Predicting Emotion from Music Audio Features Using Neural Networks

Naresh N. Vempala and Frank A. Russo

SMART Lab, Ryerson University
nvempala@psych.ryerson.ca

Abstract. We describe our implementation of two neural networks: a static feedforward network, and an Elman network, for predicting mean valence/arousal ratings of participants for musical excerpts based on audio features. Thirteen audio features were extracted from 12 classical music excerpts (3 from each emotion quadrant). Valence/arousal ratings were collected from 45 participants for the static network, and 9 participants for the Elman network. For the Elman network, each excerpt was temporally segmented into four, sequential chunks of equal duration. Networks were trained on eight of the 12 excerpts and tested on the remaining four. The static network predicted values that closely matched mean participant ratings of valence and arousal. The Elman network did a good job of predicting the arousal trend but not the valence trend. Our study indicates that neural networks can be trained to identify statistical consistencies across audio features to predict valence/arousal values.

Keywords: valence, arousal, music, emotion, machine learning, neural networks.

1 Introduction

A common reason for engaging in music listening is that music is an effective means of conveying and evoking emotions. Although these emotions may be subjective, based in part on the listener's cultural and musical background, there are commonalities in perceived emotion across different listeners based on the characteristics of the music. Several studies have attempted to predict emotion conveyed during music listening. Some studies have explored the relationship between physiological activity experienced by a listener and perceived emotion [1-2]. Others have explored the relationship between perceived emotion and the musical/acoustic features themselves [3-4]. While acknowledging that individual differences exist in the emotion conveyed by any one piece of music, we believe that it is legitimate to consider the modal appraisal and that this appraisal may be predicted on the basis of features extracted from the music.

Various methods have been used to represent emotion perceived by listeners. One common method, described by Russell's circumplex [5], involves representing emotion using a two-dimensional space with valence on the x-axis and arousal on the y-axis. Schubert [3] used the circumplex model to identify the relationship between musical features and perceived emotion. Changes in loudness and tempo were

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

positively correlated with changes in arousal, and melodic contour was positively correlated with valence. A few studies [4, 6] have used machine-learning techniques to predict discrete emotion categories based on audio features in musical excerpts. Laurier et al. [4] extracted timbral, tonal, and rhythmic audio features from film soundtrack excerpts that were evaluated by participants, for five different emotions. Based on this data, Laurier et al. used Support Vector Machines to classify excerpts into the five discrete emotions.

As opposed to classifying a musical excerpt into a discrete emotion category, our aim was to apply machine-learning techniques towards predicting valence and arousal values on the two-dimensional emotion space based on Russell's circumplex. A nonlinear regression function that predicts valence/arousal values offers a significant contribution to existing methods relating audio-based features to perceived emotion because, there are situations when participants may be unclear on the type of emotion conveyed by the music due to the overlapping and/or ambiguous nature of some emotions. In such cases, dimensional ratings provide a more effective means of representing the emotion conveyed by the music. We used machine learning, specifically feedforward neural networks, for predicting ratings on valence and arousal dimensions.

Although neural networks have been applied extensively in domains such as object recognition, speech and text recognition, they have been relatively underutilized in music cognition and music informatics. We designed two separate networks to predict listeners' mean valence and arousal ratings associated with musical excerpts. The first network was a standard, static feedforward neural network designed to predict valence and arousal ratings for the entire excerpt. The second network was an Elman network designed to predict valence and arousal ratings for 30-second increments of the excerpt while using the previous 30 seconds as context, to understand how context might influence ratings over time.

2 Feature Extraction and Data Collection

We used 12 classical music excerpts from 12 different composers as stimuli (see Table 1). Each excerpt lasted 120 seconds. These excerpts were selected 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)*. Excerpts were chosen based on previous work investigating emotional responses to music [7-8]. Using MIRtoolbox [9], we extracted 13 low- and mid-level features pertaining to dynamics, rhythm, timbre, pitch and tonality: *rms*, *lowenergy*, *eventdensity*, *tempo*, *pulseclarity*, *zerocross*, *centroid*, *spread*, *rolloff*, *brightness*, *irregularity*, *inharmonicity*, and *mode*. Values of all the features were normalized between 0 and 1. For the standard feedforward neural network, we used data from 45 participants (37 females, 2 males, 6 unknown; M_{age}=24.8, SD_{age}=8.2) with limited musical training. Participants heard each excerpt and rated two dimensions of emotion: unpleasant vs. pleasant (i.e., valence) and calm vs. excited (i.e., arousal) on a scale from 1 (least pleasant/least excited) to 9 (most pleasant/most excited). For the Elman network, we collected data from 9 participants (5 females, 4 males; M_{age}=30.4,

SD_{age}=6.8) with music training. Each of the 12 excerpts lasting 120 seconds was broken into four sequential segments of equal duration totaling 48 separate segments. For each excerpt, participants heard the four 30-second segments in sequence. After each segment, they provided their valence/arousal ratings following the same procedure as was used for the previous 45 participants. For the second, third, and fourth sequential segments of each melody, participants were told to assume that these segments were a continuation of the previous segment when providing their ratings. The same 13 features were extracted from each of the 48 segments using MIRtoolbox.

Table 1. 12 music excerpts with composers, emotion quadrants, and mean participant ratings.

Excerpt	**Composer**	**Composition**	**Quadrant**	**Mean Arousal**	**Mean Valence**
M1	Bartok	Sonata for 2 pianos and percussion (Assai lento)	Agitated	5.9	4.8
M2	Shostakovich	Symphony No. 8 (Adagio)	Agitated	7.1	4.1
M3	Stravinsky	Danse sacrale (Le Sacre du Printemps)	Agitated	6.6	4.1
M4	Beethoven	Symphony No. 7 (Vivace)	Happy	5.8	6.2
M5	Liszt	Les Preludes	Happy	6.4	6.0
M6	Strauss	Unter Donner und Blitz	Happy	7.0	6.8
M7	Bizet	Intermezzo (Carmen Suite)	Peaceful	2.5	6.3
M8	Dvorak	Symphony No. 9 (Largo)	Peaceful	2.4	5.7
M9	Schumann	Traumerei	Peaceful	2.9	5.2
M10	Chopin	Funeral March, Op. 72 No. 2	Sad	2.5	4.8
M11	Grieg	Aase's Death (Peer Gynt)	Sad	4.1	3.9
M12	Mozart	Requiem (Lacrimosa)	Sad	3.4	4.4

3 Methods

In this section we describe the linear and nonlinear regression methods that were used to (a) examine the relationship between audio features and valence/arousal ratings, and (b) predict valence/arousal ratings based on audio features[1].

3.1 Correlation of Audio Features with Emotion Ratings

As a first step towards understanding the pattern by which audio features might account for emotion ratings, we conducted correlational analyses between features and mean valence/arousal ratings of the 45 participants for the 12 excerpts. We performed a bivariate correlation analysis with the valence/arousal ratings as the first variable, and each of the 13 features as the second variable. We found a significant, strong positive correlation between arousal and five audio features: *pulseclarity* ($r(10) = .79, p < .01$), *zerocross* ($r(10) = .66, p < .05$), *centroid* ($r(10) = .80, p < .01$), *rolloff* ($r(10) = .80, p < .01$), and *brightness* ($r(10) = .73, p < .01$). For valence, apart from a significant, positive correlation with *lowenergy* ($r(10) = .59, p < .05$) and a marginally significant correlation with *mode* ($r(10) = .55, p = .06$), there was no correlation with the remaining audio features.

[1] Ground truth data is available at http://www.ryerson.ca/~nvempala/cmmr2012data.html

3.2 Multiple Regression for Predicting Emotion Ratings

Given that there was some significant correlation between a subset of the 13 audio features and valence/arousal ratings, we performed multiple linear regression to check for a linear relationship between features and ratings. We performed stepwise regression with features as independent variables (probability of *F* to enter = *.05*) and valence/arousal ratings as dependent variables. The model for arousal, Equation 1, was significant (*F(2,9) = 18.3, p < .01*) with *centroid* as the only predictor, accounting for 64.7% of the variance. The model for valence, Equation 2, was significant (*F(2,9) = 5.4, p < .05*) with *lowenergy* as the only predictor, accounting for 35.1% of the variance. Here, $y_{Arousal}$ and $y_{Valence}$ are the arousal and valence values on a scale from 1 to 9, respectively. In both equations, the addition of other variables did not lead to an increase in the explained variance. These results clearly suggest that a linear combination of the features does not account well for the valence and arousal ratings of participants. Hence we explored the possibility of predicting valence and arousal ratings through nonlinear combinations of audio features using neural networks.

$$y_{Arousal} = 4.94\, x_{centroid} + 2.14 \;. \qquad (1)$$

Here, $y_{Arousal}$ is the magnitude of arousal on a scale from 1 to 9.

$$y_{Valence} = 2.12\, x_{lowenergy} + 4.19 \;. \qquad (2)$$

3.3 Neural Networks for Predicting Emotion Ratings

3.3.1 Static Neural Network

Our first network implementation was a supervised, feedforward network with backpropagation. Our goal was to train the network to predict mean participant valence and arousal values for musical excerpts. We used one set of hidden units for our network. Network architecture consisted of 13 input units, 13 hidden units, and two output units as shown in Figure 1(a). As seen in Table 1, the mean valence/arousal ratings for each of the 12 music excerpts aligned with its expected emotion quadrant. Since valence and arousal ratings were from 1 to 9 and were plotted on the *x* and *y* axes respectively, *happy* excerpts needed to have values of $x > 5.0$, $y > 5.0$; *agitated* excerpts needed to have values of $x < 5.0$, $y > 5.0$; *sad* excerpts needed to have values of $x < 5.0$, $y < 5.0$; and *peaceful* excerpts needed to have values of $x > 5.0$, $y < 5.0$. From the 12 music excerpts, we randomly chose two out of three excerpts from each quadrant for our training set, which consisted of M1, M2 (*agitated*), M4, M5 (*happy*), M7, M8 (*peaceful*), and M10, M11 (*sad*). The test set consisted of the remaining four excerpts M3 (*agitated*), M6 (*happy*), M9 (*peaceful*), and M12 (*sad*).

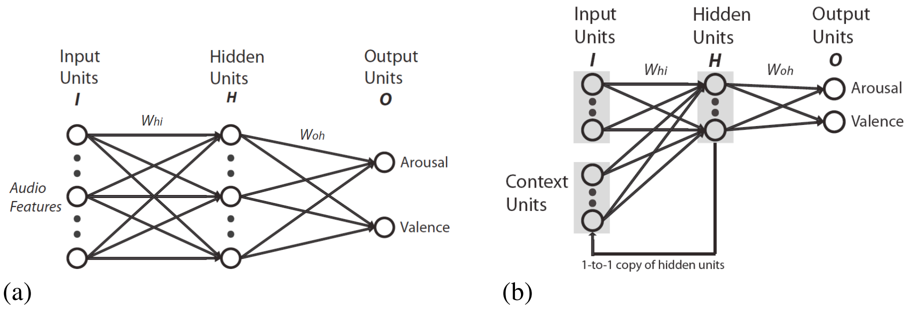

Fig. 1. (a) Static neural network with 13 input units, 13 hidden units, and two output units. W_{hi} indicates connection weights from input units to hidden units. W_{oh} indicates connection weights from hidden units to output units. Only a subset of the 13 input and hidden units are shown. (b) Elman network architecture showing input units, hidden units, output units, and context units. Hidden units from previous processing step are copied into context units for current step.

The network's task was to provide the valence and arousal values based on the 13 audio features. The output values fell within a range of 0 to 1. Since desired outputs were average valence/arousal ratings provided by participants on a scale from 1 to 9, the network outputs were rescaled back. The training set consisted of eight input and output arrays. Each input array had 13 values, one for each audio feature, and its corresponding output array had the two desired arousal and valence values. For example, if the input array being fed into the network was the feature set for excerpt M1 (Bartok), then the input array was [*rms*, *lowenergy*, … , *mode*]= $[0.5748, 0.7579, \ldots, 0.0052]^T$. Mean participant valence and arousal ratings for M1 were 5.9 and 4.8 respectively, resulting in normalized ratings of 0.61 and 0.48, respectively. Hence the desired output array would be $[\mathit{arousal}, \mathit{valence}]=[0.61, 0.48]^T$. To avoid overfitting the network, we kept the number of hidden units equal to the number of input units. The network was built, trained, and tested using the MATLAB programming language. The following procedure was used for training and testing the network:

1. Connection weights W_{hi} (input units to hidden units) and W_{oh} (hidden units to output units) were initialized to random numbers close to zero.
2. Input arrays were fed to the network from the training set in a randomized order. Inputs were passed through a sigmoidal function, multiplied with the connection weights W_{hi}, and summed at each hidden unit. Hidden unit values were obtained by passing the summed value at each hidden unit through a sigmoidal function. These values were multiplied with the connection weights W_{oh}, summed at each output unit, and passed through a sigmoidal function to arrive at the final output value for each output unit. Network outputs were compared to desired outputs and the error was computed. The backpropagation algorithm was applied and changes in connection weights were stored. At the end of the entire epoch, connection weights were updated with the sum of all stored weight changes.
3. The network was trained for approximately 10000 epochs by repeating step 2 to reduce the mean squared error to less than 0.01, and tested. During training, the learning rate parameter was initially set to 0.3 and reduced over time.

We obtained results as shown in Figure 2. The results show the network did a good job of predicting valence/arousal values for M3 (Stravinsky), M9 (Schumann), and M12 (Mozart). Although, predicted values for M6 (Strauss) fell in the expected quadrant (happy), they were not as close to the mean participant ratings. For the purpose of quantifying the network's performance, we computed the Cartesian distance between the mean participant rating and network-predicted value over all four test melodies. The network's performance error was 1.14 on average (on a scale from 1 to 9) or 14.3%, indicating that the network accuracy was 85.7%. These results clearly suggest that a nonlinear relationship exists between music audio features and their associated valence/arousal ratings.

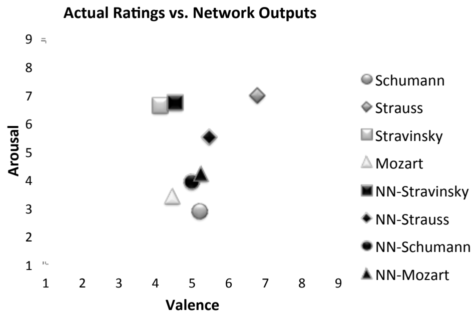

Fig. 2. Mean participant valence/arousal ratings (on a scale of 1 to 9) and corresponding neural network outputs for the four test melody excerpts. *NN* indicates neural network output.

3.3.2 Elman Neural Network for Predicting Emotion Ratings

Although the static network's performance was satisfactory, the network's implementation was based on participant valence/arousal ratings for the entire 120-second duration of each excerpt. It is reasonable to assume that a listener's appraisal of valence and arousal at any point in an excerpt is dynamic and sensitive to the previous few seconds of context. Hence, our next goal was to understand how context might influence a listener's ratings over time. Unlike a typical feedforward network, an Elman network uses context from the previous time-step as additional input for the current time-step. The architecture of the Elman network was almost identical to the static network with 13 input units for features, 13 hidden units, and two outputs units. The network had one additional component, which was a set of 13 context units, as shown in Figure 1(b). Context units were connected to the hidden units similar to input units, and had connection weights associated with them. For each step of input processing in the network, values of hidden units from the previous step were copied to the context units. This is explained below.

Each of the 12 music excerpts was broken into four equal chunks of 30-second duration, and data was collected for the 48 segments from 9 participants, as explained in Section 2. The network was trained on the same 8 music excerpts and tested on the remaining four excerpts, as chosen for the previous network to allow consistent comparison of network performance. The range of input and output values was identical to the static network. Since the network was being trained on the mean participant data for eight different melodic excerpts, and each excerpt had four 30-second segments, the training set consisted of 32 input and output arrays (four for each excerpt). The four input arrays for each excerpt were sequentially fed into the network. Context units were first initialized to 0. After the input array corresponding to the first segment was processed by the network, values of hidden units were copied into context units for processing the second segment. This process of one-to-one copy from hidden units to context units was continued for the third and fourth segments. This procedure was repeated for all eight excerpts. The network was built, trained, and tested using the MATLAB programming language. The procedure used for training and testing the network was identical to what was used for the static network.

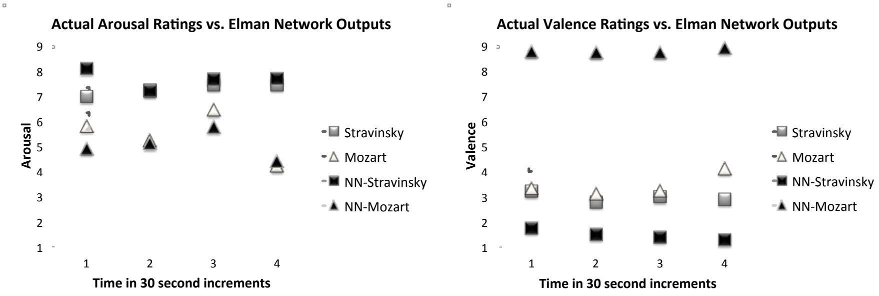

Fig. 4. Mean participant valence (*right*) and arousal (*left*) ratings (on a scale of 1 to 9) and corresponding neural network outputs for the four sequential 30-second segments of two excerpts. *NN* indicates neural network output.

We computed the mean error between participant ratings and network-predicted outputs across all segments of all four test melodies based on Cartesian distance. The network predicted at an average accuracy of 54.3% for all four segments. However, it performed better at predicting valence/arousal values for the final 30-second segment, at an average accuracy of 60%. Figure 4 shows a comparison of mean participant valence/arousal ratings and network values for excerpts M3 and M12. For arousal, the results clearly show that the network was good at predicting how participant ratings were influenced by context from previous segments – i.e., the trend over time was reasonably captured. However, for valence, although the relative changes over time are captured by the network to some extent, the absolute values are poorly predicted.

4 Conclusions and Future Directions

Our aim was to use neural networks to predict valence and arousal ratings of musical excerpts based on audio features within music. Results from the static network indicate that a network can be trained to identify statistical consistencies across audio features abstracted from music and satisfactorily predict valence/arousal values that closely match mean participant ratings. Our second goal was to highlight the role of musical context during listeners' appraisal of emotional content within music, and enable a neural network to utilize previous context during prediction. Results from the

Elman network showed that our network was more successful in capturing the trend of participant appraisals for arousal rather than valence. Three important improvements could be made to our current study involving emotion prediction.

First, having already trained a static network, we would like to identify features that are contributing most towards prediction of valence and arousal. This may be done by removing each feature and testing the network's performance in a step-by-step fashion. Second, a neural network's predictions depend largely on the size and type of training set provided. We intend to train our networks on larger datasets for improved generalizability. We also intend to develop separate static networks that will be trained on different types of musical genres and ratings drawn from different types of music listeners (e.g., trained vs untrained). This would enable us to (a) predict emotion ratings for an excerpt based on its genre and type of listener; and (b) identify salient music audio features for each genre and type of listener. Finally, we would like to improve the performance of our Elman network by training the network on data from a larger set of participants.

Acknowledgments. This research was supported by a Mitacs Elevate postdoctoral fellowship to Naresh N. Vempala co-sponsored by Mitacs and *wave*DNA, Inc. We thank Gillian Sandstrom and Christopher Lachine for assistance with data collection in this study.

References

1. Rainville, P., Bechara, A., Naqvi, N., Damasio, A. R.: Basic Emotions are Associated with Distinct Patterns of Cardiorespiratory Activity. International J. of Psychophysiology. 61, 5--18 (2006)
2. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Music Listening. IEEE Transactions on Pattern Analysis and Machine Intelligence. 30, 2067--2083 (2008)
3. Schubert, E.: Modeling Perceived Emotion with Continuous Musical Features. Music Perception. 21, 561--585 (2004)
4. Laurier, C., Lartillot, O., Eerola, T., Toiviainen P.: Exploring Relationships between Audio Features and Emotion in Music. In: 7th Triennial Conference of European Society for the Cognitive Sciences of Music, pp. 260--264. University of Jyväskylä Press, Jyväskylä (2009)
5. Russell, J. A.: A Circumplex Model of Affect. J. of Personality and Social Psychology. 39, 1161--1178 (1980)
6. Laurier, C., Herrera P.: Audio Music Mood Classification using Support Vector Machine. In: Proceedings of ISMIR. Vienna, Austria (2007)
7. Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., Dacquet, A.: Multidimensional Scaling of Emotional Responses to Music: The Effect of Musical Expertise and of the Duration of the Excerpts. Cognition & Emotion. 19, 1113--1139 (2005)
8. Sandstrom, G. M., Russo, F. A.: Music hath charms: The effects of valence and arousal on the regulation of stress. Music and Medicine. 2, 137-143 (2010)
9. Lartillot, O., Toiviainen, P., Eerola, T.: A Matlab Toolbox for Music Information Retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag (2008)