On the Relationships between Music-induced Emotion and Physiological Signals

Xiao Hu
The University of Hong Kong Shenzhen
Institute of Research and Innovation
xiaoxhu@hku.hk

Fanjie Li
Sichuan University
fjlichn@gmail.com

Jeremy T. D. Ng
The University of Hong Kong
jeremyng@hku.hk

ABSTRACT

Emotion-aware music information retrieval (MIR) has been difficult due to the subjectivity and temporality of emotion responses to music. Physiological signals are regarded as related to emotion and thus could potentially be exploited in emotion-aware music discovery. This study explored the possibility of using physiological signals to detect users' emotion responses to music, with consideration of individual characteristics (personality, music preferences, etc.). A user experiment was conducted with 23 participants who searched for music in a novel MIR system. Users' listening behaviors and self-reported emotion responses to a total of 628 music pieces were collected. During music listening, a series of peripheral physiological signals (e.g., heart rate, skin conductance) were recorded from participants unobtrusively using a research-grade wearable wristband. A set of features in the time- and frequency- domains were extracted from the physiological signals and analyzed using statistical and machine learning methods. Results reveal 1) significant differences in some physiological features between positive and negative arousal and mood categories, and 2) effective classification of emotion responses based on physiological signals for some individuals. The findings can contribute to further improvement of emotion-aware intelligent MIR systems exploiting physiological signals as an objective and personalized input.

1. INTRODUCTION

Mood-based music discovery is a typical scenario of music information retrieval (MIR). Previous research has adopted content-based [11][27], collaborative filtering [7], or semantic-based [4] approach to recognize the emotion the music expressed and therefore enable emotion-aware MIR. Beyond research, Web-based music services are also available to support searching for music based on emotions, such as MoodFuse, Musicovery, etc.

However, emotion responses to music are subjective, varying from one user to another. They are also temporal in that the same user may respond to the same music differently at different times [33]. These make it challenging to optimize emotion-aware music retrieval.

Physiological signals such as heart rate, blood pressure and skin conductance were found to be related to people's emotion status [1][12][20] and they are deemed objective compared to self-reported emotion status which has been criticized as being subjective and sometimes inaccurate [2][3]. Therefore, physiological signals could potentially be exploited in emotion-aware music discovery. In addition, with the rapid development of wearable technology in recent years, peripheral physiological signals can be collected from users through small and less noticeable devices (e.g., wristband) in naturalistic settings in unobtrusive manners [13]. This advantage is not yet comparable by methods of gathering other physiological signals such as eye tracking and electroencephalography, and thus peripheral physiological signals are currently preferable for studying users' emotion responses to music in everyday life.

This study, therefore, aims to explore to what extent physiological signals measured by a wearable device are related to users' emotion responses to music played on an MIR system. Specifically, we are interested in the following research questions:

RQ1: Among features extracted from physiological signals collected during music listening, which ones differ significantly across different emotion responses?

RQ2: To what extent can physiological signals collected during music listening be used to predict users' emotion response to music?

As emotion responses to music may vary across listeners [37], we take into consideration listeners' characteristics by asking the following question:

RQ3: To what extent do prediction performances vary across different users and user characteristics (i.e. personality, music preferences)?

To answer these questions, a user experiment was conducted to collect data of users' interactions with a novel MIR system [17]. During the experiment, participants were asked to explore the music collection in the system while users' music listening behaviors and self-reported emotion responses to the music were recorded by the system. Simultaneously, physiological signals of the users were collected using a research-grade wearable wristband. Statistical tests and machine learning classifiers were applied to analyze the data, with classification performances compared across different classification

Attribution: Hu, X. Li, F. Ng, J. "On the Relationships between Music-induced Emotion and physiological signals", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

algorithms and users. Furthermore, collected in the pre-experiment questionnaire, participants' personality and music preferences were analyzed to see if they played a role in the relationships between physiological signals and emotion responses to music. As one of the first studies exploiting peripheral physiological signals in MIR, findings of this study can shed light on the feasibility of predicting users' emotion responses to music based on physiological signals and contribute to future implementation of emotion-aware MIR systems.

2. RELATED WORK

Work related to this study can be broadly categorized into physiological signal analysis in information retrieval and emotion-based music discovery.

2.1 Physiological Signals in Information Retrieval

Although using physiological signals as an implicit measurement of users' affective states during information retrieval process is still an emerging research topic, several recent studies have demonstrated its usefulness in predicting users' relevance judgments [2][3][28] and engagement levels [12]. Moshfeghi and Jose [28] used physiological features derived from heart rate, galvanic skin response, and skin temperature, along with facial expression features and behavioral features (i.e. dwell time), to predict users' relevance judgments in video retrieval tasks. They found that the combination of dwell time and heart rate features performed better for the task with entertainment-based search intention (i.e., when the main purpose of video search was to adjust arousal level or mood). Edwards and Kelly [12] combined skin conductance, heart rate with search behavior measures to evaluate users' levels of engagement, frustration, and stress when conducting searching tasks on a Web search interface. The results suggested that heart rate might be more associated with negative arousal, and skin conductance with positive arousal.

These studies in text and video information retrieval were encouraging and inspiring, yet there is little research on MIR exploiting physiological measures. Like many videos, music is a strong stimulus in eliciting emotion from listeners [23], and many users indeed listen to music for the very purpose of emotion modulation [16][34]. Considering the close relationship between music and emotion, as well as that between emotion and physiological signals, this study aims to help bridge the gap of incorporating physiological signals in MIR.

2.2 Emotion-aware MIR

The majority of previous research on music emotion recognition adopted content-based [11][27], collaborative filtering [7], and/or semantic-based [4] approaches which may suffer various shortcomings such as ignoring individual differences and the "cold start" problem (i.e., the recommendation performance is poor when few user ratings are available) [25]. Physiological signals, on the other hand, provide a new approach to understand users' emotion response to music. Several prior studies have probed physiology-based approach in MIR and yielded promising initial results. The Affective DJ project [9] used changes in users' skin conductance to detect users' mood based on which it helped users select music and generate playlists. Their evaluation results confirmed that skin conductance has a significant correlation with perceived excitement level of a song. Oliver et. al [30] also proposed a framework of automatic playlist generation by monitoring users' purpose of music listening and physiological responses (i.e. heart rate, galvanic skin response, respiration rate, and movement) to music. As an exemplar case of the framework, the MPTrain system was designed and implemented for selecting songs for runners to accompany their exercises. More recently, an affective music player (AMP) was developed to select music for mood enhancement by modeling the effects of music based on changes in skin conductance level and skin temperature [22]. Validation of the AMP found that lower skin temperatures were related to more positive emotions induced by music listening.

Notwithstanding the impact of these existing studies, the investigation on the relationship between physiological signals and emotion responses to music is still limited. Moreover, to the best of our knowledge, few studies have probed whether and how individual differences (in personality, music preferences, etc.) may play a role in such relationships. This study aims to bridge these gaps.

3. USER EXPERIMENT OF INTERACTIVE MUSIC SEARCH

The purpose of this experiment was to collect physiological signals and self-reported emotion response to music during music searching and listening. To encourage participants to interact with music, during the experiment, participants were asked to create a playlist using a novel Web-based music retrieval system. Participants' physiological signals during music listening were collected, as well as their self-reported emotion responses to each piece of music they listened to.

3.1 The MIR System

The Moody system [18] (now in its 3rd version) is a novel music retrieval system which supports searching for songs using several criteria: Genre (e.g., folk, jazz), Occasion (e.g., party, workout), Artist, Song, Album, and presents basic metadata and album image of each retrieved song (shown in Figure 1). Users can listen to any songs they are interested in using an HTML5 music player embedded in the Web interface of the system. They can also select any songs to add into a playlist at any time. Users' interactions with the systems (e.g., search, play) are recorded in the system logs.

The music collection hosted in the system is a subset of the Jamendo dataset, one of the world's largest digital services for free music. The subset of 10K tracks was obtained through the Grand Challenge in User Experience of the Music Information Retrieval Evaluation Exchange (MIREX)[17]. All the tracks are under the CC-BY license and thus the full tracks (of 60+ Gigabytes in total) can be freely listened to by the public. Metadata of the tracks (e.g., title, album, artist) as well as free tags were also obtained from Jamendo and displayed in the system to facilitate searching and browsing.

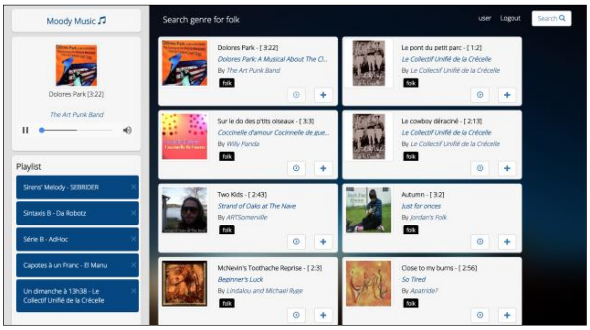

Moody Music
Dolores Park [3:22]
The Art Punk Band
Playlist
Sirens' Melody - SEBRIDER
Sintaxis B - Da Robotz
Série B - AdHoc
Capotes à un Franc - El Manu
Un dimanche à 13h38 - Le Collectif Unifié de la Crécelle
Search genre for folk
user
Logout
Search
Dolores Park - [3:22]
Dolores Park: A Musical About The C...
By The Art Punk Band
folk
Le pont du petit parc - [1:2]
Le Collectif Unifié de la Crécelle
By Le Collectif Unifié de la Crécelle
folk
Sur le do des p'tits oiseaux - [3:3]
Coccinelle d'amour Cocinnelle de gue...
By Willy Panda
folk
Le cowboy déraciné - [2:13]
Le Collectif Unifié de la Crécelle
By Le Collectif Unifié de la Crécelle
folk
Two Kids - [2:43]
Strand of Oaks at The Nave
By ARTSomerville
folk
Autumn - [3:2]
Just for onces
By Jordan's Folk
folk
McNevin's Toothache Reprise - [2:3]
Beginner's Luck
By Lindalou and Michael Rygge
folk
Close to my bums - [2:56]
So Tired
By Apatride?
folk

Figure 1. Interface of the Moody System (version 3)

3.2 Experiment Procedure

The experiment consisted of four main phases: 1) pre-experiment questionnaire; 2) instructions of the Moody system and the search task; 3) participants searching and listening to music; and 4) post-experiment questionnaire.

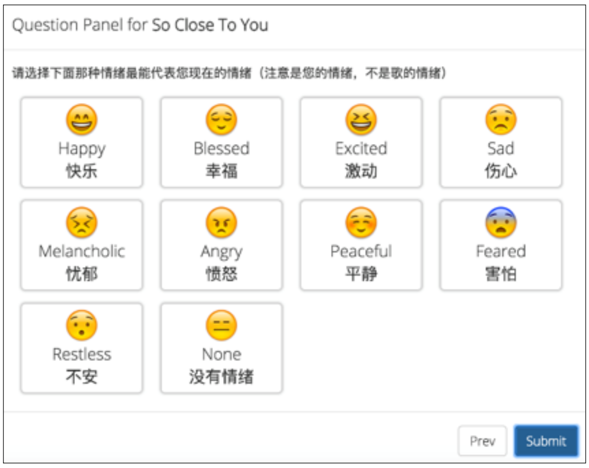
The pre-experiment questionnaire gathered information on demographics, music listening behaviors, music preference, and personality as measured by the Ten Item Personality scale (TIPI) [14] which contains 10 questions in five dimensions: extrovert - introvert; agreeable - disagreeable; open-close; stable - unstable; conscientious - unconscientious.

During phase 3, participants were asked to use the Moody system for no less than 40 minutes, looking for at least 10 songs they liked to make a playlist. They were encouraged to search for different types of music for a more diverse experience. For each music piece listened to for more than 30 seconds, participants would be prompted to answer two questions on their current emotion. The first question asked participants to give a score of arousal [32] on a scale of -10 (low arousal) to 10 (highly aroused), while the second question was to choose a mood category from a set of options adapted from [37] (Figure 2).

Also during phase 3, participants were asked to wear an Empatica E4 wristband [13] which is one of the few wearable devices designed specifically for measuring physiological signals for research purposes. This device supports real-time data acquisition and provides a secure API for raw data downloading. It has been used in emotion-related studies in psychology, health sciences, etc. with high reliability (e.g., [29]). For signal stabilization, the wristband was mounted on a participant's wrist 2 minutes before the search task started.

After conducting the task, the last phase of the experiment was for the participants to fill out a post-experiment questionnaire concerning their emotional states and general experience of the experiment including the search process.

The experiment took place in a computer classroom where participants worked on iMac computers. Earphones were used during the music searching and listening. Ethical consent forms were signed at the beginning and a nominal remuneration was paid at the end to compensate participants' time.


Question Panel for So Close To You

请选择下面那种情绪最能代表您现在的情绪（注意是您的情绪，不是歌的情绪）

Figure 2. Question on mood popped up in the Moody system. (Translation of the instruction: "Please choose one of the following moods that can best represent your current mood. (Note: This refers to your mood, not the mood expressed by the music piece).")

3.3 Data Collection

23 participants (15 males, 8 females) were recruited to join this experiment. All participants were undergraduate or graduate students in a comprehensive university in Hong Kong, whose majors ranged from Social Sciences, Science, Engineering, Business, Humanities & Arts to Medicine, with a diverse background in music knowledge and a relatively high frequency of music listening ranging from several times a week to a daily basis.

Physiological signals collected included electrodermal activity (EDA), blood volume pulse (BVP), inter beat interval (IBI), heart rate (HR) and skin temperature (TEMP). The sampling rates of EDA, BVP, HR, and TEMP are 4 Hz, 64Hz, 1Hz, and 4Hz respectively.

Among all participants, we collected arousal and mood ratings for 628 pieces of music. Each piece of music was listened to for approximately 80 seconds on average.

4. DATA ANALYSIS

4.1 Data Preprocessing

Before extracting features, we constructed two datasets: one with raw physiological signals, and the other with normalized physiological signals by z-score normalization [31]. As physiological signals vary across individuals [26], the normalization was conducted within each individual participant.

Time-series of physiological data in both datasets were then aligned with the starting and ending time of each music piece played in the experiment as recorded by the Moody system logs. Physiological data were then split into chunks corresponding to each song played by each participant in the experiment.

The emotion status reported by participants after listening to each piece was aligned with the physiological signals and taken as the ground truth labels in the classification analysis. The arousal and mood values rated by participants were then grouped into three main categories (i.e. positive, negative, neutral) for comparison using ANOVA and t-tests as well as classification. This resulted in 436 positive, 175 negative, and 17 neutral (i.e., 0) arousal ratings. For mood ratings, we combined the mood categories into positive (happy, blessed, etc.), negative (sad, fearful, etc.), and neutral moods (i.e. none), resulting in 387, 141, and 100 ratings respectively.

4.2 Feature Extraction

After data preprocessing, features of physiological signals during each music listening period were extracted based on time series and spectrum analysis. Table 1 summarizes the features extracted in this study.

Category	Features
Descriptive statistics of raw signals	Mean, Standard deviation, Median, Range [8], [20]
Time series features	Means of the absolute values of the 1st / 2nd differences of the raw / normalized signals [31]
Frequency domain features	HF, LF, LF/HF [36]
Physiological signal specific features	skin conductance response (SCR)[6], heart rate variability (HRV) [1]

Table 1. Features extracted from physiological signals.

Features considered in this study were those found closely linked to emotions by previous studies [1][6][8][20][31][36], including descriptive statistics such as median, range, standard deviation (stdev), means of the first difference in raw values (MFDR) and in normalized values (MFDN), means of the second difference in raw values (MSDR) and in normalized values (MSDN), low and high frequency (LF, HF) in frequency spectrum which was obtained through a Fast Fourier Transformation (FFT) on the time domain signals, as well as the ratio of the two (LF/HF). In addition, two features specific to physiological signals were considered. The first is skin conductance response (SCR) which depicts the phenomenon of the skin momentarily being a better conductor of electricity. The SCR is characterized by an increase in electrodermal response followed by a decrease in response [6]. It is generally related to stimulus arousal [24]. The second is heart rate variability (HRV) which measures the continuous interplay between sympathetic and parasympathetic influences on the heart rate. HRV has been found relevant to arousal as well [1].

4.3 Feature Analysis

Using a one-way ANOVA, we compared physiological features across the three arousal categories as well as the three mood categories (i.e., positive, negative, neutral). We also applied t-tests on the features between positive and negative categories of arousal and mood (i.e., without consideration of the neutral categories). As multiple comparisons were involved, Bonferroni correction [15] and Benjamini–Hochberg procedure [5]were applied to ANOVA and t-tests respectively to control Type I error. Features with significant differences across arousal and mood categories are identified.

4.4 Classification and Evaluation

A machine learning approach was applied to measure the extent to which physiological signals could be used to recognize users’ emotion responses to music listening, in positive and negative categories of arousal and mood. Specifically, we trained and compared the performance of several well-adopted classification models representative of different approaches, namely decision tree, k-Nearest Neighbor (k-NN), naïve Bayes and SVM.

As the sample distribution across the positive and negative categories is unbalanced, for each classifier and each category pair (i.e., arousal, mood), we constructed a balanced dataset by randomly selecting samples from the larger categories and performed a classification experiment on the balanced dataset. This process was repeated 10 times and within each time a 10-fold cross-validation was applied to evaluate the performances.

Besides classification based on the whole feature sets, the forward feature selection method was applied in combination with the classifiers to remove redundant and noisy features and improve classification performances.

In addition, to examine whether prediction performances vary across different user characteristics, we also conducted a classification experiment on data partitioned by participants, their personality, as well as their music preferences.

5. RESULTS AND DISCUSSION

5.1 Features with Significant Differences across Emotion Categories

Features found with significant differences in the ANOVA and t-test results after Bonferroni correction are shown in Table 2.

For arousal, both tests indicated that BVP, HR and EDA features differed significantly across categories. For mood, HR_range showed consistent significance across the two tests. In addition, HR and EDA seem more promising than other physiological signals in predicting and/or monitoring listeners' emotion responses to music. However, TEMP features, SCR, and time domain measures of HRV were not significant in either test.

Feature	**Arousal**		**Mood**	
	ANOVA	t-test	ANOVA	t-test
BVP_median	0.034	0.012	0.004	-
BVP_HF	0.049	0.007	-	-
HR_stdev	0.022	0.002	-	0.005
HR_range	0.010	< 0.001	0.039	0.003
HR_LF	0.022	< 0.001	-	0.004
HR_HF	0.028	0.001	-	0.006
EDA_MFDN	0.020	0.003	-	-
EDA_MSDN	0.017	0.003	-	0.008
EDA_LF/HF	0.036	-	-	-
IBI_median	-	-	0.006	-
IBI_mean	-	-	0.006	-

Table 2. Significant results (*p* values) of ANOVA and t-tests across extracted features.

5.2 Classification on the Dataset of All Listeners

Table 3 shows the performances of different classifiers with feature selection, on datasets balanced by repeated random sampling. Both accuracy and Cohen's kappa are used as performance measures. In general, the classification performances on the dataset aggregated across all users were low, with the best performances (k-NN) being around 60% in accuracy (baseline 50%) and lower than 0.2 in Cohen's kappa (indicating low agreement [35]).

Classifier	**Arousal**		**Mood**	
	Accuracy	Kappa	Accuracy	Kappa
Decision Tree	55.43%	0.109	60.04%	0.201
k-NN	59.97%	0.199	60.78%	0.216
Naïve Bayes	57.74%	0.155	58.83%	0.177
SVM	58.40%	0.168	59.50%	0.190

Table 3. Classification results on balanced datasets consisting of all listeners.

5.3 Classification on Individual Listeners

To examine whether and how prediction results differ across participants, the k-NN classifier was applied to datasets of individual participants. As the sample sizes of some participants were not sufficient for constructing balanced datasets of non-trivial sizes, these sets of experiments on individual participants were performed on unbalanced datasets. Therefore, F1 measure and Cohen's kappa were used and reported in Table 4. From the Cohen's kappa values in the results, we can see that the performances on individual listeners were much better than those on the dataset of all participants (Table 3). This difference implies that individual variability on physiological signal analysis might be too large to build generic classifiers that work for most (if not all) listeners.

The results also indicate that, for some participants, e.g. Users 5, 8, and 18, the prediction worked well for both arousal and mood, whereas for other participants, such as users 11, neither prediction exhibited good results (Cohen's kappa values were lower than 0.2). The variability of prediction performances across participants corroborates with findings in existing research that physiological signals are highly individual dependent [19][26].

User	No. of songs	Arousal		Mood	
		F measure	Kappa	F measure	Kappa
User 1	24	90.00%	0.400	78.57%	0.294
User 2	24	81.48%	0.577	62.50%	0.262
User 3	28	78.57%	0.571	90.00%	0.800
User 4	45	80.00%	0.537	93.33%	0.700
User 5	43	98.70%	0.788	93.55%	0.604
User 6	30	81.25%	0.490	84.85%	0.516
User 7	29	86.36%	0.435	93.03%	0.516
User 8	17	96.54%	0.767	96.55%	0.767
User 9	24	94.44%	0.694	96.30%	0.765
User 10	33	87.80%	0.670	92.31%	0.747
User 11	21	93.75%	-0.063	97.14%	0.000
User 12	18	88.89%	0.722	93.33%	0.843
User 13	25	94.74%	0.614	90.48%	0.405
User 14	29	91.30%	0.580	95.00%	0.750
User 15	31	80.00%	0.427	83.72%	0.440
User 16	30	92.59%	0.259	92.31%	0.423
User 17	32	88.89%	0.595	85.00%	0.475
User 18	33	98.36%	0.784	98.31%	0.784
User 19	33	84.45%	0.507	90.32%	0.750
User 20	35	84.00%	0.380	91.67%	0.586
User 21	16	91.67%	0.673	96.00%	0.818

Table 4. Classification performance on individual participants.

5.4 Classification on Participant Groups

Table 5 shows classification performance of users with different personalities. Personality was determined based on responses to the TIPI questionnaire [14] which consisted of five personality dimensions. For each dimension, each user was categorized to either end based on their answers to the two question items in that dimension. This set of experiments were also conducted on balanced datasets, with accuracy and Cohen's kappa values reported (Table 5). Compared to performances on the dataset of all listeners (Table 3), classification performances on some of the personality dimensions were better. In particular, predictions on users with Agreeable personality reached Cohen's kappa values of 0.581 (for arousal) and 0.682 (for mood) which are deemed as medium and high agreement levels respectively [35].

Another observation is that the personality dimensions with relatively high classification performances had lower numbers of users compared to other personality dimensions. A correlation analysis revealed significant negative correlations between the number of users and classification performances (r = - 0.78 for both measures of arousal, $p = 0.008$; and r = - 0.62 for both measures of mood prediction, $p = 0.054$). This again implies the significance of individual differences in physiological signal analysis. A suggestion for future research is thus to analyze physiological signals within individual users.

Personality	No. of users	**Arousal**		**Mood**	
		Accuracy	Kappa	Accuracy	Kappa
Extrovert	12	62.42%	0.248	65.00%	0.300
Introvert	11	67.56%	0.351	63.65%	0.273
Agreeable	3	79.03%	0.581	84.09%	0.682
Disagreeable	20	62.50%	0.250	61.64%	0.233
Open	7	64.51%	0.290	66.58%	0.332
Close	16	62.96%	0.259	74.69%	0.494
Stable	9	64.07%	0.281	69.38%	0.388
Unstable	14	62.39%	0.248	62.41%	0.248
Conscientious	6	69.75%	0.395	68.91%	0.378
Unconscientious	17	60.65%	0.213	61.05%	0.221

Table 5. Classification performances of each personality dimension.

Besides personality, users' music preference might play a role in their emotion responses to music [21]. Therefore, we grouped the participants based on their self-reported genre preferences using the k-means clustering algorithm, with the optimal k value selected by the Davies Bouldin index [10]. The results yielded three clusters corresponding to preferences as shown in Table 6. We then conducted classification experiments on participants in each of the clusters, using the k-NN algorithm and balanced datasets. Prediction performances (Table 6) were higher than those on the whole dataset (Table 3), and the performances on the mood classification of cluster 2 were comparable to those of some individual users (Table 4). These results indicate that certain music preferences might play a role in predicting emotion responses to music based on physiological signals. Future work is needed to further investigate this phenomenon, preferably with larger samples.

Cluster	Preferences	No. of users	**Arousal**		**Mood**	
			Accuracy	κ	Accuracy	κ
0	Pop only	7	64.26%	0.285	71.32%	0.426
1	Classical, Folk, Pop	10	65.17%	0.303	64.06%	0.281
2	Electronica, Rock, Pop	6	69.02%	0.380	76.76%	0.535

Table 6. Classification performances of participant clusters based on music preferences.

6. CONCLUSION AND FUTURE WORK

This paper presented a study towards recognizing users' emotion response to music using physiological data obtained from wearable sensors. ANOVA and t-tests revealed that heart rate (HR) and electrodermal activity (EDA) features were consistently significant in both arousal and mood dimensions. This finding provides empirical evidence for feature extraction and selection in future studies. In predicting emotion responses to music based on physiological signals during music listening, predictions on individual participants showed promising performances as well as large performance differences across individuals. These results verified that the predictability of emotion responses to music based on physiological signals may vary from person to person. Additionally, the classification experiments conducted on data partitioned by personality dimensions and music preferences illustrated that classification based on physiological signals might be more effective for users with certain personality traits or genre preferences, such as being agreeable or preferring Electronica and Rock music. However, these results were confounded with the sample size, as the number of users in each personality category was negatively correlated with performance measures, further indicating variability across users and suggesting that individual-based analysis might be more fruitful for exploiting physiological signals. The results reported in this paper demonstrate the potential of physiological sensing techniques in emotion-aware MIR. This opens up a number of possibilities in future MIR systems and services, such as recommending music based on users' current physiological measures and maintaining mood-based playlists which can be adjusted in real time based on changes in physiological signals, etc.

Future studies will be conducted to further investigate in which circumstances physiological signals are more effective in predicting emotion responses to music. Combinations of factors will be considered such as the matching between music preferences and the music pieces being listened to. In the next stage of our research, we will also explore the effect of incorporating music features (e.g., acoustic, emotion, occasion, etc.) in the prediction as well as users' behavioral logs recorded in the user experiment. Besides, the experiment in this study was run in laboratory settings. To achieve a higher level of ecological validity, future experiments can be extended to the everyday environment of the participants and for longer time spans.

7. ACKNOWLEDGEMENT

The authors acknowledge the support given by the National Natural Science Foundation of China: 61703357, and the Research Grants Council of the HKSAR Government, #T44-707/16/N, under the Theme Based Research Scheme. We also thank the anonymous reviewers for their helpful suggestions.

8. REFERENCES

[1] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional

responding.," *Review of General Psychology*, Vol. 10, No. 3, pp. 229–240, 2006.

[2] I. Arapakis, I. Konstas, and J. M. Jose: "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," *Proc. of the 17th ACM international conference on Multimedia*, pp. 461-470, 2009

[3] O. Barral, M. J. A. Eugster, T. Ruotsalo, M. M. Spapé, I. Kosunen, N. Ravaja, S. Kaski, and G. Jacucci: "Exploring peripheral physiology as a predictor of perceived relevance in information retrieval," *Proc. of the 20th International Conference on Intelligent User Interfaces*, pp. 389-399, 2015.

[4] M. Barthet, G. Fazekas, A. Allik, and M. Sandler: "Moodplay: an interactive mood-based musical experience," *Proc. of the Audio Mostly 2015 on Interaction with Sound*, p. 3, 2015.

[5] Y. Benjamini, Y. Hochberg: "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289-300, 1995.

[6] W. Boucsein: "Electrodermal Activity," *Plenum Series in Behavioral Psychophysiology and Medicine, Plenum Press*, 1992.

[7] J. Broekens, A. Pronker, and M. Neuteboom: "Real time labeling of affect in music using the affect button". *Proc. of the 3rd international workshop on Affective interaction in natural environments*, pp. 21-26, 2010.

[8] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun: "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," *Proc. of the 12th international conference on Entertainment and media in the ubiquitous era*, pp. 13-17, 2008

[9] F. Dabek, J. Healey, and R. Picard: "A new affect-perceiving interface and its application to personalized music selection," *Proc. from the 1998 Workshop on Perceptual User Interfaces,* 1998

[10] D. L. Davies and D. W. Bouldin: "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, pp. 224–227, Apr. 1979.

[11] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen: "Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation," *ACM Transactions on Interactive Intelligent Systems*, Vol. 5, No. 1, pp. 1–36, Mar. 2015

[12] A. Edwards and D. Kelly: "Engaged or Frustrated: Disambiguating Emotional State in Search," *Proc. of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 125-134, 2017.

[13] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti: "Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," *Proc. of the 4th International Conference on Wireless Mobile Communication and Healthcare*, pp. 39-42, 2014

[14] S. D. Gosling, P. J. Rentfrow, and W. B. Swann: "A very brief measure of the Big-Five personality domains," *Journal of Research in Personality*, Vol. 37, No. 6, pp. 504–528, Dec. 2003.

[15] S. Holm: "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics* pp. 65-70, 1979.

[16] X. Hu, N. Kando, "Evaluation of Music Search in Casual-Leisure Situations," *Proc. of Search for Fun Workshop at the Information Interaction in Context conference (IIiX)*, 2014.

[17] X. Hu, J. Lee, D. Bainbridge, K. Choi, P. Organisciak and J. Downie, "The MIREX grand challenge: A framework of holistic user-experience evaluation in music information retrieval", *Journal of the Association for Information Science and Technology*, Vol. 68, no. 1, pp. 97-112, 2017.

[18] X. Hu, V. Sanghvi, B. Vong, P. J. On, C. Leong, and J. Angelica. "Moody: A web-based music mood classification and recommendation system", *Proc. of 9th International Conference on Music Information Retrieval*, Philadelphia, U.S. 2008.

[19] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D'Mello: "Affect detection from multichannel physiology during learning sessions with AutoTutor," *International Conference on Artificial Intelligence in Education*, pp. 131-138, 2011.

[20] M. S. Hussain, R. A. Calvo, and F. Chen: "Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference," *Interacting with Computers*, Vol. 26, No. 3, pp. 256–268, Jun. 2013.

[21] M. Iwanaga and Y. Moroki: "Subjective and Physiological Responses to Music Stimuli Controlled Over Activity and Preference," *Journal of Music Therapy*, Vol. 36, No. 1, pp. 26–38, Mar. 1999.

[22] J. H. Janssen, E. L. van den Broek, and J. H. D. M. Westerink, "Tune in to your emotions: a robust personalized affective music player," *User Modeling and User-Adapted Interaction*, Vol. 22, No. 3, pp. 255–279, Oct. 2011.

[23] P. Kenealy: “Validation of a music mood induction procedure: Some preliminary findings,” *Cognition & Emotion*, Vol. 2, No. 1, pp. 41–48, Mar. 1988.

[24] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon: “Event-related skin conductance responses to musical emotions in humans,” *Neuroscience Letters*, Vol. 328, No. 2, pp. 145–149, Aug. 2002.

[25] S. Khusro, Z. Ali, and I. Ullah: “Recommender Systems: Issues, Challenges, and Research Opportunities,” *Information Science and Applications (ICISA) 2016*, pp. 1179–1189, 2016.

[26] A. M. Kiviniemi, H. Hintsala, A. J. Hautala, T. M. Ikaheimo, J. J. Jaakkola, S. Tiinanen, T. Seppanen, and M. P. Tulppo: “Impact and management of physiological calibration in spectral analysis of blood pressure variability,” *Frontiers in Physiology*, Vol. 5, Dec. 2014.

[27] F. F. Kuo, M. F. Chiang, M. K. Shan, and S. Y. Lee: “Emotion-based music recommendation by association discovery from film music,” *Proc. of the 13th annual ACM international conference on Multimedia*, pp. 507-510, 2005.

[28] Y. Moshfeghi and J. M. Jose: “An effective implicit relevance feedback technique using affective, physiological and behavioral features,” *Proc. of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 133-142, 2013.

[29] S. C. Müller and T. Fritz: “Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress,” *Proc. of the 37th International Conference on Software Engineering IEEE*, Vol 1, 2015.

[30] N. Oliver, L. Kregor-Stickles: “PAPA: Physiology and Purpose-Aware Automatic Playlist Generation,” *Proc. of 7th International Conference on Music Information Retrieval*, pp. 250-253, Victoria, Canada. 2006

[31] R. W. Picard, E. Vyzas, and J. Healey: “Toward machine emotional intelligence: analysis of affective physiological state,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 10, pp. 1175–1191, 2001.

[32] J. A. Russell: “A circumplex model of affect.”, *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161–1178, 1980.

[33] L. Su, C. C. M. Yeh, J.-Y. Liu, J.-C. Wang, and Y.-H. Yang: “A Systematic Evaluation of the Bag-of-Frames Representation for Music Information Retrieval,” *IEEE Transactions on Multimedia*, Vol. 16, No. 5, pp. 1188–1200, Aug. 2014.

[34] T. F. M. Ter Bogt, J. Mulder, Q. A. W. Raaijmakers, and S. Nic Gabhainn: “Moved by music: A typology of music listeners,” *Psychology of Music*, Vol. 39, No. 2, pp. 147–163, Aug. 2010.

[35] A. J. Viera and J. M. Garrett: “Understanding interobserver agreement: the kappa statistic,” *Physical Therapy*, Mar. 2005.

[36] W. Wu, Y. Gil, and J. Lee: “Combination of Wearable Multi-Biosensor Platform and Resonance Frequency Training for Stress Management of the Unemployed Population,” *Sensors*, Vol. 12, No. 12, pp. 13225–13248, Sep. 2012.

[37] Y. H. Yang and Y. C. Teng: “Quantitative Study of Music Listening Behavior in a Smartphone Context,” *ACM Transactions on Interactive Intelligent Systems*, Vol. 5, No. 3, pp. 1–30, Sep. 2015.