Psychophysiological Measures of Emotional Response to Romantic Orchestral Music and Their Musical and Acoustic Correlates

Konstantinos Trochidis, David Sears, Diêu-Ly Trân, and Stephen McAdams

Schulich School of Music, MGill University
{Konstantinos.Trochidis,David.Sears,
Dieu-Ly.Tran}@mail.mcgill.ca, smc@music.mgill.ca

Abstract. This paper examines the induction of emotions while listening to Romantic orchestral music. The study seeks to explore the relationship between subjective ratings of felt emotion and acoustic and physiological features. We employed 75 musical excerpts as stimuli to gather responses of excitement and pleasantness from 20 participants. During the experiments, physiological responses of the participants were measured, including blood volume pulse (BVP), skin conductance (SC), respiration rate (RR) and facial electromyography (EMG). A set of acoustic features was derived related to dynamics, harmony, timbre and rhythmic properties of the music stimuli. Based on the measured physiological signals, a set of physiological features was also extracted. The feature extraction process is discussed with particular emphasis on the interaction between acoustical and physiological parameters. Statistical relations among audio, physiological features and emotional ratings from psychological experiments were systematically investigated. Finally, a forward step-wise multiple linear regression model (MLR) was employed using the best features, and its prediction efficiency was evaluated and discussed. The results indicate that merging acoustic and physiological modalities substantially improves prediction of participants' ratings of felt emotion compared to the results using the modalities in isolation.

1 Introduction

With the recent advances in diverse fields of technology there is an emerging interest in recognizing and understanding the emotional content of music. Music emotion recognition plays an important role in music retrieval, mood detection, health care, and human-machine interfaces. Moreover, the entire body of music collections available to humans is increasing rapidly, and there is a need to intelligently classify and retrieve music according to the emotions they elicit from listeners. Indeed, emotion recognition is considered a key issue in integrating emotional intelligence within advanced human-machine interaction. Thus, there is strong motivation for developing systems that can recognize music-evoked emotions. In the following, we briefly review some of the work related to music emotion recognition based on acoustical and physiological features.

M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 44–57, 2013.

The emotions elicited during music listening are influenced by a number of structural music characteristics, including tempo, mode, timbre, harmony and loudness [1, 2]. In a pioneering publication [3], Li and Ogihara used acoustic features to classify music into mood categories. They achieved an accuracy of 45% using a database of 499 music clips selected from different genres annotated by a subject. They used a SVM-based multilabel classification method and determined the accuracy of their model using micro and macro-averaged precision. In [4] the authors used a similar variety of acoustic features for 800 classical music clips and achieved a recognition accuracy of 85%. Within the framework of Music Information Research Evaluation eXchange (MIREX), Tzanetakis reported an accuracy of 63.5% using a limited number of acoustic features [5]. Within the same framework, Peeters used a larger number of acoustic features and reported only a slight improvement [6], whereas in the next year Kim et al. proposed a system that reached a recognition accuracy of 65.7% [7].

Music emotion recognition has employed a number of approaches. In [8] the automatic detection of emotion in music was modeled as a multi-label classification task. A series of multi-label classification algorithms were tested and compared, with the predictive power of different audio features reaching an average precision of 81%. However, recent research in music emotion recognition from audio has shown that regression approaches can outperform existing classification techniques. In [12] the effectiveness of emotion prediction using different musical datasets (classical, film and popular music) was investigated. Their model had low generalizability between genres for valence (16%) and moderate generalizability between genres for arousal (43%), suggesting that valence operates differently depending on the musical genre. In [9] the authors used multiple acoustic features to predict pleasure and arousal ratings for music excerpts. They found that audio features are better for predicting arousal than valence and that the best prediction results are obtained for a combination of different features. In [10] a regression approach with combinations of audio features was employed in music emotion prediction. They found that the best performing features were spectral contrast and Mel-frequency cepstral coefficients (MFCC). The best performance, however, was achieved by a combination of features. In a recent publication [11], audio-based acoustical features for emotion classification were evaluated. A data set of 2090 songs was used, different audio features were extracted, and their predictive performance was evaluated. The results suggest that a combination of spectral, rhythmic and harmonic features yields the best results.

Despite the progress achieved on emotion recognition using audio features alone, the success of these various models has reached a glass ceiling. In order to improve the recognition accuracy of audio-based approaches, many studies have exploited the advantages of using additional information from other domains. This approach has led to the development of methods combining audio and lyrics [13-16], audio and tags [17], and audio and images [18], all of which result in moderate increases in recognition accuracy. There is a large body of studies establishing the relationship between physiological responses and musical emotions during music listening. Several studies have attempted to demonstrate whether the basic emotions induced by music are related to specific physiological patterns [19-23]. The relation between

discrete emotions and emotion-specific physiological response patterns predicted by theorists, however, still remains an open problem.

Indeed, the attempt to provide robust, incontrovertible evidence of emotional induction during music listening remains a tremendous challenge. The adoption of psychophysiological measures provides one possible solution, as they offer direct, objective evidence of autonomic and somato-visceral activation. Physiological responses during music listening include variations in heart rate, respiration electrodermal activity, finger temperature, and surface electromyography. Little attention, however, has been paid to the effect of physiological signals in music emotion recognition. The main problem of using physiological signals is the difficulty of mapping physiological patterns to specific emotional states. Furthermore, recording physiological signals requires the use of sensors and the analysis of signals that often reflect innervation by distinct branches of the autonomic nervous system (ANS). On the other hand, physiological signals have certain advantages, as they provide an objective measure of the listener's emotional state without relying on participant self-reports.

In [24] the authors used movie clips to induce emotions in 29 subjects, and combining physiological measures and subjective ratings achieved 83% recognition accuracy. In [25] the authors recorded four biosignals from subjects listening to songs and reached a recognition accuracy of 92%. Kim [26] used music excerpts to spontaneously induce emotions, measured electromyogram, electrocardiogram, skin conductivity and respiration changes, and then extracted the best features, achieving a classification accuracy of 70% and 90% for subject-independent and subject-dependent classification, respectively. Recently, in [27] a multimodal approach was based on physiological signals for emotion recognition, using music video clips as stimuli. They recorded EEG signals, peripheral physiological signals and frontal video. A variety of features was extracted and used for emotion recognition by using different fusion techniques. The results, however, demonstrated only a modest increase in recognition performance, indicating limited complementarity of the different modalities.

An important issue in musical emotion recognition is the modeling of perceived musical emotions. The two main approaches to modeling emotions in music-related studies are the categorical and the dimensional approach. According to the categorical approach, emotions are conceptualized as discrete entities, and there are a certain number of basic emotions, such as happiness, sadness, anger, fear and disgust, from which all subsequent emotional states are ultimately derived [28]. In music-related studies, emotion researchers often employ music-specific emotion labels (awe, frisson), or they use emotion terms that are more suitable to everyday musical experience (peacefulness, tenderness). Whereas the categorical model often employs these apparently distinct labels, in the dimensional approach all of the emotions experienced in everyday life are characterized (or supported) by two underlying dimensions: valence, which is related to pleasure-displeasure, and arousal, which is related to activation-deactivation. Thus, all emotions can be characterized in terms of varying degrees of valence and arousal [29, 30]. Both approaches have been recently investigated in relation to musical emotions [31], and their limitations were analyzed and discussed. In our study, the dimensional approach was employed because existing research in psychophysiology can find little evidence to suggest that there are emotion-specific physiological descriptors [21]. Rather,

psychophysiological responses appear to be related to the underlying dimensions of arousal and valence [32].

To the best of our knowledge, a combination of audio and physiological features has not been used in music emotion recognition research. There are, however, studies combining speech and physiological features for emotion recognition. In [33] and [34] the authors used combined voice data and physiological signals for emotion recognition. By fusing the features from both modalities, they achieved higher recognition accuracy compared with recognition results using the individual modalities.

The primary aim of the present work is to investigate the acoustic and physiological effects on the induction of emotions by combining audio and physiological features for music emotion recognition. Following [35] and [36], we argue that there is a possible route of emotion elicitation by peripheral feedback, and thus, that physiological arousal may influence the intensity and valence of emotions. In our study, we want to investigate the possibility of increasing the prediction rate of felt emotion through peripheral feedback by using acoustic and physiological features. The emotion recognition task is formulated as a regression problem, in which the arousal and valence ratings for each musical excerpt are predicted using a forward step-wise multiple linear regression model. During the experiment, music excerpts were employed as stimuli and the physiological responses of the listeners were measured, which included blood volume pulse, respiration rate, skin conductivity, and facial electromyographic activity. Both audio and physiological features were extracted, and the best features were combined and used for emotion recognition.

To combine the two modalities, it is important to determine at which stage in the model the individual modalities should be combined, or *fused*. A straightforward approach is to simply merge the features from each modality, called *feature-level fusion*. The alternative is to fuse the features at the decision level based on the outputs of separate single classifiers, called *decision-level fusion*. The existing literature on bimodal emotion recognition using speech features and physiological changes [34] demonstrates that feature-level fusion provides higher recognition accuracies compared to decision-level fusion. Therefore, in our study we employed feature-level fusion.

2 Methods

Participants. Twenty non-musicians (10 females) were recruited as participants (mean age = 26 years). The participants reported less than one year of training on an instrument over the past five years and less than two years of training in early childhood. In addition, all participants reported that they liked listening to Classical and Romantic music. The participants also filled out a demographic questionnaire and passed an audiometric test in order to verify that their hearing was normal.

Stimuli. Seventy-five music excerpts from the late Romantic period were selected for the stimulus set. The excerpts were 35 to 45 seconds in duration and selected by a

music theorist from the Romantic, late Romantic, or Neo-classical period (from 1815 to 1900). These genres were selected under the assumption that music from this time period would elicit a variety of emotional reactions along both dimensions of the emotion model. Moreover, each excerpt was selected to clearly represent one of the four quadrants of the two-dimensional emotion space formed by the dimensions of arousal and valence. Ten excerpts were chosen from a previous study [37] and 65 excerpts from our own personal collection. Aside from the high-arousal/negative-valence quadrant, which had 18 excerpts, the other three quadrants contained 19 excerpts each.

Procedure. During the experiment, five physiological signals were measured, including facial electromyography (EMG) of the smiling (zygomaticus major) and frowning (corrugator supercilii) muscles, skin conductance (SC), respiration rate (RR), and blood volume pulse (BVP). EMG measures the muscle activity through surface voltages generated when muscles contract. It is often employed to index emotional valence [38]. EMG sensors were placed above the zygomaticus major and corrugator supercilli muscles. SC is typically employed to index the physiological arousal of participants [38]. It measures the skin's ability to conduct electricity as a result of variations in sweat-gland activity. To measure SC, we positioned electrodes on the index and ring fingers of the non-dominant hand. RR is one of the characteristics of respiration change. A stretch sensor attached around the torso was used to record the breathing activity of the listeners. Heart rate variability (HRV) is the corresponding characteristic of heart rate activity derived from blood volume (BVP) pulse, which is measured with a plethysmograph attached to the middle finger of the non-dominant hand.

During the experiment the participants were asked to sit in a comfortable and relaxed position. They were told that it was crucial not to move during the baseline recordings and while the excerpts were playing. Following a practice trial to familiarize the participants with the experimental task, there was a two-minute baseline period in which their physiological measurements were taken. To remove inter-individual variability, seven additional one-minute baselines were recorded after each block of ten excerpts. Following each excerpt, participants rated their level of experienced excitement and pleasantness on 7-point continuous-categorical Likert scales.

3 Audio Feature Extraction

A theoretical selection of musical features following [12] was made based on musical characteristics such as dynamics, timbre, pitch, harmony, rhythm and structure using the MIR Toolbox for MATLAB [40]. For all features a series of statistical descriptors was computed, such as the mean, the standard deviation and the linear slope of the trend across frames. A total of 58 descriptors related to these features was thus extracted from the musical excerpts. Table 1 lists the various acoustic features and statistical descriptors extracted.

Table 1. The acoustic feature set extracted from the audio signals

Domain	No.	Name
Dynamics	1-3	RMS[1,2,3]
Timbre	4-18	Spectral Centroid[1,2,3] Spectral Flux[1,2,3] Spectral Spread[1,2,3] Spectral Entropy[1,2,3] Roughness[1,2,3]
Pitch	19-24	Chromagram[1,2,3] Pitch[1,2,3]
Tonality	25-36	Key Clarity[1,2,3] Key Strength[1,2,3] Harmonic Change Detection Function[1,2,3] Mode[1,2,3]
Rhythm	37-49	Fluctuation Pattern[1] Attack Times[1,2,3] Event Density[1,2,3] Tempo[1,2,3] Pulse Clarity[1,2,3]
Structure	50-58	Spectral Novelty[1,2,3], Rhythmic Novelty[1,2,3], Tonal Novelty[1,2,3]

Mean[1] Standard deviation[2] Slope[3]

3.1 Dynamics

We computed the RMS amplitude to examine whether the energy is evenly distributed throughout the signals, or to determine whether certain frames are more contrasted than others.

3.2 Timbre

A set of 5 features related to musical timbre were extracted from the Short-term Fourier Transform: Spectral Centroid, Spectral Flux, Spectral Spread and Spectral Entropy. Spectral Centroid represents the degree of timbre brightness. Spectral Flux is related to the degree of temporal evolution of the spectral envelope. Spectral Spread indicates the breadth of the spectral envelope. Spectral Entropy is used to capture the formants and the "peakedness" of the spectral distribution. Roughness was also derived from the peaks in the spectrogram based on the model in [41] and represents the sensory dissonance of the sound.

3.3 Pitch

Two pitch features were derived. The Chromagram represents the energy distribution of the signals wrapped around the 12 pitch classes. The Pitch was also computed using an advanced pitch extraction method which divides the audio signal into two channels below and above 1000 Hz and computes the autocorrelation of the low channel, the envelope of the high channel, and sums the autocorrelation functions [45].

3.4 Tonality

The signals were also analyzed according to their harmonic characteristics. A Chromagram representing the distribution of pitch-classes is created. Key Strength

computes the cross-correlation of the Chromagram with each possible major or minor key. The Key Clarity is the Key Strength of the key with the highest Key Strength out of all 24 keys [42]. The Harmonic Change Detection Function is a measure of the flux of the Tonal Centroid, and it captures the tonal diversity across time [43]. Finally, to model the Mode of each piece, a computational model that distinguishes major and minor excerpts was employed. It calculates an overall output that continuously ranges from zero (minor mode) to one (major mode) [44].

3.5 Rhythm

Fluctuation Pattern represents the rhythmic periodicity along auditory frequency channels) [46], and Attack Times refers to the estimation of note onset times. The Event Density measures the overall amount of simultaneous events in a musical excerpt. The tempo of each excerpt in beats per minute (BPM) was estimated by first computing a spectral decomposition of the onset detection curve. Next, the autocorrelation function was translated into the frequency domain in order to be compared to the spectrum curve, and the two curves were subsequently multiplied. Then a peak-picking algorithm was applied to the spectrum representation to select the best candidate tempo. The Pulse Clarity, a measure of the rhythmical and repetitive nature of a piece, was finally estimated by the autocorrelation of the amplitude envelope.

3.6 Structure

A degree of repetition was estimated through the computation of novelty curves [47] based on the spectrogram, the autocorrelation function, the key profiles and the Chromagram, each representing a different aspect of the novelty or static temporal nature of the music, such as Spectral, Rhythmic, and Tonal Novelty.

4 Physiological Feature Extraction

From the five psychophysiological signals, we calculated a total of 44 features, including conventional statistics in both the time and frequency domains. Table 2 lists the various physiological features extracted.

Table 2. The feature set extracted from the physiological signals

Domain	No	Name
Blood volume pulse	1-6	BVP[1,2,3,4,5,6]
Heart-rate	7-21	Heart-rate[1,2,3,4,5,6,7,8,9] SDNN[1,2,3,4,5,6]
Respiration-rate	22-26	BR**V**[1,2,3,4,5]
Skin conductivity	27-32	Skin conductivity[1,2,3,4,5,6]
Electromyography (Corrugator-Zygomaticus)	33-44	EMGc[1,2,3,4,5,6] EMGz[1,2,3,4,5,6]

Mean[1] Standard deviation[2] Median[3] Maximum[4] Minimum[5] Derivative[6] SpecVLF[7] SpecLF[8] SpecHF[9]

4.1 Blood Volume Pulse (BVP)

First, we normalized the blood volume pulse (BVP) signal by subtracting the preceding baseline from the signal. From the normalized BVP we computed time-series statistics, such as the mean, standard deviation, median, max, min and the derivative. To obtain HRV (heart rate variability) from the initial BVP signal, each signal was filtered, the QRS complex was detected, and finally the RR intervals (all intervals between adjacent R waves) or the normal-to-normal (NN) intervals (all intervals between adjacent QRS complexes resulting from sinus node depolarization) were determined. In the time-domain representation of the HRV time series, we calculated statistical features, including the mean, the standard deviation of all NN intervals (SDNN), the standard deviation of the first derivative of the HRV, the number of pairs of successive NN intervals differing by greater than 50 ms (NN50), and the proportion derived by dividing NN50 by the total number of NN intervals. In the frequency-domain representation of the HRV time series, three frequency bands are typically of interest: the very-low frequency (VLF) band (0.003-0.04 Hz), the low frequency (LF) band (0.04-0.15 Hz), and the high frequency (HF) band (0.15-0.4 Hz) [26]. From these sub-band spectra, we computed the dominant frequency and mean power of each band by integrating the power spectral densities (PSD) obtained using Welch's algorithm.

4.2 Respiration Rate

After detrending with the mean value of the entire signal and low-pass filtering with a cut-off frequency of 2.2 Hz, we calculated the Breath Rate Variability (BRV) by detecting the peaks in the signal. From the BRV time series, we computed the mean, standard deviation, median, max, min and derivative values.

4.3 Skin Conductivity (SC)

The mean, median, standard deviation, max, min, and derivative were extracted as features from the normalized SC signal and the low-passed SC signal, which used a 0.3 Hz cut-off frequency. In order to remove DC drift caused by physical processes like sweat evaporation off the surface of the skin, the SC signal was detrended by removing continuous, piecewise linear trends in the two low-passed signals: the very low-passed (VLP) signal was filtered with a 0.08 Hz cutoff frequency, and the low-passed (LP) signal was filtered with a 0.2 Hz cutoff frequency.

4.4 Electromyography (EMG)

From the EMG signals we took a similar approach to the one we employed for the SC signal. From the normalized and low-passed signals, the mean, median, max, min, and derivative of the signal were extracted as features.

5 Results

For the 75 excerpts a forward step-wise multiple linear regression (MLR) model between the acoustical and physiological descriptors and participant ratings was computed to gain insight into the importance of features for the arousal and valence dimensions of the emotion space. Table 3 provides the regression estimates and variance inflation factors (VIF) for each of the excitement and pleasantness ratings. The VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. Table 4 shows the outcome of the corresponding analysis of the physiological features. Finally, Table 5 shows the outcome of the analysis of the combined acoustic and physiological features.

Table 3. Mean audio features and standardized beta weights of the regression analysis for excitement and pleasantness

Excitement	***β***	**VIF**	**Pleasantness**	***β***	**VIF**
RMS **	.17	2.30	Key Clarity **	.51	1.06
Spectral Novelty **	-.21	1.56	Pitch **	.32	1.06
Spectral Spread **	-.41	2.10	Key Mode **	-.30	1.00
Spectral Entropy **	.24	1.15	Attack Times *	-.19	1.00
Spectral Centroid **	.25	1.13			
Pulse Clarity **	.18	2.00			

R^2 = .84 for Excitement. R^2 = .42 for Pleasantness. * $p < .05$, ** $p < .01$

Table 4. Physiological features and standardized beta weights of the regression analysis for excitement and pleasantness

Excitement	***β***	**VIF**	**Pleasantness**	***β***	**VIF**
SDNN[1]**	-.42	1.32			
Bvp[3]**	-.27	1.08	Heart-rate[2]**	-.37	1.00
Skin C[4]**	-.31	1.17	EMGc[4]**	-.28	1.00
EMGz[1]**	.25	1.11			
Skin C[1]*	.21	1.07			
Heart-rate[5]*	.20	1.08			

Mean[1] Standard deviation[2] Maximum[3] Minimum[4] SpecHF[5]

R^2 = .55 for Excitement. R^2 = .21 for Pleasantness. * p<.05, ** p<.01

Shown in Table 3, the regression model provides a good account of excitement (R^2 = .84) using only acoustic features (means of RMS energy, spectral centroid, spread, entropy and pulse clarity). Four features significantly predicted the pleasantness ratings (R^2 =.42): the means of Key Clarity, Mode, Pitch and Attack Times. Thus the results show that features related to characteristics of harmony, pitch, and articulation contribute most to pleasantness.

Table 5. Combined audio and physiological features and standardized beta weights of the regression analysis for excitement and pleasantness

Excitement	β	VIF	Pleasantness	β	VIF
RMS[1] **	.16	2.28	Key clarity[1] **	.46	1.00
Spectrum Novelty[1] **	-.21	1.59	Pitch[1] **	.23	1.06
Spectral Spread[1] **	-.34	2.29	Key mode[1] **	-.41	1.07
Spectral Entropy[1] **	.23	2.24	EMGZ[3] **	.36	1.06
Spectral Centroid[1] **	.26	1.40	Attack Time[1]**	-.24	1.06
SDNN[2]**	-.21	1.21	Heart-rate[4]**	-.22	1.13
Pulse clarity[1]**	.19	1.57			

Mean[1] Minimum[2] Derivative[3] SpecLF[4]

R^2 = .87 for Excitement. R^2 = .56 for Pleasantness. * p<.05, ** p<.01

Using only physiological features, the model provides an account of excitement with R^2 =.55 (see Table 4). The standard deviation of the NN intervals (SDNN) in the heart rate signals contributes most to excitement, along with the max value of the BVP and the mean and minimum of the skin conductance and EMGz signals. The power spectrum of the heart rate in the high frequency band (0.15-0.4 Hz) also contributes to this dimension. For the pleasantness dimension the model provides R^2 = .21 using the standard deviation of the heart rate signals and the minimum of the EMGc signals. Finally, using combined acoustical and physiological information (means of RMS energy, Spectral Centroid, Spread, Entropy, Pulse Clarity and the maximum value of SDNN), the model provides an account of excitement with R^2 = .87. The corresponding estimates for pleasantness use acoustic features related to Key Clarity, Mode, Pitch and the attack slope, and physiological features related to the EMGz and heart rate (R^2 = .56).

6 Discussion

In the present paper, the relationships among acoustic features and physiological features in emotional reactions to Romantic music were investigated. Our goal was to determine the importance of acoustic features in predicting the global emotional experience with music as measured with subjective ratings provided after each stimulus, and to explore the extent to which physiological activity may increase the prediction rate of emotion felt through peripheral feedback. A regression model based on a set of acoustic parameters and physiological features was systematically explored. The correlation analysis demonstrates that low- and mid-level acoustic features, such as RMS energy, Spectral Centroid, Spectral Spread, Spectral Entropy, and Pulse Clarity, significantly predict emotional excitement. The corresponding best features for the prediction of pleasantness are Key Clarity, Mode, Pitch and Attack Times. This result is in agreement with existing work on acoustic feature selection for emotion classification [10]. As far as the physiological features are concerned, the results indicate that features obtained from time and frequency analysis of the HRV series (SDNN, BVP), along with features of skin conductance, are decisive in the

prediction of participant ratings of excitement. Furthermore, features such as heart rate and corrugator EMG are important for pleasantness prediction. These findings are in agreement with previous research on music emotion recognition using physiological signals [26] and also support the findings of previous studies, according to which SC is linearly correlated to the intensity of arousal [22].

To the best of our knowledge a combination of audio and physiological features has not been employed in music emotion recognition tasks, and thus, we cannot compare our results with existing studies. There are, however, previous studies combining speech features and physiological responses for emotion recognition [33, 34]. The results of these studies show that the combination of speech and physiological features results in a moderate improvement of 3% for both valence and arousal. In our case the corresponding improvements are 3% and 14%, respectively, suggesting that the combination of acoustic and physiological features can provide more complementary information compared to the combination of speech and physiological features.

Existing results show that combined acoustic features provide better prediction for arousal than for valence [11, 10]. Therefore, the significant increase of pleasantness prediction by employing both acoustic and physiological features in our study is noteworthy here. It seems that EMG measures and spectral features of HRV play a significant role for the correct differentiation of positive and negative valence, and thus contribute substantially to improved valence prediction. This result is of particular importance, as valence is an otherwise elusive and opaque dimension in music emotion research. Moreover, MIR approaches thus far have only considered objective acoustical/musical features in an emotion recognition task, thereby failing to account for the role of physiological responses in the evocation of subjective feelings. Thus, any attempt to model a listener's affective state must also account for how subjective ratings of emotional experience may interact with the internal physiological state of the listening individual. Indeed, we hypothesize that our autonomic and somato-visceral reactions during music listening may influence the intensity and valence of our emotions through a process of peripheral feedback.

7 Future Work

There are several aspects in the work presented here that need to be addressed in future research. It remains to be investigated whether this particular model can be applied to other music-listening populations using other musical styles. Indeed, we believe that this approach could lead to fundamental advances in different areas of research because it may provide consistent descriptions of the emotional effects of particular musical stimuli. This, in turn, will have important implications for a number of disciplines, such as psychology and music therapy. In our study, feature-level fusion was employed. However, it appears that simply combining modalities with equal weighting does not always result in improved recognition accuracy. An alternative approach would be to decompose an emotion recognition problem into sub-problems, treating valence and arousal separately. For valence recognition, audio features could be used, whereas for arousal recognition physiological changes could be used.

Acknowledgments. Konstantinos Trochidis was supported by a post-doctoral fellowship by the ACN Erasmus Mundus network. This research was funded by grants to Stephen McAdams from the Social Sciences and Humanities Research Council of Canada and the Canada Research Chairs program. David Sears was supported by a Richard H. Tomlinson fellowship and a Quebec doctoral fellowship from the Programme de Bourses d'Excellence pour Étudiants Étrangers. The authors thank Bennett K. Smith for valuable technical assistance during the experiments.

References

1. Gabrielsson, A., Lindström, E.: The role of structure in the musical expression of emotions. In: Juslin, P.N., Sloboda, J.A. (eds.) Music and Emotion: Theory, Research, Applications. Oxford University Press, Oxford (2010)
2. Gomez, P., Danuser, B.: Relationships between musical structure and physiological measures of emotion. Emotion 7(2), 377–387 (2004)
3. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of the International Conference for Music Information Retrieval (ISMIR), Baltimore (2003)
4. Lu, L., Lu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signal. IEEE Transactions on Audio, Speech and Language Processing 14(1), 5–18 (2006)
5. Tzanetakis, G.: Marsyas submission to MIREX 2007. MIREX (2007)
6. Peeters, G.: A generic training and classification system for MIREX08 classification tasks: Audio music mood, audio genre, audio artist, audio tag. MIREX (2008)
7. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D.: Music emotion recognition: a state of the art review. In: Proceedings of the International Conference for Music Information Retrieval (ISMIR), pp. 255–266 (2010)
8. Trochidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. In: Proceedings of the International Conference for Music Information Retrieval, ISMIR (2008)
9. Song, Y., Simon, D., Pears, M.: Evaluation of musical features for emotion classification. In: Proceedings of the International Conference for Music Information Retrieval (ISMIR), pp. 523–528 (2012)
10. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proceedings of the International Conference for Music Information Retrieval (ISMIR), pp. 267–274 (2010)
11. Mc Dornan, K.F., Ough, S., Ho, C.C.: Automatic emotion prediction of song excerpts: Index construction, algorithm design and empirical comparison. Journal of New Music Research 36(4), 281–299 (2007)
12. Eerola, T.: Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres. Journal of New Music Research 40(4), 349–366 (2011)
13. Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H.H.: Toward multi-modal music emotion classification. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 70–79. Springer, Heidelberg (2008)
14. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representation from social tags. In: Proceedings of the International Conference for Music Information Retrieval, ISMIR (2009)

15. Hu, Y., Chen, X., Yang, D.: Lyrics based song emotion detection with affective lexicon and fuzzy clustering method. In: Proceedings of the International Conference for Music Information Retrieval, ISMIR (2009)
16. Schuller, B., Dorfner, J., Rigoll, D.: Determination of non-prototypical valence and arousal in popular music: features and performances. EURASIP Journal on Audio, Speech and Music Processing 2010, 1–20 (2010)
17. Turnbull, D., Barrington, L., Torres, D., Lanckiert, G.: Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech and Language Processing 16(2), 455–462 (2010)
18. Biscoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sodo, M.: Proceedings of the International Conference for Music Information Retrieval, ISMIR (2009)
19. Dunker, P., Nowak, S., Begau, N., Lanz, C.: Content-based mood classification framework and evaluation approach. In: Proceedings of ACM, New York (2008)
20. Nyklicek, I., Thayer, J., Van Doornen, L.: Cardiorespiratory differentiation of musically-induced emotion. Journal of Psychophysiology 11, 304–321 (1997)
21. Krumhansl, C.: An explanatory study of musical emotion and psychophysiology. Canadian Journal of Experimental Psychology 51, 336–352 (1997)
22. Khalfa, S., Peretz, I., Blondin, J., Manon, R.: Event-related skin conductance responses to music al emotion in humans. Neuroscience Letters 328, 145–149 (2002)
23. Lundquist, L., Carlsson, F., Hilmersson, P.: Facial electromyography, autonomic activity and emotional experience to happy and sad music. Paper Presented at the International Congress of Psychology (2002)
24. Nasoz, F., Lisetti, C.L., Alvarez, K., Finkelstein, N.: Emotional Recognition from Physiological Signals for User Modeling of Affect. In: Proceedings of the 3rd Workshop on Affective and Attitude User Modeling (2003)
25. Wagner, J., Kim, J., Andre, E.: From physiological signals to emotion. In: International Conference on Multimedia and Expo, pp. 940–943 (2005)
26. Kim, J.: Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(12), 2067–2083 (2008)
27. Koelstra, S., Muehl, C., Soleymani, M., Lee, J.D., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing 3(1), 18–31 (2011)
28. Ekman, P.: Are there basic emotions? Physiological Review 99(3), 550–553 (1992)
29. Russel, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
30. Withvliet, C.V., Vrana, S.R.: Play it again Sam: repeated exposure to emotionally evocative music polarizes liking and smiling responses and influences the affective reports, facial EMG and heart rate. Cognition & Emotion 21(1), 1–23 (2006)
31. Eerola, T., Vuoskoski, J.K.: A comparison of discrete and dimensional models of emotion in music. Psychology of Music 31(1), 18–49 (2010)
32. Bradley, M.M., Lang, P.J.: Emotion and Motivation. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, 3rd edn., pp. 581–607. Cambridge University Press, New York (2008)
33. Kim, J., Andre, E., Rehm, M., Vogt, T., Wagner, J.: Integrating information from speech and physiological signals to achieve emotion sensitivity. INTERSPEECH 2005, 809–812 (2005)
34. Kim, J., Andre, E.: Emotion recognition using physiological and speech signals in short term observation. In: ICGI 2006. LNCS (LNAI), vol. 4201, pp. 53–64. Springer, Heidelberg (2006)

35. Dibben, N.: The role of peripheral feedback in emotional experience with music. Music Perception 22(1), 79–115 (2002)
36. Scherer, K., Zentner, M.: Emotional effects of music: Production rules. In: Juslin, P., Sloboda, J. (eds.) Music and Emotion: Theory and Research, Oxford University Press, Oxford (2001)
37. 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(8), 1113–1139 (2005)
38. Dimberg, U.: Facial electromyography and emotional reactions. Psychophysiology 27(5), 481–494 (1990)
39. Rickard, N.: Intense emotional responses to music: a test of the physiological arousal hypothesis. Psychology of Music 32(4), 371–399 (2004)
40. Lartillot, O., Toiviainen, P.: MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio. In: International Conference on Music Information Retrieval, Vienna (2007)
41. Sethares, W.: Tuning, Timbre, Spectrum, Scale. Springer, Berlin (1998)
42. Gomez, E.: Tonal description of polyphonic audio for music content processing. INFORMS Journal on Computing 18(3), 294–304 (2006)
43. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio. In: Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, Santa Barbara, CA, pp. 26–31 (2006)
44. Saari, P., Eerola, T., Lartillot, O.: Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. IEEE Transactions in Audio, Language, and Speech Processing 19(6), 1802–1812 (2011)
45. Tolonen, T., Karjalainen, M.: A computationally efficient multipitch analysis model. IEEE Transactions on Speech and Audio Processing 8(6), 708–716 (2000)
46. Pampalk, E., Rauber, A., Merkl, D.: Content based organization and visualization of music archives. In: Proceedings of the 10th ACM International Conference on Multimedia, Juan les Pins, France, pp. 579–585 (2002)
47. Foote, J., Cooper, M.: Media segmentation using self-similarity decomposition. In: Proceedings of SPIE Storage and Retrieval for Multimedia Databases, vol. 5021, pp. 167–175 (2003)