to incisively assess the model reliability. The distribution of the stimuli amongst the training and test sets for both music and speech are as follows (the numbers indicate the stimuli ID's indicated in Table 1 and Table 2): Training set (Music)={3, 4, 5, 7} (462 s); Test set (Music) = {1, 2, 6, 8} (401 s); Training set (Speech)={1, 3, 6, 7, 8} (376 s); Test set (Speech)={2, 4, 5, 9} (375 s).

Results - Behavioural Phase

We first explore data from the behavioural phase of the research before proceeding to the computational stage.

Reliability of the means

The internal consistency of participants' ratings of emotions perceived was tested using Cronbach's Alpha, a measure of the reliability of the mean, calculated across all participants. The average reliability across all stimuli was high for both reported arousal (.95) and valence (.80), although participant responses were more consistent for the first. Cronbach's Alpha scores for each individual piece and speech sample showed very high consistency (all >.80), with the exception of the arousal score for speech sample 8 (.65) and the valence scores for speech samples 6 (-.73) and 9 (.15). These findings indicate that the emotion ratings are sufficiently consistent as to be suitable for further analysis and modelling.

Stimuli

We tested the representativeness of the stimuli as regards coverage of the 2DES affect space. The music stimuli elicited responses in the predicted quadrants of the 2DES (see Table 1). As intended, responses for individual pieces showed changes during the time course of the stimuli and responses across the set of stimuli covered all four quadrants of the 2DES (see Figure 2a). Ratings are slightly skewed towards the higher half of the arousal dimension in the 2DES, which may reflect the particular stimuli chosen.

Similarly, the speech extracts elicited responses in the predicted quadrants of the 2DES (see Table 2). Responses for individual pieces showed changes during the time course of the stimuli and responses across the set of extracts covered all four quadrants of the 2DES (see Figure 2b).

Overall, these results indicate that the stimuli provide a good foundation from which to model emotion communicated by music and speech: the responses show internal consistency and elicit emotions in different areas of, and temporally distributed across, the 2DES affect space.

Results - Computational Study

Simulations

The term "simulation" refers to the training of an ENN to output the emotion features for the training stimuli set as close as possible to the emotional dimensions reported by human subjects. Each simulation consists of a set of 40 trials in which the same model is trained using different initial conditions (randomized weights for all connections: values distributed between -0.05 and 0.05, except for the connections from the hidden to the memory layer which are set constant to 1.0). This procedure verifies the consistency of the results across simulations. Each trial consists of 120000 iterations of the learning algorithm, implemented using a standard back-propagation technique (Rumelhart et al., 1986). During training the same learning rate (0.1) and momentum (0.0) were used for each of the three connection matrices. Further details related to the learning algorithm and procedure used can be found in Coutinho & Cangelosi (2010).

Having fixed the network parameters and defined the output vector, the only aspect that changes from one simulation to the next is the set of inputs used. By trying different vectors of features as inputs to the model we test varied configurations of psychoacoustic features and evaluate the model performance for each. The quantification of the deviation of the model outputs from the values observed experimentally is implemented using the root mean square error (*rmse*), which is a measure of precision. For each trial the training stop-point was estimated *a posteriori* by calculating the ideal number of training iterations so as to minimize the model output error (i.e., the *rmse*) for the test set, thus avoiding the over fitting of the training set. After identifying the iterations leading to the best model predictions, we then picked the five best trials (lowest *rmse* for the test set) from each simulation and averaged their outputs. The error statistics presented in this section correspond to the deviations between experimental data (averaged across participants) and model outputs (averaged across the five best trials).

Simulations for music and speech stimuli were conducted separately since the use of speech stimuli is a new addition to our modelling work and because some of the input features vector may differ from that for music (e.g., tempo vs. speech rate). In the first battery of simulations, we start with a set comprising only variables describing speech and music stimuli in terms of loudness, tempo/speech rate and melodic/prosodic contour, and we monitor the extent to which each addition of the remaining psychoacoustic cues improves the model performance. Thus, in each simulation, the input vector consists of a basic set of three psychoacoustic groups – loudness (i1), duration (i2m/i2s), and pitch (i3m/i3s) – alone, plus each of the remaining features (i4-i10) (adding up to eight initial simulations; see Table 4 and Table 5, simulations #1 to #8). The decision to start with these three features as the basis for the input vector is due to their consistent association with the perception of emotion in music (Gabrielsson & Lindstrom, 2010; Coutinho & Cangelosi, 2009, 2011) and prosody (Juslin & Laukka, 2003). In these initial simulations we monitored how much each single feature produced an effect on the model performance when compared to the model output that only used the basic set of features ({L T mC} for music, and {L SR pC} for speech). Performance-enhancing variables were selected for further testing.

In consecutive simulations we evaluated several combinations of the selected input features, until a final input vector was found: the one that leads to the lowest test set error, and a good balance between arousal and valence outputs prediction accuracy. Given that in all simulations only the composition of the input vector is changed (training and test sets are the same; and the network parameters, except the weights, are kept constant across simulations), the selection of the final model (the one with lowest error) is directed towards the identification of the set of psychoacoustic features that permits the most accurate predictions of the emotional responses to the stimuli used. In what follows we present the preliminary simulations used to determine the set of acoustic features which best predicts the ratings of emotion, followed by a detailed analysis of each model.

Music preliminary simulations: input vector optimization

The results corresponding to the first battery of simulations with music stimuli is shown in Table 4 (simulations #1 to #8). To evaluate the contribution of each new variable added to the basic set, we compared the error of simulations #2 to #8 with the error of simulation #1 (L, T, and mC) to evaluate the contribution that each new input brings to the model. In particular, we are interested in identifying those features that lead to an improvement in the model performance for the test data set, i.e., the generalization of the knowledge extracted from the training set to a new sample of music.

-- Insert Table 4 here --

In the table, we highlighted in bold those values that are at least equal to the basic model (sim. #1) error. Those that produced lower error are additionally underlined. Concerning the arousal dimension of the training data set, all simulations produced very similar errors (see Table 4). With respect to the test data set error, little is learned from the variables tested, which suggests that the basic input set (L, T, mP) contains the essential information that the model can use to generalize the knowledge extracted (considering that the improvements to the training set error didn't lead to better generalization). The results pertaining to valence show a different picture. Indeed, the features used in some of the simulations contain additional information with relevance for valence predictions.

Having found improvements by adding new input features to the model, we identified those that led to the best results. To do so we focused on those simulations that produced improvements in the generalization performance. With this premise, the features from simulations #2, #3, #5 and #6 are the most obvious candidates, because, as mentioned before, they lead to better generalization of valence ratings. In the remaining simulations (#4, #7, and #8), although there were some improvements to the training error, they did not lead to a generalization performance, hence can be interpreted as an over-fitting of the training data and, consequently, the feature tested was not used in further simulations.

Having selected four input features for further testing - SC, SF, Sa, and SDhk - we proceeded by creating a new battery of simulations to evaluate a new set of input vectors. They consist of the basic set (L, T, and mC), plus every combination of two variables selected in the first set of simulations (SC, SF, Sa, SDhk). This sums to a total of six new simulations. The input vectors and results for the new simulations are shown in Table 4 (simulations #9 to #14). To identify the best simulations we compared the error values in each new simulation with the corresponding minimum errors in the simulations, taking each variable separately. The minimum errors are those values highlighted in bold in Table 4[iii].

The error figures show that several of the new simulations were able to bring together the information from the new sets of inputs, and match and improve the performance of the model using the input variables separately (see Table 4). Simulations #9 and #10 have some apparent advantages in relation to other simulations in terms of arousal and valence. They conciliate the lowest errors for training and test arousal, with the lowest error for test valence. Simulation #14 was particularly good for arousal (lowest training and test errors across all simulations so far). The input vector includes the basic set of variables plus S_a and SD_{hk}, but the performance for valence ratings is worse than in other simulations. These three simulations have in common four input features: two share SC, two share S_a, one contains SF, and another SD_{hk}. These variables were selected for further analysis.

Finally, to choose the best model we tested the subset of four variables together in the same simulation. There is only one doubt and this pertains to the SD_{hk} input: in all simulations in which it appeared the test valence errors were worse than other simulations (see simulations #11, #13, and #14). Consequently, we conducted two simulations instead of one: L, T, mP + SC, SF, S_a (sim. #15) and L, T, mP + SC, SF, S_a, SD_{zf} (sim. #16). Results are shown at the bottom of Table 4 (again we highlighted in the best error conditions following the procedure explained above). Simulation #15 clearly produced better results than simulation #16. In addition to this, the input set used in simulation #15 led to the lowest error predictions of valence ratings when compared to all other simulations. As expected, the predictions of arousal ratings are not improved (note that earlier we observed that none of the new features improved the predictions of arousal), reinforcing the idea that the new features do not contribute distinct information leading to the inference of relevant rules about arousal dynamics.

The final set of inputs selected for the music model consists of the input vector used in simulation #15, which included: loudness (L), tempo (T), melodic contour (mC), spectral flux (SF), spectral centroid (SC), and sharpness (S_a). They represent four main psychoacoustic groups: dynamics (L), duration (T), pitch (mC and SF), and timbre (SC and S_a).

The following section describes the simulations focused on speech stimuli. Later the music model is analysed in more detail and compared to that for speech.

Speech preliminary simulations: input vector optimization

The analysis for speech simulations was performed identically to that for music. First, we tested each variable independently in separate simulations. Then we selected those features that potentially improved the model performance and tested all possible combinations of this subset of features and again evaluated how much they improve the model performance. This process was repeated until the best set of input features was found.

-- Insert Table 5 here --

The results corresponding to the first battery of simulations with speech stimuli are shown at the top of Table 5 (simulations #1 to #8). All simulations produced similar errors for the training set stimuli. In fact, little new information seems to be extracted from the added input features. Through the observation of the performances in shown in Table 5, we chose the features included in the input vectors of simulations #2, #5, #7, and #8: SC, S_a, SD_s, and R. We dropped S_{zf}, used in simulation #4 with good results, because it quantifies the same quantity as S_a (sim. #5) but using a different algorithm. On top of that, S_a led to a better performance than S_{zf}

for test valence. SF, used in simulation #3, was also dropped because it worsened the result for test valence and it does not improve the performance for arousal (compared to sim. #1).

The next step was to set up another battery of simulations to test all possible pair-wise combinations of the newly selected variables (plus the basic set of features; see Table 5, simulations #9 to #14). Again we highlighted the lowest error values using the same method described for the music model input vector optimization. Simulations #9, #11 and #13 yielded the lowest test valence error of all simulations, a more balanced error between arousal and valence, and the lowest error values for the test set. The input vectors for these three simulations include: SC (#9 and #11), Sa (#9 and #13), and R (#11 and #13). These variables are chosen for further testing.

The last stage is to test all the performance enhancing variables together in the same input vector of a single simulation. Results are again shown in Table 5 (see simulation #15). This simulation produced the lowest error for the test arousal and valence. The input vector used in this simulation is composed of variables describing five psychoacoustic dimensions: loudness (L), duration (SR), pitch (pC), timbre (SC and S_a), and roughness (R).

The following section provides a detailed analysis of the results obtained for the music and speech models using the selected input vectors. We also compare the results obtained for both domains.

Analysis

The following sections provide a detailed analysis of each model for all stimuli. It provides a closer look at the model performance and resemblance to subjects' responses using the input sets optimized in the previous section for the music and speech models. We chose four measures to depict in more depth the model performance and to allow subsequent comparison of speech and music models: 1) similarity between model outputs and observed values - Pearson linear correlation coefficient (*r*); 2) explained variance (r^2); 3) precision of predictions - root mean squared error (*rmse);* and 4) standardized precision of predictions - normalized *rmse* (*nrmse*). This last measure was introduced to interpret the magnitude of the *rmse* in relation to the range of observed values since its interpretation depends on the variation of each piece. Furthermore, it allows us to compare directly speech and music models which is central to our aim of identifying acoustic cues used in music and speech.

Music model

The performance measures calculated for the music model are shown in Table 6.

-- Insert Table 6 here --

The model is able to explain a very large proportion of the variance in arousal of the training (97%) and test (75%) sets, which suggests a strong resemblance between model predictions and observed data. In regard to the similarity between predicted and observed values, the residuals were estimated to correspond to 21% of the total range of values (averaged across all pieces, after being calculated independently for each piece). Both measures indicate a very good performance for arousal, although the model more accurately predicted the outputs for the training set (15%) than for the test set (27%).

In regard to valence, the model was able to explain 83 % of the variance in the training set and 43% of the test set. These figures are lower than those for the arousal output, which indicates less similarity between model predictions and observed values (although this mainly derives from the small amount of explained variance in pieces 2 and 8). Nevertheless, the precision was similar to that of arousal: the *nrmse* was 21% for the training set and 30% for the test one.

Concerning this discrepancy, it should be noted that there are important limitations when using r as a measure to quantify the similarities between functional data sets (and consequently r^2 to estimate the explained variance), particularly when modelled with a non-linear framework. With flat curves (i.e., small variance) the best linear fit is a horizontal line passing through the mean. Considering that the model can output nonlinear, translated into small random variations, then the linear correlation coefficient becomes a non robust performance measure. Indeed, those deviations from linearity will increase the total sum of squared distances from the regression line even if they represent a very close relationship between the two variables. This can be made worse by the presence of just a few outliers.

In this context it is helpful to observe the model outputs in relation to the subjects' responses for all pieces, represented in Figure 2a which displays arousal and valence as orthogonal dimensions in a two-dimensional Cartesian space. To gain an even better insight into the continuous ratings predicted by the model see Figure 3a, which shows the arousal and valence time series for two sample pieces from the test set.

-- Insert Figure 2 here --

-- Insert Figure 3 here --

The overlap of the human participant responses and the model outputs in Figure 2a shows how closely the model resembles the participants' responses. Consistent with the statistics presented in Table 6, the individual charts show that the model is able to capture the "affective journey" of the different pieces. The accuracy is quite remarkable for several of them, as can been seen by inspecting the results for pieces 1, 3, 4 or 7. It is also evident the model fails to accurately predict the emotional qualities of some sections of a few pieces. For instance, the model tends to predict higher valence than expected for pieces 5 and 8, and fails to identify correctly the arousal level for a portion of piece 6.

These results indicate that the model can extract relevant information from psychoacoustic features to predict emotional characteristics of music, and that the knowledge derived is likely to represent meaningful relationships between psychoacoustic structure and emotions conveyed. In that sense, the model performance is quite remarkable. Nevertheless the model cannot account for aspects of the emotions perceived by listeners, either due to intrinsic mathematical limitations of the model or, very likely, because the input vector does not contain enough information to explain that variance.

Speech model

The model performance for each speech sample is shown in Table 6. Figure 2b shows the 2DES representation of the participants' ratings as well as the model outputs (similarly to the representation of the results for music).

-- Insert Table 7 here –

The model was able to explain 89% and 81% of, respectively, the training and test arousal, which is a large proportion of the observed variance. The similarity between participants' responses and model predictions was also high. Indeed, the *nrmse* for arousal predictions was 10% (training set) and 21% (test set). The explained variance in arousal is higher than that of valence for both training and test sets. The value of r^2 averaged across all samples for the arousal dimension is 86%, whereas for valence it is 67%. In terms of similarity we have a different picture: both models have very similar *nrmse* (16% for arousal and 15% for valence).

These statistics indicate a good fit of the speech model in predicting the human participants' responses. Such observation can be verified in Figure 2b, which superimposes in the 2DES the continuous model predictions and (target) participants' continuous judgment of emotions perceived in each speech sample.

The most obvious exceptions are sample 5, in which arousal was underrated, and sample 2, where both arousal and valence were overrated. In order to observe the unfolding in time of the reported emotions in more detail, the second-by-second model predictions and participants' responses for two of the test set samples were plotted in Figure 3b.

Comparison between music and speech models

In the previous sections, we have shown that it is possible to model continuous ratings of emotions perceived in music and speech stimuli using low-level psychoacoustic features and explain with a good level of precision a large proportion of the observed variance. Furthermore, both models performed very satisfactorily when predicting the emotion ratings for the unknown stimuli, which indicates the presence of underlying rules linking low-level psychoacoustic features and the communication of emotions in music and speech.

At this point, it seems natural to consider the similarities between the models. In the context of this research, we propose two methods to compare music and speech models. The first is to compare the composition of the input vectors, which reveals the relevant perceptual features used by the model to predict the subjective feelings of emotion. The second is to compare the performance figures, which can indicate how accurately these responses can be modelled for each domain. The following paragraphs detail both comparisons.

Input vector

Regarding the composition of the input vector, both models' input vectors share five features: L (loudness), T/SR (tempo/speech rate), mC/pC (melodic/prosodic contour), SC (spectral centroid), and S_a (sharpness). These features represent four psychoacoustic groups: loudness/intensity, duration/rate, pitch/frequency (mC and pC), and timbre/frequency. In addition to this, each model has one other distinct input feature: the music input vector includes SF (spectral flux) and the speech input includes R (roughness).

The most obvious implication of these results is that emotional cues are encoded as psychoacoustic spatio-temporal patterns in both mediums. Moreover, the subset of psychoacoustic cues that carry emotional "meaning" is very similar for both mediums. Such a result is consistent with the hypothesis that both music and speech prosody communicate emotion through structured modulations in the intensity, duration, and frequency components of sound. Perceptually, we suggest that those affective cues are perceived in both domains by means of spatio-temporal patterns in psychoacoustic percepts.

The extra input found for each model (R: speech; SF: music) can be interpreted as reflecting the relevance of roughness for conveying emotion in speech (and not in music), and the relevance of spectral flux for conveying emotion in music (but not in speech). However, there is at least one determinant to consider before accepting this interpretation of the results: the limitations of our experiment regarding the small sample of stimuli used. Indeed, the fact that the model learns how to compute the outputs from the inputs that describe the stimuli in the training set, binds the model's computational space to the information extracted from it. If, for instance, roughness is sufficiently similar across all training pieces, then the model may not be able to construe the output stream (and predict the emotions conveyed) by picking up information from that variable. This does not mean that the variable is irrelevant in general, but rather that it can be left out in the musical "universe" considered. The same can happen, for instance, when two variables (e.g., loudness and sharpness) are highly correlated (see also Schubert, 1999a).

In this context, one has to assume that the variables excluded during the input are not necessarily unimportant for emotional expression, but rather that the affective information they convey is redundant, insufficient or not representative of the actual relationships between the variable and the affective features. This could justify the fact that roughness, which is an important component in emotional responses to music from early on in life (Trainor & Heinmiller, 1998), does not form part of the music model.

In relation to the spectral flux input in the music model, it is likely that the possible changes in consecutive spectra of the voice are much more limited than those in music. Indeed, music has a greater range and goes through more pronounced frame-to-frame changes than speech (in point of fact spectral flux has been used as a feature to discriminate music and speech; see Scheirer & Slaney, 1997). It therefore seems reasonable that spectral flux is less informative in the case of speech signals.

Performance

In terms of explained variance in arousal and valence ratings, the music model outperforms the speech one for the training set stimuli (97% vs. 89% for arousal, and 83% vs. 70% for valence). The opposite happens with the test stimuli, such that more variance in the emotional perceived in speech stimuli was explained than in the music stimuli (75% vs. 81% for arousal, and 43% vs. 63% for valence). These figures suggest that the some of the variance in the training stimuli explained by the music model is the result of *overfitting* (that is, it uses rules extracted from the training set that apply to it but are not necessarily generalizable to other stimuli). Conversely, the larger amount of variance of the test set stimuli explained by the speech model, indicates better generalization of the speech model (over the music model) to novel stimuli. In summary, the most salient difference between the two models' performance (in terms of explained variance) is the fact that the speech model is better able to generalize the valence ratings to novel stimuli.

The *nrmse* (introduced earlier) allowed us to estimate the magnitude of the prediction errors normalized to the range of the observed values. Here, we revert to the *nrmse* figures presented earlier in Table 6 (music) and Table 7 (speech) to compare both models. Analysis of the average *nrmse* for the training set and test set indicates that the speech model predicts arousal more precisely than the music model (15% vs. 10% for the training set and 27% vs. 24% for the test set), and even more so for valence (21% vs. 8% for the training set and 30% vs. 22% for the test set). This means that the standard deviation of the error residuals normalized to the range of observed values was smaller for speech, which indicates a better fit.

Taking both measures together, we conclude that both models achieved similar performances, although there is a tendency for the speech model to predict more accurately a larger amount of the variance in valence conveyed from novel stimuli. This suggests that valence is more consistently predictable in speech signals.

Discussion and Conclusions

This investigation of emotional communication in music and speech combined a behavioural and computational study to identify a set of psychoacoustic cues associated with judgments of arousal and valence in the auditory domain. This research produced a model of psychoacoustic cues to emotion communication comprising a set of core features common to both music and speech, plus one extra feature unique to each. The five psychoacoustic features implicated in judgments of emotion in music and speech are loudness, tempo and speech rate, melodic and prosodic contour, spectral centroid, and sharpness. The features distinct to each domain are spectral flux, in the case of music, and roughness, in the case of speech.

These results indicate that emotional cues are encoded as psychoacoustic spatio-temporal patterns in both mediums. Furthermore, the psychoacoustic cues relevant to judgments of emotion perceived are very similar for both mediums, which is consistent with the hypothesis and previous findings that communication of emotion in the auditory domain arises from structured modulations in the intensity, duration, and frequency components of sound. Whereas previously pitch height, rate and intensity have been identified as likely candidates (Ilie & Thompson 2006, 2011), and a vast array of other attributes implicated in the communication of discrete emotions in music and speech (Juslin & Laukka, 2003), our study provides a parsimonious yet powerful model which pinpoints a feature set of 5 shared plus one additional unique acoustic attribute in each domain. A further strength of our findings is that they arise from ecological rather than artificial stimulus materials, and are based upon continuous ratings of emotion, thus capturing the dynamic character of emotional experience with music and speech. Our study provides supporting evidence for the idea that emotional content of music and speech is decoded, at least partially, by a shared processor that responds to psychoacoustic features regardless of the type of sound source (Juslin & Laukka, 2003; Ilie & Thompson, 2006, 2011), and specifies more precisely the particular acoustic features involved.

These insights are particularly noteworthy given that this is the first time, to our knowledge, that emotional communication by psychoacoustic cues in both speech and music have been modelled using the same mathematical framework and continuous measurements. This approach permitted a direct comparison of both mediums, as well as the investigation and modelling of mild and nuanced emotions (the most common states communicated in everyday life) as well as full-blown emotional states.

In this paper we do not elaborate on the relationships between specific acoustic cues (or groups of cues) and emotion qualities. This is because analytical tools which would allow us to infer the rules embedded in the model currently do not exist; this kind of modelling requires apriori knowledge about the inputs and/or output which is not available in our case. It is possible to describe the model as a set of equations, but that would not be very helpful in terms of informing us about the links between music, speech prosody, and emotion. In fact, to find these straightforward links (typically reported as the relationship between extreme levels of a specific cue and some sort of emotion label) would be counterproductive since the very premise of our research is that these relationships are nonlinear (related to the existence of redundant information amongst features and the temporal dimension). We are currently developing analytical methods that permit analysis of the network and the extraction of relevant information, but this is a separate project and beyond the scope of this paper. These techniques include reducing the dimensionality of the network's hidden layer activations (using dimensionality reduction techniques such as as PCA or ICA) and evaluating the canonical correlations between this reduced set of variables, the inputs, and the outputs. This approach highlights linear relationships between specific psychoacoustic cues, and arousal and valence but ignores the temporal component. To consider the temporal dimension, we are testing the use of moving correlations between inputs, hidden activations, and outputs, to detect specific moments of causal linkage between psychoacoustic cues and emotion dimensions. The development of these analytical tools is therefore dealt with elsewhere.

Despite these limitations in terms of model interpretability, it should be remembered that the model itself is the representation of the relevant emotional features obtained from acoustic features and used to predict emotional responses. With this in mind our results have several important implications beyond identifying specific psychoacoustic cues to emotion communication in music and speech.

The difference between the psychoacoustic cues implicated in the model for music and for speech is instructive because it suggests that judgments of emotion in music and speech will call on different psychoacoustic dimensions according to their relevance in the individual context. Brunswik's (1956) lens model, as modified by Scherer (1982) for nonverbal communication, and Juslin for music performance (1997), provides a helpful framework within which to understand this phenomenon. According to this perspective, when one cue is unavailable in a performance another cue can be used to convey a similar affect, and so listeners also adopt a flexible approach to their decoding. Ilie & Thompson (2006) argue that differences in the influence of affective cues in music and speech may be because the listener is using different attentional strategies, allocating more or less attention to particular cues depending on which source they are listening to, and that, because of this, stimulus properties are not good predictors of specific emotions. However, one of the significant advantages of our model is that it seems able to model these different attentional strategies. Furthermore, the good performance of our models suggests that a more stimulus-driven (or ecologically mutualistic) approach can be a successful predictor of perceived emotion. This same principle underlies the pertinence of different psychoacoustic cues in different genres of music because different repertoires are associated with different acoustic characteristics: for example, it is difficult to vary loudness on a harpsichord and therefore loudness is unlikely to operate as a cue to emotion in music performed on a harpsichord. Evidence supporting this observation comes from preliminary tests with multiple genres which suggest that the model performs less well and with poorer generalization when applied to multiple genres: when dance, rock, pop and death metal were used as a training set and instrumental and vocal classical, and film music were used as the test set a significant proportion of listeners responses was predicted from the psychoacoustic features but with lower performance than in this case where a single genre is used for both training and test sets (Coutinho, 2010).

As this suggests, psychoacoustic features are not orthogonal dimensions and hence they share information about the stimulus that can be seen as redundant. One way of understanding this overlap of available information is to see it as providing a safeguard, which keeps crucial auditory information alive even in the event of an impairment to a particular cognitive function. From an evolutionary perspective this makes sense because it means that the organism is able to utilize what would otherwise be redundancy in the acoustic cues to emotion to avoid individual and social disadvantage (e.g. the ability to perceive emotional meaning in sound despite impairment of timbre perception).

This distinction between shared and domain-specific attributes of music and speech is highlighted by Ilie & Thompson (2011: 260-261) in their overview of emotional communication in the auditory domain. They speculate that cues which are shared across auditory domains are likely to resist enculturation in comparison to domain specific cues, and that enculturation of these domain specific cues may lead to fractionation of emotional communication systems, thus accounting for cross-cultural differences in emotional coding. For example, they argue that the association of higher pitch with motherese during child development may lead to high pitch in speech being associated with greater pleasantness than high pitch in music. Thus, while we argue that our results provide evidence of the importance of acoustic cues to emotional responses to music we recognize that cues are also subject to historical, social and technological contingencies.

Our results support Ilie and Thompson's (2006) observation that music stimuli tend to elicit stronger affective responses than speech stimuli. The maximum arousal and valence scores (averaged across all participants) across all music stimuli was almost double that of the speech samples (music: .53/.31, compared to speech: .28/.14), whereas the minimums were almost identical. This difference in the strength of emotional response may reflect differences in the emotional "work" performed by these two different domains in everyday life: as Ilie and Thompson (2006) remark, people more commonly listen to music for reasons of mood regulation and pleasure than they do to speech alone. In addition to this, we suggest that the larger range of sounds qualities as well as their organization and production in instrumental and electroacoustic music, permits communication of more intense states.

A possible criticism of the approach adopted here is that listeners may report changes in musical energy and characteristics rather than perceived emotions. In other words, participants may be subject to demand characteristics within the experimental context which encouraged them to model "emotional response to music" using the only information available to them – psychophysical cues. In counter-argument to this claim are two types of evidence. First, other studies have adopted this behavioural approach and provide physiological data consistent with the induction of perceived (Coutinho & Cangelosi, 2009) and felt (Coutinho & Cangelosi, 2011) emotion in listeners. Second, listener responses are internally consistent and show a high degree of agreement with each other, which indicates that acoustic cues act as indices of emotion even if this is seen as a measure of how music communicates emotion rather than emotion which is felt.

A further question concerns the status of the averaged participant responses on which the model is based. The model predicts listener arousal and valence based on an average of listener responses; in other words we are collecting shared emotional evaluations. Application of Cronbach's Alpha indicates that this shared evaluation has good internal consistency. Investigation of individual differences lies beyond the scope of this paper and are the subject of ongoing research.

In addition, the number of stimuli may seem relatively small when compared to the possible combinations across the 2-DES space. In fact, the 17 stimuli represent around 15 minutes of music and 15 minutes of speech, which, having used a 1Hz sample rate, equates to 900 input stimuli to the model. Nontheless, and it was not possible to have an exact match of the number of times each quadrant occurred in the final set. This raises the possibility that there may be over- or under-fitting of some areas of the emotion space limiting the generalisability of the model and highlights the need for a set of stimuli that is able to represent the acoustic diversity of a wide range of music styles and emotional states. Such diversity can only be achieved through multiple experiments (due to experimental restrictions) and the strength of this study is that by using continuous data we have provided a more complete picture than has hitherto been possible.

Lastly, criticisms of the early-stop back propagation algorithm used here are that it is very much dependant on the training and test sets chosen, which in turn may lead to an overestimation of the model performance. Indeed, due to the fact that the optimal model is chosen based on the analysis of the error of the test set, it may lead to an optimization of the model performance to that particular set. As a result the model may over estimate the extent to which it generalises to other data sets, that is, the amount of variance explained by the model may be smaller than what our results show. One approach to avoid this in future research would be to evaluate the generalisation using a third ‘validation’ set, not used as a criteria to select the best model, but that can be used on it to obtain an unbiased estimate for the predicted error. It should be noticed nevertheless that comparable performances were obtained in previous modeling work using two different data sets (Coutinho & Cangelosi, 2009, 2011)

These conclusions highlight specific topics for future research. First, it is desirable to produce a model which reflects a fuller range of musical stimuli (including various genres and cultures) and therefore of psychophysical cues to emotion, in order to explore how much of the variance in responses can be explained by psychoacoustic cues. Second, the generalizability of the model needs to be determined through cross-cultural studies with listeners from different languages and musical cultures. Third, it should also be possible to determine which psychophysical cues are important within particular musical styles. In addition to the study of film music reported here, the model has been applied successfully to Western classical music (Coutinho & Cangelosi, 2011), romantic music (Coutinho & Cangelosi, 2009), and popular music (Coutinho 2010). Replication of the study across further repertoires will provide insight onto the variability of acoustic cues to emotion across genres.

This project has a number of implications and applications. This research provides a new method for the analysis of emotional communicated by speech and music stimuli, and reveals the existence of acoustic and perceptual schema underlying the perception of emotion in both domains. We identify particular acoustic characteristics that appear to be responsible for the perception of emotion in music and speech prosody, furthering knowledge of emotion communication in the auditory domain. This study, for the first time, captures the dynamic aspect of emotional experience with auditory phenomena, by using continuous measurement. A final innovation is the identification of group differences in continuous report of emotion perception across auditory domains, namely gender and emotional intelligence, emotional stability, and musical training. We show that these factors play a role in continuous evaluations of emotion perception with the potential to be integrated into future models. It is worth noting that it is not only the global model performance which is important, but also the errors at each moment in the stimuli. It is plausible to assume that higher error in specific section/portions of music or speech stimuli are at least partially due to the lack of information from the current model inputs and/or inherent limitations of the limited set of stimuli used. In the case of the first, it could indicate that other low-level features or, very likely, "higher-level" features (either stimulus related or individual) would be necessary to describe the emotions perceived by subjects.

The work also has applications to practice. For example, the model provides a means to analyse the potential effects of particular pieces and aid the systematic selection of music for applications in settings such as health and wellbeing. The model has potential to be used as an instrument for the diagnosis of different psychiatric conditions that can be inferred from speech, such as depression and schizophrenia (e.g., Tolkmitt, Helfrich, Standke & Scherer, 1982). The model could also be used to improve hearing aid systems in their response to relevant acoustic features in emotional communication, by adjusting their electronic and electro-acoustic parameters to optimize speech parameters. Areas involved in human-computer interactions may also benefit, since our model could be integrated into emotion recognition systems and used in the synthesis of emotional speech.

Beyond the specific applications suggested above, this work provides new evidence that speech and music are processed by general-purpose brain mechanisms which are responsive to acoustic features regardless of their modality (e.g., Patel, 2010). In the longer-term such evidence can contribute to an understanding of the evolutionary origins of language and music, as part of a converging picture of commonalities between speech and music.

References

Aures, W. (1985). Ein Berechnungsverfahren der Rauhigkeit. *Acustica*, 58, 268-281.

Bachorowsky, J.-A. & Owren, M. J. (2001). Not all laughs are alike: voiced but not unvoiced laughter elicits positive affect. *Psychological Science*, 12, 252-257.

Bachorowsky, J.-A. & Owren, M. J. (2003). Sounds of Emotion: Production and Perception of Affect-Related Vocal Acoustics. *Annals of the New York Academy of Sciences*, 1000, 244-265.

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

Blood, A. & 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 of the United States of America*, 98 (20), 11818-11823.

Bradley, M. & Lang, P. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25 (1), 49–59.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Chalupper, J. & Fastl, H. (2002). Dynamic Loudness Model (DLM) for Normal and Hearing-Impaired Listeners. *Acta Acustica united with Acustica*, 88 (3), 378-386.

Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Zhu, H., Kangarlu, A., Duan, Y., Russell, J., & Peterson, B. (2010). Neural systems subserving valence and arousal during the experience of induced emotions. Emotion, 10 (3), 377-389.

Coutinho, E. (2010). Modeling psycho-physiological measurements of emotional responses to multiple music genres. In S. M. Demorest, S. J. Morrison, & P. S. Campbell (Eds.), *Proceedings of the 11th International Conference of Music Perception and Cognition (ICMPC11)* (p. 53). Seattle, WA, USA.

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

Coutinho, E. & Cangelosi, A. (2010). A Neural Network Model for the Prediction of Musical Emotions. In S. Nefti-Meziani & J.G. Grey (Eds.). *Advances in Cognitive Systems* (pp. 331-368). London: IET Publisher. ISBN: 978-1849190756.

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11(4). Advance online publication. Retrieved from http://psycnet.apa.org/journals/emo/11/4/

Cowie, R. & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40, 5–32.

Daniel, P. & Weber, R. (1997). Psychoacoustical roughness: implementation of an optimized model. *Acustica*, 83, 113-123.

De Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41 (2), 385-390.

Dittmar, C., Dressler, K., & Rosenbauer, K. (2007). A toolbox for automatic transcription of polyphonic music. In *Proceedings of audio mostly: 2nd conference on interaction with sound* (pp. 58-65). Ilmenau, Germany: IMDT.

Dixon, S. (2006). MIREX 2006 audio beat tracking evaluation: BeatRoot. http://www.music-ir.org/evaluation/MIREX/2006 abstracts/BT dixon.pdf.

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.

Ekman, P. (1992). Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science,* 3 (1), 34-38.

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

Frick, R.W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97 (3), 412-429.

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

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

Gabrielsson, A., & Juslin, P.N. (2003). Emotional expression in music. In R.J. Davidson, H.H. Goldsmith & K.R. Scherer (Eds.), *Handbook of Affective Sciences* (pp. 503-534). New York: Oxford University Press.

Gabrielsson, A. & Lindström, E. (2010). The role of structure in the musical expression. In P. N. Juslin & J. A. Sloboda, J. (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 367-400). Oxford, UK: Oxford University Press.

Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2007b). Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, 7 (4), 774-788.

Hutchinson, W. & Knopoff, L. (1978). The acoustic component of Western consonance. *Journal of New Music Research*, 7 (1), 1-29.

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.

Ilie, G. & Thompson, W. F. (2011). Experiential and cognitive changes following seven minutes exposure to music and speech. *Music Perception,* 28 (3), 247-264.

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

Juslin, P. N. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin*, 129 (5), 770–814.

Juslin, P. N. & Sloboda, J. A. (Eds.). (2010). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford, UK: Oxford University Press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). International affective picture system (IAPS): Technical manual and affective ratings. Gainesville, FL, USA: University of Florida, The Center for Research in Psychophysiology.

Larsen, J. T. & McGraw, A. P. (2011). Further evidence for mixed emotions. *Journal of Personality and Social Psychology,* 100 (6), 1095-1110.

Mertens, P. (2004). Le prosogramme: une transcription semi-automatique de la prosodie. *Cahiers de l'Institut de Linguistique de Louvain*, 30 (1-3), 7-25.

Owren, M. J. & Rendall, D. (1997). An affect-conditioning model of nonhuman primate signaling. In D. H. Owings, M. D. Beecher, & N. S. Thompson (Eds.) *Perspectives in Ethology* (*Volume 12): Communication* (pp. 299-346). New York, NY, USA: Plenum Press.

Patel, A. D. (2010). *Music, language, and the brain*. New York, NY, USA: Oxford University Press.

Patel, A.D., Peretz, I, Tramo, M. & Labreque, R. (1998). Processing Prosodic and Musical Patterns: A Neuropsychological Investigation. *Brain and Language*, 144 (61), 123-144.

Peretz, I. (2001). Brain specialization for music: new evidence from congenital amusia. *Annals of the New York Academy of Sciences*, 930 (1), 153.

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

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

Schachner, A. & Hannon, E. E. (2011). Infant-directed speech drives social preferences in 5-month-old infants. *Developmental Psychology*, 47 (1), 19-25.

Scheirer, E. & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In B. Werner (Ed.), *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)* (vol. 2, pp. 1331-1334). Los Alamitos, CA, USA: IEEE Computer Society Press. DOI: 10.1109/ICASSP.1997.596192.

Scherer, K. R. (1982). Emotion as a Process: Function, Origin, and Regulation, *Social Science Information,* 21, 555–570.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin* , 99, 143-165.

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

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 (3), 154-165.

Sethares, W. A. (1999). *Tuning, timbre, spectrum, scale*. Springer-Verlag: London, UK.

Thompson, W. F. & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica*, 158, 407-424.

Tolkmitt, F., Helfrich, H., Standke, R. & Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of Communication Disorders*, 15 (3), 209-222.

Trainor, L. J., & Heinmiller, B. M. (1998). The development of evaluative responses to music: Infants prefer to listen to consonance over dissonance. *Infant Behavior and Development*, 21, 77–88.

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance*. Unpublished doctoral dissertation. University of California Los Angeles.

Webster, D. and Weir, C. G. (2005). Emotional Responses to Music: Interactive Effects of Mode, Texture, and Tempo. *Motivation and Emotion*, 29, 1, 19-39.

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

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: facts and models*. New York, NY, USA: Springer-Verlag.

Discography

Bibel TV das Gespräch: Eine Stimme, die berührt, Njeri Weth (Interview with Njeri Weth at Bibel.tv) http://www.youtube.com/watch?v=8SaoEUf5rCw Accessed 9 May 2011.

Cinema Century: A Musical Celebration of 100 Years of Cinema. Silva Screen, 1999. (Audio CD)

Das Leben Der Anderen (The Lives of Others) (Dir. Florian Henckel von Donnersmarck 2006). Albert Jerska (actor Volkmar Kleinert) speaking to Georg Dreyman (actor Sebastian Koch).

"Die Lehre der vier Temperamente: Her (sic) Cholerix" (The doctrine of the four temperaments: Mr. Cholerix). Andreas Konzack. http://www.youtube.com/watch?v=R90rCdplA-M Accessed 9 May 2011.

"Die Lehre der vier Temperamente: Herr Sanguinix" (The doctrine of the four temperaments: Mr. Sanguinix). Andreas Konzack. http://www.youtube.com/watch?v=iXid2kdchzE Accessed 9 May 2011.

"Ich Bin Ein Wunder" (I Am A Miracle) Interview with Jenny Spritzer (Auschwitz survivor).

"Mad as hell" speech from the film *Network* (Dir. Sidney Lumet 1976). Dubbed into German. http://www.youtube.com/watch?v=nPbash9jKH8 Accessed 9 May 2011.

"Orphische Bucht" (Orphic Bay). Poem by Erich Arendt (recited by an unknown female). http://www.gedichte-finden.de/gedichte/Orphische-Bucht--7047.html. Accessed 26 July 2011.

Sci-Fi At the Movies. 5th Continent, 1997. (Audio CD)

Simply Soundtracks. Simply, 2009. (Audio CD)

Simply Film Themes. Simply, 1999. (Audio CD)

“Three things that make me happy". Interview with Edda Raspé (summer of 2004). http://video.google.de/videoplay?docid=-6311105687124251 40&ei=Lh5PS6TDBMeP-AacjpT7CQ&q=gl%C3%BCcklicher+Interview&hl=de# Accessed 9 May 2011.

"Woman secrets" Interview with Charlotte Roche by *Stern* magazine. http://www.youtube.com/watch?v=4s8ZiQBrNiQ Accessed 9 May 2011.

Table 1

Pieces of music used in the empirical study. The extracts are numbered consecutively, so as to serve as aliases for reference in this article. For each extract we give the piece title, its duration, and the 2DES quadrant corresponding to the emotional response we expect the extract to elicit in listeners based on pre-testing.

ID	*Piece*	Duration	Expected quadrant
1	*(Bram Stoker's) Dracula,* Vampire Hunters	2:00	Q2
2	*Bride of Frankenstein,* Main Title	1:24	Q1 Q2 Q3 Q4
3	*Guns for San Sebastian*	1:54	Q1 Q3
4	*Hellraiser*, Main Title	1:50	Q2 Q3
5	*Krull*, Love Theme	2:10	Q1 Q4
6	*Minority Report*, Main Theme	1:48	Q1 Q4
7	*The Quiet Earth*, Finale	1:48	Q2 Q3
8	*The Searchers*, Suite	1:29	Q1 Q2 Q4

Table 2

Speech samples used in the experiment. The excerpts are numbered consecutively, so as to serve as aliases for reference in this article. For each excerpt we indicate the source, its duration, and the 2DES quadrant corresponding to the emotional response we expect it to elicit in listeners, based on pre-testing.

ID	Sample	Duration	Expected quadrant
1	Sketch: "The doctrine of the four temperaments: Mr. Sanguinix".	0:45	Q1
2	Interview: Charlotte Roche "Woman secrets".	2:36	Q1 Q4
3	Speech: Howard Beale (actor Peter Finch) delivering his "mad as hell" speech from the film *Network*.	1:39	Q2
4	Sketch: "The doctrine of the four temperaments: Mr. Cholerix".	0:58	Q2
5	Interview: Jenny Spritzer "Ich Bin Ein Wunder/I Am A Miracle".	1:16	Q2 Q3
6	Poetry: “Orphische Bucht” (Orphic Bay). Poem by Erich Arendt (recited by an unknown female).	1:40	Q3 Q4
7	Speech: Albert Jerska (actor Volkmar Kleinert) speaking to Georg Dreyman (actor Sebastian Koch) in the film *The Lives of Others/Das Leben Der Anderen*.	1:03	Q3
8	Interview: Njeri Weth "A voice that touches".	1:28	Q1 Q4
9	Interview: Edda Raspé “Three things that make me happy”.	1:46	Q3 Q4

Table 3

Psychoacoustic variables considered in this study. The time series obtained were down-sampled from the original sample rates (which vary from feature to feature) to 1Hz in order to obtain second by second values.

Psychoacoustic Group	Feature	ID	Alias
Loudness	Dynamic Loudness	i1	L
Duration	Tempo (only music stimuli)	i2m	T
	Speech rate (only speech stimuli)	i2s	SR
Pitch	Melody contour (only music stimuli)	i3m	mC
	Prosody contour (only speech stimuli)	i3s	pC
	Spectral Flux	i4	SF
Timbre	Sharpness (Zwicker and Fastl)	i5	Szf
	Sharpness (Aures)	i6	Sa
	Power Spectrum Centroid	i7	SC
Roughness	Psychoacoutical Roughness	i8	R
	Auditory dissonance (Hutchinson and Knopoff)	i9	Dhk
	Auditory dissonance (Sethares)	i10	Ds

Table 4

Error statistics for the simulations: Music.

Simulation	Inputs	*rmse*			
		Arousal		Valence	
		Training	Test	Training	Test
1	L, T, mC	.05	.08	.04	.10
$2^{\wedge}$	L, T, mC + SC	**.05**	.09	.05	**.09**
$3^{\wedge}$	L, T, mC + SF	.07	**.08**	**.03**	**.09**
4	L, T, mC + S_{zf}	**.04**	.08	.08	.10
$5^{\wedge}$	L, T, mC + S_{a}	.07	.09	.06	**.09**
$6^{\wedge}$	L, T, mC + SD_{hk}	.06	.09	.08	**.09**
7	L, T, mC + SD_{s}	**.05**	.09	**.03**	**.10**
8	L, T, mC + R	.07	.09	**.04**	**.10**
9^{+}	L, T, mC + SC, SF	**.05**	**.08**	.06	**.09**
10^{+}	L, T, mC + SC, S_{a}	**.05**	**.09**	.06	**.09**
11	L, T, mC + SC, SD_{hk}	.06	**.09**	.06	.10
12^{+}	L, T, mC + SF, S_{a}	**.07**	.09	.05	**.09**
13	L, T, mC + SF, SD_{hk}	.08	.09	.07	.10
14	L, T, mC + S_{a}, SD_{hk}	**.04**	**.07**	**.05**	.10
15^{*}	L, T, mC + SC, SF, S_{a}	**.05**	**.08**	.06	**.08**
16	L, T, mC + SC, SF, S_{a}, SD_{hk}	**.04**	.09	.07	**.09**

Table 5

Error statistics for the simulations: Speech.

Simulation	Inputs	*rmse*			
		Arousal		Valence	
		Training	Test	Training	Test
1	L, SR, pC	.02	.05	.02	.10
$2^{\wedge}$	L, SR, pC + SC	.03	.06	**.02**	**.08**
3	L, SR, pC + SF	**.02**	**.05**	**.02**	.11
4	L, SR, pC + S_{zf}	**.02**	.06	**.02**	**.07**
$5^{\wedge}$	L, SR, pC + S_{a}	**.02**	.07	**.02**	**.06**
6	L, SR, pC + SD_{hk}	.03	.06	**.02**	.13
$7^{\wedge}$	L, SR, pC + SD_{s}	**.02**	**.04**	**.02**	.11
$8^{\wedge}$	L, SR, pC + R	.03	**.05**	.05	**.08**
9^{+}	L, SR, pC + SC, S_{a}	.03	**.06**	**.02**	.07
10	L, SR, pC + SC, SD_{s}	**.02**	**.04**	**.02**	.11
11^{+}	L, SR, pC + SC, R	.04	**.05**	.05	**.08**
12	L, SR, pC + S_{a}, SD_{s}	**.02**	.05	**.02**	.09
13^{+}	L, SR, pC + S_{a}, R	**.02**	.06	**.01**	**.06**
14^{+}	L, SR, pC + SD_{s}, R	**.02**	**.04**	**.02**	.12
15*	L, SR, pC + SC, S_{a}, R	.03	**.04**	.02	**.06**

Table 6

Precision (rmse and nrmse), similarity (r), and explained variance (r^2) values for all music pieces used in the experiment. The statistics shown pertain to the model used in Simulation #15 which includes loudness (L), tempo (T), melodic contour (mC), spectral centroid (SC), spectral flux (SF) and sharpness (S_a) as inputs. The nrmse corresponds to the rmse normalized to the range of observed values (participants responses to target values) for each emotion dimension.

Piece	Arousal				Valence			
	r	r^2 (%)	rmse	nrmse (%)	r	r^2 (%)	rmse	nrmse (%)
1	.93	86	.03	9	.91	82	.04	19
2	.73	54	.10	39	.24*	6	.09	32
3	.98	96	.05	13	.82	68	.06	25
4	.98	96	.07	24	.95	90	.05	17
5	.99	98	.05	12	.93	86	.08	26
6	.85	72	.09	45	.87	75	.07	28
7	.99	98	.04	10	.95	90	.04	15
8	.94	89	.07	17	.32*	10	.12	40
av. Train		97	.05	15		83	.06	21
av. Test		75	.07	27		43	.08	30
total av.		86	.06	21		63	.07	25

Note. * p < .05; p = 0 for all others

Table 7

Precision (rmse and nrmse), similarity (r), and explained variance (r^2) values for all speech samples used in the experiment. The statistics shown pertain to the model used in Simulation #15 which includes loudness (L), speech rate (SR), prosodic contour (pC), spectral centroid (SC), sharpness (S_a) and roughness (R) as inputs. The nrmse corresponds to the rmse normalized to the range of observed values (participants responses to target values) for each emotion dimension.

Sample	Arousal				Valence			
	r	r^2 (%)	rmse	nrmse (%)	r	r^2 (%)	rmse	nrmse (%)
1	.98	97	.02	6	.98	96	.01	2
2	.92	85	.04	21	.89	80	.08	19
3	.99	97	.05	12	.99	99	.03	4
4	.94	88	.04	13	.97	94	.04	7
5	.83	68	.06	42	.86	74	.07	18
6	.98	95	.01	5	.25*	6	.01	21
7	.99	98	.01	6	.87	75	.01	7
8	.77	59	.01	20	.87	76	.02	7
9	.91	82	.03	19	.22*	5	.05	45
av. Train		89	.02	10		70	.02	8
av. Test		81	.04	24		63	.06	22
total av.		86	.03	16		67	.04	15

Note. * p < .05; p = 0 for all others

Figure captions

Figure 1: Model architecture and information flow showing the three-layered recurrent neural network, which takes encoded psychoacoustic features as input and outputs predicted valence and arousal.

Figure 2: Plot showing the second-by-second values of the self-reported emotional arousal and valence averaged across all participants for each music piece (a) and speech sample (b) against the model predictions. Each pair of values is represented by their corresponding location in the 2DES.

Figure 3: Model predictions for music pieces 1 and 6 (a) and speech samples 4 and 9 (b). The arousal and valence time series predicted by the model for the stimuli are shown together with the averaged 2DES rating from participants.

Figure 1

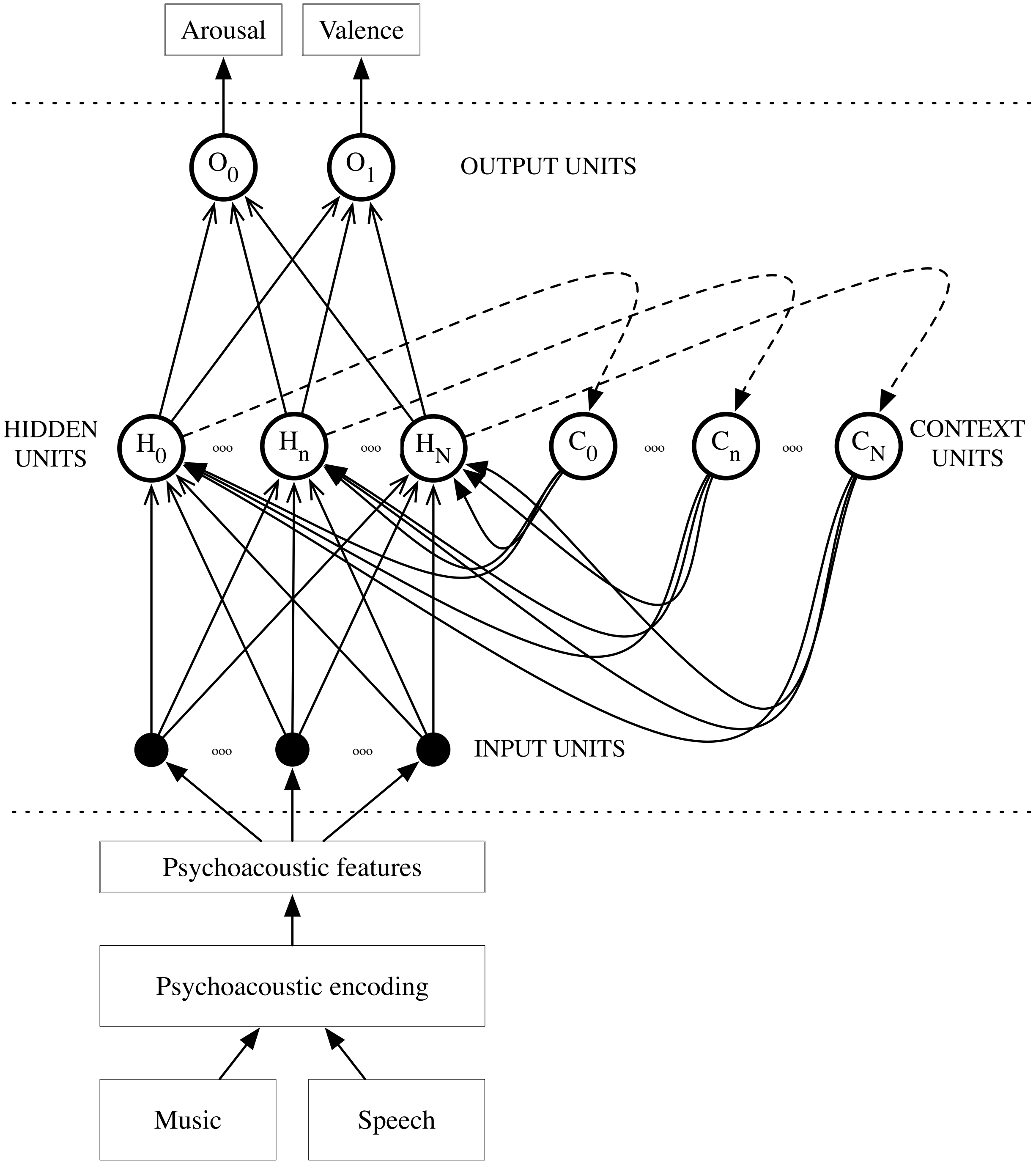
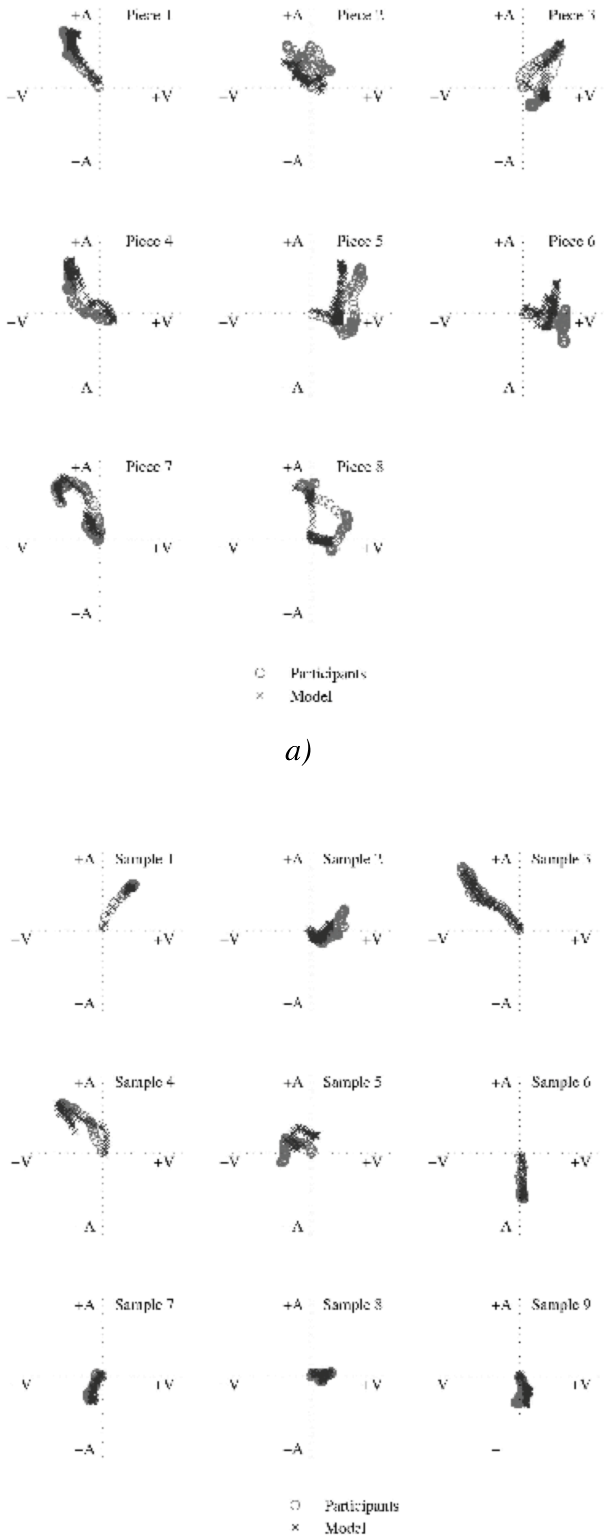

Arousal
Valence
O_0
O_1
OUTPUT UNITS
HIDDEN UNITS
H_0
H_n
H_N
C_0
C_n
C_N
CONTEXT UNITS
INPUT UNITS
Psychoacoustic features
Psychoacoustic encoding
Music
Speech

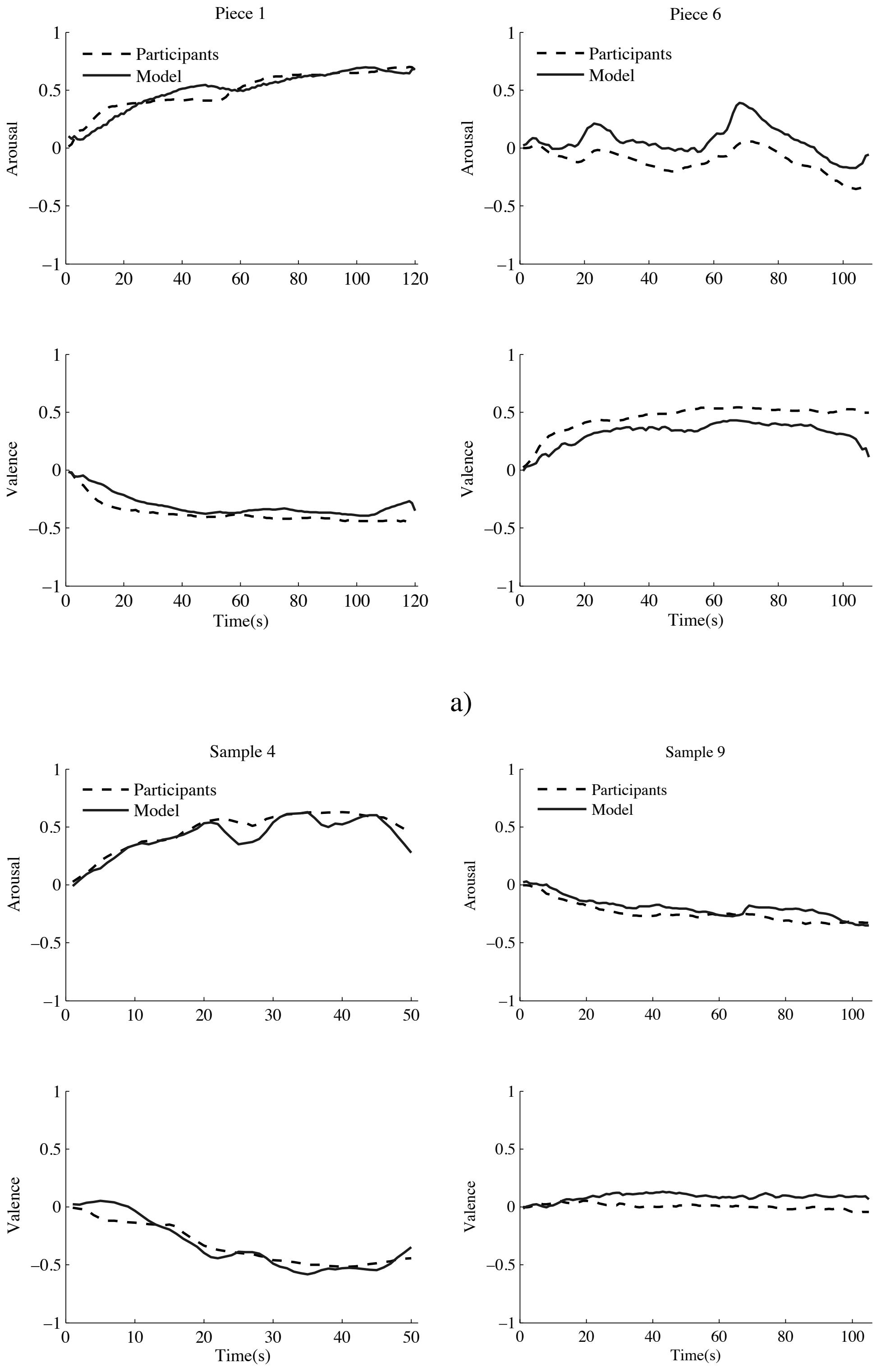
Figure 2


Piece 1
Piece 2
Piece 3
Piece 4
Piece 5
Piece 6
Piece 7
Piece 8
+A
−A
−V
+V
Participants
Model
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Participants
Model

a)

b)

Figure 3


Piece 1
Piece 6
Participants
Model
Arousal
Valence
Time(s)
1
0.5
0
−0.5
−1
0
20
40
60
80
100
120
Sample 4
Sample 9
10
30
50

a)

b)

Author Notes

This research was supported by a Postdoctoral Fellowship from the Portuguese Foundation for Science and Technology (SFRH/BPD/62850/2009) and a Small Research Grant from the Calouste Gulbenkian Foundation to the first author.

Notes

[i] A description of the model can be found in Coutinho (2010). Please refer to Elman (1990) for a detailed functional and mathematical description of the ENN.

[ii] Calculated as the Euclidean distance between the two normalised spectra in consecutive spectral frames.

[iii] For instance, to evaluate simulation #9 (L, T, mP + SC and L, T, mP + SF), the error reference for the training set arousal ratings was .05 – the minimum training error for arousal ratings between simulations #2 (L, T, mP + SC) and #3 (L, T, mP + SF).