Exploring Cognitivist and Emotivist Positions of Musical Emotion Using Neural Network Models

Naresh N. Vempala (nvempala@psych.ryerson.ca)
Department of Psychology, Ryerson University, 350 Victoria Street
Toronto, ON, M5B 2K3, Canada

Frank A. Russo (russo@psych.ryerson.ca)
Department of Psychology, Ryerson University, 350 Victoria Street
Toronto, ON, M5B 2K3, Canada

Abstract

There are two positions in the classic debate regarding musical emotion: the cognitivist position and the emotivist position. According to the cognitivist position, music expresses emotion but does not induce it in listeners. So, listeners may recognize emotion in music without feeling it, unlike real, everyday emotion. According to the emotivist position, listeners not only recognize emotion but also feel it. This is supported by their physiological responses during music listening, which are similar to responses occurring with real emotion. When listeners provide emotion appraisals, if the cognitivist position were true, then these appraisals might be based on audio features in the music. However, if the emotivist position were true, then appraisals would be based on the emotion experienced by listeners as opposed to what they perceived in the audio features. We propose a hypothesis combining both positions according to which, listeners make emotion appraisals based on a combination of what they perceive in the music as well as what they experience during the listening process. In this paper, we explore all three positions using connectionist prediction models, specifically four different neural networks: (a) using only audio features as input, (b) using only physiological features as input, (c) using both audio and physiological features as input, and (d) using a committee machine that combines contributions from an audio network and a physiology network. We examine the performance of these networks and discuss their implications as possible cognitive models of emotion appraisal within listeners.

Keywords: musical emotion; connectionist models; neural networks.

Introduction

There are two positions in the classic debate regarding musical emotions: the *cognitivist* position and the *emotivist* position. According to the cognitivist position, music can express emotion but it does not actually induce it in the listener (Konečni, 2007; Meyer, 1956). So, the listener may recognize the emotion without feeling it as they would in the case of everyday emotion. However, other studies (Gabrielsson, 2002; Juslin & Västjfäll, 2008) have shown that listeners, in most cases, not only recognize emotions within music but also undergo physiological changes during music listening. These physiological changes are characteristic of changes that occur with real, everyday emotion. This has been used as evidence by the emotivist position that music induces emotion that is quite similar to everyday emotion, and in effect, genuine.

When a listener is asked to provide an appraisal of emotion experienced while listening to a music excerpt, he or she has to make an assessment at a cognitive level. If the cognitivist position is true, then the emotional appraisal can be based on an assessment of audio features (deep and/or surface-level). However, if the emotivist position is true, then the listener makes his or her appraisal on the basis of the emotion he or she experiences. We propose a third position that involves a combination of the cognitivist and emotivist positions. Our hypothesis is that appraisals of musical emotion are made by listeners based on a combination of what they perceive in the music and what they experience during the listening process.

In this paper, we explore all three positions using connectionist models, specifically feedforward neural networks (multilayer perceptrons). If the cognitivist position were true, then audio features alone would be sufficient to capture the emotion appraisals made by human listeners. Likewise, if the emotivist position were true, then physiological features alone would be sufficient to capture the emotion appraisals made by human listeners. On the other hand, if emotions were appraised by a listener based on a combination of what the listener perceives as well as what he or she feels, the ideal model would involve a combination of audio and physiological features.

We used four neural networks as models of musical emotion. The first network takes only audio features as network inputs. The second network takes only physiological responses of participants as network inputs. The third and fourth network combine audio and physiological features in different ways. The third network uses a combination of audio and physiological features as inputs whereas the fourth network is a meta-level network (i.e. a committee machine), which takes the outputs from two expert networks and combines them to form appraisals of musical emotion. We examine the performance of these networks and discuss their implications as possible cognitive models of emotion appraisal within listeners.

Discrete and dimensional methods have been used to capture emotion assessments of listeners during music listening. While discrete methods label emotions into specific categories such as *happy* and *angry*, dimensional

Table 1: 12 music excerpts with composers, emotion quadrants, mean valence/arousal ratings, and standard deviations.

				Valence		Arousal	
Excerpt	Composer	Composition	Quadrant	Mean	SD	Mean	SD
M1	Bartok	Sonata for 2 pianos and percussion (Assai lento)	Agitated	5	2.13	6.35	2.21
M2	Shostakovich	Symphony No. 8 (Adagio)	Agitated	3.35	1.84	7.45	1.43
M3	Stravinsky	Danse sacrale (Le Sacre duPrintemps)	Agitated	3.95	1.99	7.15	1.95
M4	Beethoven	Symphony No. 7 (Vivace)	Happy	6.6	1.6	6.35	1.57
M5	Liszt	Les Preludes	Happy	5.75	1.52	6.25	1.48
M6	Strauss	Unter Donner und Blitz	Happy	6.8	1.94	7.5	1.19
M7	Bizet	Intermezzo (Carmen Suite)	Peaceful	6.6	1.6	2.85	1.93
M8	Dvorak	Symphony No. 9 (Largo)	Peaceful	5.95	1.85	2.65	1.66
M9	Schumann	Traumerei	Peaceful	5.75	1.86	2.8	1.91
M10	Chopin	Funeral March, Op. 72 No. 2	Sad	4.85	1.87	2.55	1.54
M11	Grieg	Aase's Death (Peer Gynt)	Sad	4.05	1.93	4.15	2.25
M12	Mozart	Requiem (Lacrimosa)	Sad	4.3	1.78	3.75	2.45

models allow emotions to be characterized on an n-dimensional space. We used a 2-dimensional space, comprised of valence and arousal, to capture appraisals of musical emotion (Russell, 1980). Valence refers to the hedonistic aspect of felt emotion, ranging from pleasant to unpleasant. Arousal refers to the activation aspect of felt emotion, ranging from calm to excited. Using this method of capturing felt emotion provided us with two advantages: (a) the ability to compute the emotional distance between musical excerpts, and (b) the ability to avoid categorizing an excerpt by a specific emotion label in cases where the emotion is ambiguous.

Stimuli and Data Collection

Our stimuli consisted of 12 classical music excerpts from 12 different composers. These excerpts were chosen based on previous work investigating emotional responses to music (Bigand et al., 2005; Nyklicek et al., 1997; Sandstrom & Russo, 2010). As shown in Table 1, they were chosen such that three excerpts represented each of the four emotion quadrants in Russell's circumplex: high arousal, positive valence (*Happy*); high arousal, negative valence (*Agitated*); low arousal, negative valence (*Sad*); and low arousal, positive valence (*Peaceful*).

We used data from 20 participants (17 females, 1 male, 2 undeclared). On average, the participants were 25 years of age (SD_{age} = 9.2) and had limited music training: 1.7 years of individual training (SD = 2.9) and 2 years of group training (SD = 2.8). Participants heard the 12 music excerpts in a single session. Each excerpt was preceded by 30 seconds of white noise and followed by 50 seconds of silence. Participants were randomly assigned to any one of four randomized orders of the excerpts.

Physiological responses were collected from five different channels during music listening using the Biopac MP100 data acquisition system. The five physiological channels included heart rate (HR), skin conductance level (SCL), respiration rate (Resp), and facial muscle activity from zygomaticus major (Zyg) and corrugator supercilii (Corr). After hearing each excerpt, participants provided subjective appraisals of felt emotion using two Likert-type scales ranging from 1 to 9: valence (least pleasant/most pleasant), and arousal (least excited/most excited).

Neural Network Models

Previous studies have explored various methods for predicting a listener's assessment of emotion from music. Laurier et al. (2009) used Support Vector Machines to predict discrete emotion categories from timbral, tonal, and rhythmic audio features. Kim and André (2008) created an automatic emotion recognition system based on physiological inputs of listeners. They extracted physiological features that were correlated with emotions and used an extended linear discriminant analysis to classify these emotions. Coutinho and Cangelosi (2011) used neural networks combining audio features and physiological features to predict emotion.

Our goal is not to provide a prediction method that adds to these existing approaches, but to provide a theoretical explanation on how musical emotion might be perceived, felt, and assessed by a listener through a suitable computational modeling approach. Towards this end, supervised feedforward neural networks with backpropagation (i.e. multilayer perceptrons) (Rumelhart, Hinton, & Williams, 1986; Haykin, 2008) are useful connectionist models that not only act as nonlinear regression functions for emotion prediction but also allow us to explore cognitive theories pertaining to emotion. The architecture of our feedforward networks is simple with one hidden layer providing the necessary level of nonlinearity. All the networks were implemented in Matlab.

Audio Network

From each of the 12 music excerpts, we extracted 13 low- and mid-level audio features relevant to rhythm, timbre, pitch, tonality, and dynamics using MIRtoolbox (Lartillot, Toiviainen, & Eerola, 2008): *rms*, *lowenergy*, *eventdensity*, *tempo*, *pulseclarity*, *zerocross*, *centroid*, *spread*, *rolloff*, *brightness*, *irregularity*, *inharmonicity*, and *mode*. We used the same training and test sets for all four networks to allow consistent comparisons of performance. Our training set consisted of eight of the 12 excerpts; we randomly chose two out of three excerpts from each emotion quadrant to obtain a good representation of emotion. The eight excerpts were M1, M2 (*agitated*), M4, M5 (*happy*), M7, M8 (*peaceful*), and M10, M11 (*sad*). The test set consisted of the remaining four excerpts, which were M3 (*agitated*), M6 (*happy*), M9 (*peaceful*), and M12 (*sad*).

The audio network's architecture was similar to the one we used for a previous study (Vempala & Russo, 2012) involving valence and arousal predictions based on training data from a set of 45 participants. The audio network was a supervised, feedforward network that consisted of 13 input units (i.e. one unit for each audio feature), one hidden layer with 13 units, and two output units, as shown in Figure 1. Our training set consisted of eight input and output vectors, corresponding to the eight training melodies. Each input vector had 13 values, one for each audio feature. The output vector consisted of mean arousal and valence values of all 20 participants corresponding to each input melody. Values of all the features, including arousal and valence, were scaled between 0 and 1 for efficient network learning (Bishop, 1996). The number of hidden units was set to be equal to the number of input units so as to avoid overfitting the network to the training set. Connection weights were initialized to random numbers between -0.05 and 0.05. The eight input vectors were randomly fed to the network. For each input vector, network predicted outputs were compared with desired outputs (i.e. mean participant ratings). The error was computed and the backpropagation algorithm was applied. Weight changes were stored, summed together at the end of each epoch, and applied to the connection weights. The learning rate was set to 0.1. The network was trained till the mean squared error was less than 0.012 (140,000 epochs).

We obtained results as shown in Figure 2. Results of the network are shown in comparison to the means of ratings provided by participants. Our results showed that the audio network performed well in predicting valence/arousal ratings for three of the four excerpts: M3 (Stravinsky), M6 (Strauss), and M12 (Mozart). To quantify the network's overall performance, we computed Euclidian distances between mean participant ratings and network-predicted values as a measure of performance error for each of the four test melodies. The network's performance error was 1.6 on average (on a scale from 1 to 9) or 14.2%, indicating that the network accuracy was 85.8%. On the valence dimension, the network accuracy was 91.6%. On the arousal dimension, the network accuracy was 83.2%. The largest error contribution came from the arousal dimension for M9 (Schumann).

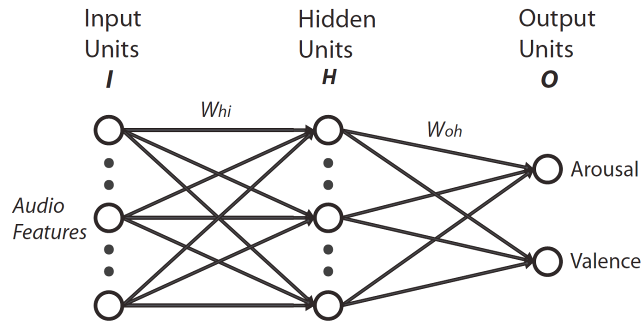

Figure 1: Audio network with 13 input units, 13 hidden units, and two output units. Connection weights W_{hi} and W_{oh} are from input to hidden layer, and hidden to output layer respectively. A subset of the 13 units are shown here.

Physiology Network

For each participant, each channel of physiological data was first standardized into z-scores. Physiological responses for a given participant were determined by subtracting the mean of baseline values (white noise) from the mean of excerpt values. The mean physiological response values (collapsed across the 20 participants) were used as inputs for this network. A detailed description of the physiology network is provided in Russo, Vempala, and Sandstrom (under review).

The physiology network's architecture was similar to the audio network, with one input layer, one hidden layer, and one output layer. It had five input units (i.e. one unit for each physiological feature/channel), five hidden units, and two outputs units. Again, the number of hidden units was set to be equal to the number of input units to avoid overfitting. The same training set with eight input and output vectors, corresponding to the eight training excerpts was used. Values of all five physiological features were scaled between 0 and 1. Instead of 13 audio feature values, each input vector consisted of five physiological feature values. Again, the output vector had the same, scaled, mean valence and arousal values corresponding to each excerpt. The procedure used for training was exactly similar to what was used for the audio network, with learning rate set to 0.1. The network was trained till the mean squared error was less than 0.012 (80,000 epochs).

Results of the network are shown in comparison to participant ratings, in Figure 2. The network predicted valence/arousal ratings for M3 (Stravinsky) and M9 (Schumann) well, and M6 (Strauss) reasonably well. Again, we quantified the network's overall performance by computing Euclidian distances between mean participant ratings and network-predicted values. The network's performance error was 1.48 on average (on a scale from 1 to 9) or 13%, indicating that the network accuracy was 87%. On the valence dimension, the network accuracy was 90.1%. On the arousal dimension, the network accuracy was 87.3%.

Audio-Physiology Network

Since overall prediction performances of both the audio network and the physiology network were more or less similar (85.8% vs. 87%), our next step involved combining both sets of features (i.e. 13 audio features and five physiological features) together into one input vector of 18 features. Our goal was to see if a combined feature set resulted in better prediction performance. Similar to the architecture of the previous two networks, the audio-physiology network had one input layer, one hidden layer, and one output layer. However, it had 18 input units, 13 hidden units, and two outputs units. The same training set was used with eight input and output vectors, corresponding to the eight training excerpts. The 18 input units in each input vector consisted of the same set of 13 audio and five physiological values used in the previous two networks, for each music excerpt. The output vector had the same, scaled, mean valence and arousal values corresponding to each excerpt. The procedure used for training was exactly similar to what was used for the previous networks, with the learning rate set to 0.1. The network was trained until the mean squared error was less than 0.012 (12,000 epochs). Results of the network are shown in comparison to participant ratings, in Figure 2. The network performed well with M3 (Stravinsky) and M9 (Schumann), and reasonably well with M12 (Mozart). Its worst prediction performance was with M6 (Strauss). Again, we quantified the network's overall performance by computing Euclidian distances between mean participant ratings and network-predicted values. The network's performance error was 1.28 on average (on a scale from 1 to 9) or 11.3%, indicating that the network accuracy was 88.7%. On the valence dimension, the network accuracy was 89.6%. On the arousal dimension, the network accuracy was 89.1%.

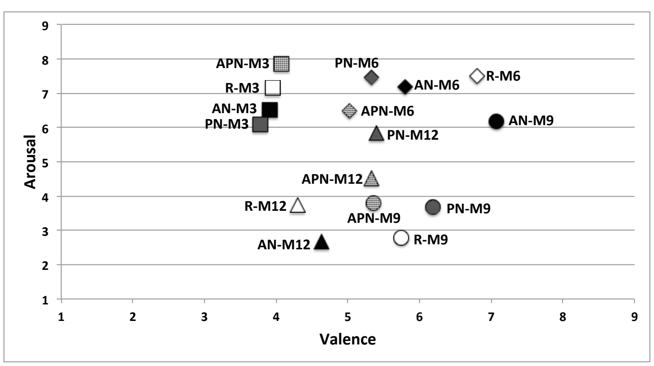

Figure 2: Participant valence/arousal ratings (R) on a scale of 1 to 9, and corresponding outputs from audio network (AN), physiology network (PN), and audio-physiology network (APN) for the four test excerpts M3, M6, M9, and M12.

Committee Machine

The audio-physiology network performed slightly better than both individual networks: audio network and physiology network, although the improvement in performance was relatively small. Two additional phenomena that can be observed are that (a) the training time for this network was considerably shorter, with only 12,000 epochs, and (b) the network performance was good for both arousal and valence dimensions. This suggests that there is emotion information exclusive to each network, and important for both dimensions, that may be shared when both sets of features are combined. The richer, combined feature set facilitates earlier convergence of the network towards a learned solution. However, this combination resulted in a small advantage as regards prediction performance. So, one plausible model of emotion assessment would be the following. Listeners make separate emotion assessments based on what they perceive from the music and what they feel when listening. Their final appraisal of emotion is based on a weighted judgment that takes contributions from both sources. This led us to implement our final model for emotion assessment: a committee machine (Haykin, 2008). The committee machine is a meta-level network, as shown in Figure 3. It combines outputs from each individual predictor to arrive at the overall output.

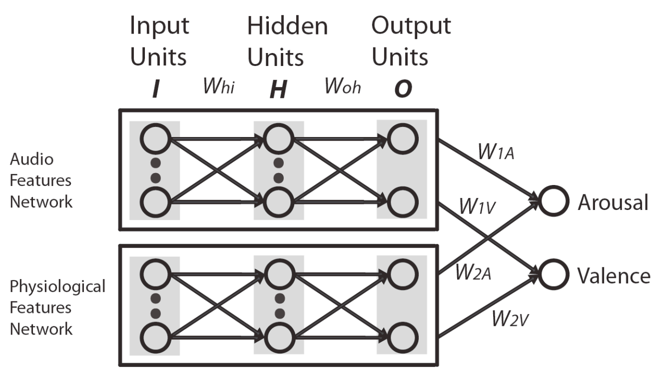

Figure 3: Committee machine consisting of a meta-level network that combines weighted outputs from the audio network and the physiological network. W_{1A} and W_{1V} are output weight contributions for arousal and valence from the audio network, and W_{2A} and W_{2V} are output weight contributions for arousal and valence from the physiology network.

As a first step, we used a basic committee machine that combined outputs through ensemble averaging. The meta-level network computed the average valence and arousal outputs from the audio network and the physiology network. In other words, the weight contributions for each individual network were 0.5. Results of this network are shown in comparison to participant ratings, in Figure 4. The network performed well with M3 (Stravinsky) and M12 (Mozart), and reasonably well with M6 (Strauss). Its worst prediction performance was on the arousal dimension for M9 (Schumann). Based on Euclidian distances between mean participant ratings and network-predicted values, the network's performance error was 1.32 on average (on a scale from 1 to 9) or 11.7%, indicating that the network accuracy was 88.3%. On the valence dimension, the network accuracy was 90.8%. On the arousal dimension, the network accuracy was 88.6%.

The results did not show any improvement in performance between the committee machine with ensemble averaging and the audio-physiology network. Hence, to obtain an optimal linear combination of the weights (Hashem, 1997) we performed multiple linear regression with outputs from individual networks as independent variables and mean participant valence/arousal ratings as dependent variables, for the four test excerpts. The model for arousal is provided in Equation 1. The model for valence is provided in Equation 2. Here, x_{1A} and x_{1V} refer to the arousal and valence outputs from the audio network on a scale from 0 to 1. Likewise, x_{2A} and x_{2V} refer to the arousal and valence outputs from the physiology network on a scale from 0 to 1. y_A and y_V refer to the overall arousal and valence outputs of the committee machine on a scale from 0 to 1.

$$y_A = 0.527\, x_{1A} + 1.166\, x_{2A} - 0.464 \quad (1)$$
$$y_V = 0.941\, x_{1V} - 0.357 x_{2V} + 0.199 \quad (2)$$

Based on Equation 1, for arousal, the meta-level network applies a weight of 0.527 to the audio network output, 1.166 to the physiology network output, and has a bias unit of weight -0.464. Likewise, for valence, based on Equation 2 the meta-level network applies a weight of 0.941 to the audio network output, -0.357 to the physiology network output, and has a bias unit of weight 0.199. To understand the significance of each individual network's contribution to the overall prediction, we computed their proportion contributions while ignoring the signs of the weights and the intercepts. For arousal, the weight contributions were 31.1% from the audio network, and 68.9% from the physiology network. For valence, the weight contributions were 72.5% from the audio network, and 27.5% from the physiology network.

Results of this network are shown in comparison to participant ratings, in Figure 4. The network performed well with M3 (Stravinsky), M9 (Schumann), and M12 (Mozart), and reasonably well with M6 (Strauss). Based on Euclidian distances between mean participant ratings and network-predicted values, the network's performance error was 0.86 on average (on a scale from 1 to 9) or 7.6%, indicating that the network accuracy was 92.4%. On the valence dimension, the network accuracy was 92.2%. On the arousal dimension, the network accuracy was 93.6%. These results clearly indicate that the committee machine with optimal linear combination of weights had the best prediction performance across all networks.

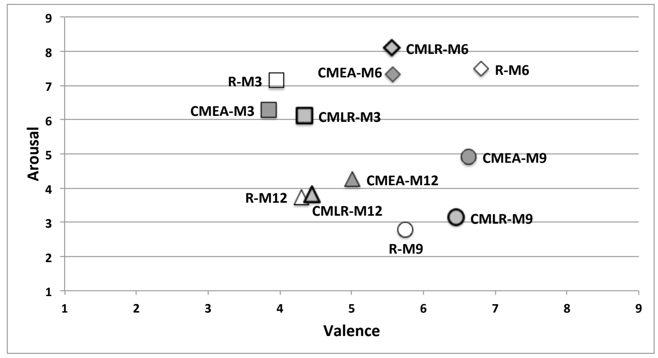

Figure 4: Participant valence/arousal ratings (R) on a scale of 1 to 9, and corresponding outputs from committee machine with ensemble averaging (CMEA), and committee machine with linear regression (CMLR) for the four test excerpts M3, M6, M9, and M12.

Discussion and Conclusions

Our results showed that the audio network had a performance of approximately 86%, and the physiology network had a performance of 87%. This suggests that there is sufficient information contained in both the audio features as well as the physiological responses for emotion prediction. So, if we look at the two models in isolation, then both positions are plausible for emotion assessment based on the performances; audio network favors the cognitivist position and physiology network favors the emotivist position. However, we notice a marginal improvement in prediction performance in the audio-physiology network (89%) when both kinds of features are combined as inputs to one network. This suggests that there is some complementary information regarding emotion that is available through feature combination, which is not accessible from individual networks. So, perhaps emotion appraisals are being made by listeners at a higher cognitive level through the weighting of separate processes, and not through a single combined representation. In other words, only after emotion assessments are formed through perceived features and felt emotion does the listener make an appraisal regarding the excerpt's emotional content through weighted contributions from both. A committee machine captures this theory well and illustrates how these weighted contributions might occur in the mind of the listener.

There are various limitations in this modeling study that we plan to address in the future. First and foremost, our data set is relatively small with just 12 excerpts. Hence, the committee machine had to be trained on the same test set that was used for testing individual networks, thus reducing its generalizability. We intend to expand our modeling to larger data sets where both individual networks learn using one training set, and the committee machine learns the optimal linear combination of weights through another training set that is separated from the test set. Next, in addition to having separate data sets for both types of training, we also intend to expand the size of each data set. Eight training melodies may not fully capture variance across participants. It is also likely that participants will vary in their physiological responses, depending on their preferences, training, and tendency to be absorbed by music (Sandstrom & Russo, 2013). Our next goal involves training separate networks that are tuned to different types of listeners. We contend that an approach that is sensitive to listener type will influence the respective weight of physiological and audio contributions to the combined network. Finally, all of the excerpts tested here were within the domain of classical music. We intend to include excerpts that include a wider range of musical genres in our future studies.

Despite these limitations, we believe that our model makes a significant contribution to the literature on music and emotion. To the best of our knowledge, no previous study has explored the cognitivist vs. emotivist debate through the lens of cognitively based computational modeling. Our model suggests that emotion appraisals are made by listeners through a combination of what they perceive from the music as well as what they feel during the listening process. We have modeled this using a meta-level network, thus supporting the view that the final appraisal is likely the outcome of a higher-level decision making process that combines together independent assessments from perception and feeling.

Acknowledgments

This research was supported by an NSERC Collaborative Research and Development grant awarded to Frank Russo co-sponsored by NSERC and WaveDNA, Inc. We thank Gillian Sandstrom for conducting experiments capturing participant physiological responses and emotion appraisals, and Christopher Lachine for assistance with data collection.

References

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

Bishop, C. M. (1996). *Neural networks for pattern recognition*. Oxford University Press.

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*, 921-937.

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

Hashem, S. (1997). Optimal linear combinations of neural networks. *Neural Networks, 10*, 599-614.

Haykin, S. (2008). *Neural networks and learning machines*. NJ: Prentice Hall.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioural and Brain Sciences, 31*, 559-621.

Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music sistening. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 2067-2083.

Konečni, V. J. (2008). Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 115-129.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag.

Laurier, C., Lartillot, O., Eerola, T., & Toiviainen P. (2009). Exploring relationships between audio features and emotion in music. *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music* (pp. 260-264). University of Jyväskylä Press, Jyväskylä.

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

Nyklicek, I., Thayer, J. F., & Van Doornen, L. J. P. (1997). Cardiorespiratory differentiation of musically-induced emotions. *Journal of Psychophysiology, 11*, 304-321.

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

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

Russo, F. A., Vempala, N. N., & Sandstrom, G. M. (under review). *Predicting musically induced emotions from physiological inputs: Linear and neural network models.* Manuscript submitted for publication.

Sandstrom, G. M., & Russo, F. A. (2010). Music hath charms: The effects of valence and arousal on the regulation of stress. *Music and Medicine, 2*, 137-143.

Sandstrom, G. M., & Russo, F. A. (2013). Absorption in music: A scale to identify individuals with strong emotional responses to music, *Psychology of Music, 41*, 216 - 228.

Vempala, N. N., & Russo, F. A. (2012). Predicting emotion from music audio features using neural networks. *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval, Lecture Notes in Computer Science*. London, UK.