Affect Detection and Classification from Non-Stationary Physiological Data

Omar AlZoubi and Davide Fossati
Computer Science
Carnegie Mellon University in Qatar
Doha, Qatar
oalzoubi@cmu.edu, dfossati@cmu.edu

Sidney D'Mello
Department of Computer Science
The University of Notre Dame
Notre Dame, IN, USA
sdmello@nd.edu

Rafael A. Calvo
School of Electrical and Information Engineering
The University of Sydney
Sydney, Australia
Rafael.Calvo@sydney.edu.au

***Abstract*— Affect detection from physiological signals has received a great deal of attention recently. One arising challenge is that physiological measures are expected to exhibit considerable variations or non-stationarities over multiple days/sessions recordings. These variations pose challenges to effectively classify affective sates from future physiological data. The present study collects affective physiological data (electrocardiogram (ECG), electromyogram (EMG), skin conductivity (SC), and respiration (RSP)) from four participants over five sessions each. The study provides insights on how diagnostic physiological features of affect change over time. We compare the classification performance of two feature sets; pooled features (obtained from pooled day data) and day-specific features using an updatable classifier ensemble algorithm. The study also provides an analysis on the performance of individual physiological channels for affect detection. Our results show that using pooled feature set for affect detection is more accurate than using day-specific features. The corrugator and zygomatic facial EMGs were more reliable measures for detecting valence than arousal compared to ECG, RSP and SC over the span of multi-session recordings. It is also found that corrugator EMG features and a fusion of features from all physiological channels have the highest affect detection accuracy for both valence and arousal.**

Keywords—Affect; emotion; classifier ensembles; physiological; Non-Stationary

I. INTRODUCTION

There is increased attention on using physiological signals in affect detection systems that are able to detect either discrete emotional categories or affective dimensions of valence and arousal [1-3]. Physiological responses such as facial muscle activity, skin conductivity, heart activity, and respiration, have all been considered as potential physiological markers for recognizing affective states. Despite high classification rates achieved under laboratory conditions [4], the changing nature of physiological signals introduces significant challenges when one moves from the lab and into the real world. In particular, physiological data is expected to exhibit non-stationarities or day variations due to factors such as: electrode drift, changes in the electrode impedance, and modulations by other mental states such as attention and motivation of subjects [2, 5].

Non-stationarity of the physiological signal indicates that the signal changes its statistical characteristics as a function of time. These changes then propagate to feature values extracted from signals over time. Non-stationarities or day variations of physiological data represents a major problem for building reliable classification models of affect (i.e., generalizability within an individual). This is because classification methods assume that training data is obtained from a stationary distribution. In real world contexts, however, this assumption of stationarity is routinely violated. According to Kuncheva [6], every real-world classification system should be equipped with a mechanism to adapt to changes in the environment. Therefore, this study utilizes an updatable ensemble classification approach (winnow), discussed in more detail in section II.

Understanding the nature of these day variations is essential for developing reliable affect detection systems that can be deployed in real world affective computing applications. There is a critical need for basic research on how physiological signals vary over time before effective solutions can be proposed. In this study we try to address two related research questions. First, the day variation in physiological data might indicate that diagnostic features of affect vary from one day/session to another. We test this issue and evaluate a new classification approach that uses day-specific features for affect detection. The performance of day-specific features is then compared to a classification approach that uses pooled features obtained from pooled day

data. Second, we test the performance and reliability of individual physiological channels for detecting both valence and arousal affective dimensions over the span of multiple day recording sessions.

A. Background and Related Work

Research on emotion referred to the existence of a set of discrete emotional prototypes (e.g. happiness, sadness, etc) [7]. As opposed to the notion or existence of discrete emotions, Russell [8] suggested that affective experience is best described in the two-dimensional space of *valence* and *arousal*. The arousal dimension ranges from highly activated to highly deactivated, and the valence dimension from highly pleasant to highly unpleasant. Take happiness for example as an emotion; it is modelled with positive valence and high arousal. On the other hand, Sadness is modelled with a negative valence and low arousal.

Recent research has utilized physiological signals for affect detection of both affective dimensions of valence and arousal. Kim and Andre [1] conducted an experiment to detect the levels of valence and arousal of subjects during music listening using physiological signals. They recorded ECG, facial EMG, SC, and RSP from three participants. They achieved 89% classification accuracy for 1-2 degrees of valence and 77% for 1-2 degrees of arousal using an LDA classifier. Similarly, Lichtenstein, et al. [4] recorded ECG, SC, EMG, RSP, and skin temperature from 41 subjects while watching emotionally charged films. They were able to detect 1-2 degrees of arousal with 82% accuracy and 72% for 1-2 degrees of valence using an SVM classifier. Likewise, Picard, et al. [2] recorded physiological data over a period of 20 days from one subject (actor), who was asked to self-elicit a set of eight emotions. They faced the problem of degrading classification performance when data from multiple days were combined. They attempted to address the problem of day variation by including day information as additional classification features; however, this did not yield a significant improvement in accuracy.

The above mentioned studies have used traditional batch static classification techniques without any updating mechanism to the classifier. On the other hand; adaptive and updatable classification techniques have been used to handle non-stationarities in two close domains of study; which are speech recognition and Brain Computer Interfaces (BCI). For example, Maier-Hein, et al. [9] implemented an adaptive approach to detect non-audible speech using seven EMG electrodes. They realized that a major problem in surface EMG based speech recognition ensue from repositioning electrodes between recording sessions, environmental temperature changes, and skin tissue properties of the speaker. In order to reduce the impact of these factors, they investigated a variety of signal normalization and model adaptation methods. An average word accuracy of 97.3% was achieved using seven EMG channels and the same electrode positions. The performance dropped to 76.2% after repositioning the electrodes if no normalization or adaptation is performed. However, by applying the adaptation methods they managed to restore the recognition rates to 87.1%.

Adaptive and dynamic classification approaches have also been employed in BCI research. BCI aims at giving the ability to control devices through mere thoughts by utilizing brain signals such as electroencephalogram (EEG). For example Lowne, et al. [10] compared the performance of a dynamic classification approach to a static classifier and an MLP classifier on an online BCI experiment. They used EEG data from eight subjects during a wrist extension exercise; 20% of the data were labelled with true labels (movement, non-movement). The three classifiers were then tested on the EEG data in a sequential manner (timely ordered) to detect one of the two classes. The performance of the dynamic classifier was significantly higher than that of the static classifier and MLP. This shows that classifier adaptation is more effective compared to static classification when dealing with changing data such as physiological data. This gives the rational for the use of an adaptive and updatable classification approach such as the winnow updatable ensemble algorithm to classify our corpus of affective data.

The remaining of this paper is organized as follows. Section II describes the procedure of collecting affective physiological data and the computational methods employed for feature extraction and classification. Section III presents and discusses our results, and section IV provides concluding remarks and directions for future work.

II. Measures, Data and Methods

A. Participants and Measures

Participants were six students from the University of Sydney (five males and one female) between 24 and 39 years of age. Participants were paid for their participation in the study. The physiological sensors used for recording physiological activity from participants were: ECG, SC, EMG, and RSP. The physiological signals were acquired using a BIOPAC MP150 system and AcqKnowledge software with a sampling rate of 1000 Hz for all channels. The ECG signal was collected with two electrodes placed on both wrists. EMG was recorded from the corrugator (eyebrow) and zygomatic (cheek) facial muscles. The SC was recorded from the index and middle fingers of the non-dominant hand, and a respiration belt fixed around the participant chest was used to measure respiration activity.

The affect-inducing stimulus consisted of set of 400 images selected from the International Affective Picture System (IAPS) collection [11]. The images were selected on the basis of their normative valence and arousal scores. The mean valence norm scores ranges from 1.40 to 8.34, and mean arousal norm scores ranges from 1.72 to 7.35 (on a scale from 1 to 9). Images were selected from the four quadrants: PositiveValence-LowArousal (mean IAPS valence norm > 6.03 and mean IAPS arousal norm < 5.47), PositiveValence-HighArousal (mean IAPS valence norm > 6.03 and mean IAPS arousal norm > 5.47),

NegativeValence-HighArousal (mean IAPS valence norm < 3.71 and mean IAPS arousal norm > 5.47), and NegativeValence-LowArousal (mean IAPS valence norm < 3.71 and mean IAPS arousal norm < 5.47). The idea was to select images from the extremes of both valence and arousal in order to maximize the differences of participants' physiological responses. The set of 400 images was then divided into 5 sets of 80 images each (20 images from each category). However for classification procedures described in section III, we classify each of the valence and arousal dimensions separately.

Only four participants were able to complete the five recording sessions. The female participant and one male participant reported their inability to continue because some images had explicit content that were overwhelming to them. Therefore, only data from four participants were used in the current study. This indicates the difficulty of obtaining a affective physiological data that is intended to track variations over multiple sessions. We note that even though the sample size is small, each participant was recorded over 5 sessions. This is consistent with the present goal of tracking variations within an individual rather than across individuals.

B. Procedure

participants sat in a quite dimmed room and were asked to sign a consent form before the start of the session. They then viewed a set of emotionally charged IAPS images, during which their physiological signals were continuously recorded. Each recording session lasted approximately 60 minutes. Each emotional trial consisted of presenting each image for 12 seconds, followed by a screen that showed a 2 X 2 affective grid which lasted few seconds and allowed participants to rank their levels of valence (positive, negative) and arousal (low, high). A blank screen was presented afterwards for 8 seconds to allow physiological activity to return to baseline neutral levels before a new image was presented. Five images were presented consecutively from each category in order to maintain a stable emotional state for that category. This protocol was designed to suit the intended goals of our study and is based on previous research [11]. Each subject participated in five recording sessions each separated by one week. A different set of images were presented for each session in order to prevent habituation effects. The same setup was used in all recording sessions.

TABLE I. THE WINNOW ENSEMBLE ALGORITHM

- Initialization: Given a classifier ensemble D = (D1,….,Dn), Initialize all classifier weights; wi = 1. i =1:n.

- Classification: For a new example x, calculate the support for each class as the sum of the weights of all classifiers Di that suggest class label ck for x. Set x to the class with largest support. k=1:number of classes.

- Updating: if x is classified correctly by classifier Di then its weight is increased (promotion) by wi = alpha * wi,, where alpha > 1. If classifier Di incorrectly classifies x, then its weight is decreased by wi = wi / alpha (demotion).

C. Feature Extraction

The MATLAB-based Augsburg Biosignal Toolbox [12] was used to preprocess and extract features from the raw physiological data. A total of 214 statistical features (e.g. mean, median, standard deviation, maxima and minima) were extracted from the five physiological channels using window size of 12 seconds (the length of the emotional trial). In general, SC responses (SCR) can be observed 1-3 seconds after stimulus presentations. EMG responses are substantially faster, however, the frequency of the muscle activity can be summed up over a period of time to indicate a change in behavioral pattern. ECG and respiration responses are considered slower, however, estimating cardiac and respiratory patterns form short term periods is common in psychophysiology research area [13]. Eighty-four features were extracted from ECG, 21 from SC, 21 from each of the EMG channels, and 67 from the RSP channel. A complete description of these features can be found in [12].

D. Classification Methods

The present study used the winnow updatable ensemble algorithm described in detail in Table. I. It is an ensemble based algorithm that is similar to a weighted majority voting algorithm. It combines decisions from ensemble members based on their weights. However, it utilizes a different updating approach for member classifiers. This includes promoting ensemble members that make correct predictions and demoting those that make incorrect predictions. Updating of the weights is done automatically based on incoming data, which makes this approach suitable for online applications that operate in non-stationary environments .

We set the parameter alpha value to 2 (alpha = 2); which is the parameter used to update the weights of classifier members. Acceptable results have been achieved using an alpha value of two in previous research [6]. In this study we use hard labeled data only. The ensemble relies on regular feedback where weights are updated on the basis of error. This is a prerequisite for the updating mechanism used by winnow. However, in real world applications immediate feedback might not be available all the time. In case that feedback or true class labels are unknown, other approaches could be explored. For example the use of semi-supervised techniques that combine labeled and unlabeled data for training a classifier. Additionally, feedback could be obtained on-demand by asking users of an affect-aware system in away similar to self-reports. In this case the frequency at which the ensemble is updated could depend on some performance threshold.

The WEKA data mining package, and PRTools 4.0 [14], a pattern recognition MATLAB library, were used for classification. Chi-square feature selection was used for dimensionality reduction in order to avoid problems associated with large feature spaces. WEKA's support vector machine (SMO) classifier with a linear kernel was utilized for training classification models. Many successful

applications of SVMs have been demonstrated in previous research [15]. Choosing the SMO classifier as a base classifier is independent from the classification approach adopted by the ensemble algorithm. In future work, additional classifiers could be evaluated to determine their effect on the performance of the ensemble algorithm.

E. Day Datasets

Day datasets were constructed separately for the two affective measures valence (positive/negative) and arousal (low/high). Additionally, separate datasets were constructed for both IAPS mapped categories (as described in section II. A.) and self-reports of participants. In total there were 80 (4 participants x 5 recording sessions x 2 affective measures (valence and arousal) x 2 ratings (IAPS and self-reports)) datasets with 80 instances in each dataset. IAPS ratings datasets had a balanced distribution of labels 40:40 for positive/negative valence or low/high arousal. On the other hand, self-reports datasets had unbalanced distribution of classes, so a down sampling procedure (WEKA's SpreadSubsample which produces a random subsample of a dataset) was applied to obtain a balanced distribution of classes. This is done in order to avoid classifier bias towards predicting majority class. On average, 34% of data was lost for self-reports arousal datasets and 15% of data was lost for self-reports valence datasets. Therefore, baseline classification accuracy is (50%) for both types of data sets. A preliminary analysis showed that the top five ranked features -using chi-square feature selection- were sufficient to produce consistent classification results without sacrificing performance. Therefore, the top five features were selected from each dataset and used in all subsequent analysis.

III. RESULTS AND DISCUSSIONS

We first tested the effectiveness and reliability of the IAPS images of inducing both valence and arousal using Cohen's kappa. We then explored the issue of how diagnostic features of affect change over time by applying feature ranking (chi-square) to separate day datasets, and how this issue could affect the building of reliable classification models. Next, we carried out two classification experiments. The first experiment compared three feature selection and training strategies utilizing the winnow ensemble algorithm and a day-cross validation procedure that resembles baseline accuracy. The effect of using day-specific and pooled features on classification accuracy was tested. The second experiment involved testing the effect of individual physiological channels on affect detection accuracy. This experiment built on the results of the first experiment by using pooled features only for affect classification. In all classifier training strategies described in the following sections, we adopted a day cross-validation strategy in order to test on all available data.

A. The effectivness of IAPS of inducing Affect

In order to test the effectiveness of the IAPS stimuli in inducing both valence and arousal, we tested the level of agreement between participants' self-reports and IAPS normative ratings using Cohen's kappa. Participants' self-reported valence showed higher agreement with IAPS normative ratings compared to arousal. The kappa score for valence was 0.89, and kappa score for arousal was 0.41. It is evident that the IAPS stimuli were quite successful in eliciting valence, but was much less effective in influencing arousal. However, both ratings schemes will be used to assess affect detection accuracy.

B. Day-Specific Features

As an example of how diagnostic features change across days, Table II presents the results of chi-square feature selection applied to participant S1 (applied to each day data separately). The chi-square value represents the degree of relevance of a feature to class category. It can be seen from the results that the diagnostic features are different for each day. This is a reflection of the changing nature of physiological data. Table III presents the features selected from one participant using IAPS ratings only (for space limitations and as a demonstration), however other subjects data showed similar behavior. An interesting observation is that there are frequent features which reoccur on different days. This is promising as it allows for easier calibration of affect detection classification models. However this leaves us wondering whether classification models that are built from these day-specific features are more accurate than those built using pooled features selected from pooled day data. We address this issue in the next section.

C. Winnow Results on Day-Specific and Pooled Features

The winnow ensemble algorithm was run with an ensemble of four base classifiers each trained on a separate day-specific features dataset. Testing was done on the remaining day-data. The procedure was repeated five times to test on all available data. We also tested the performance of pooled

TABLE II. TOP FIVE SELECTED FEATURES PERFORMED ON DAY DATA SEPRATELY FOR PARTICIPANT S1 WITH VALENCE (IAPS) AS CLASS LABEL

Chi Square/ Feature Name				
Day 1 Features	***Day 2 Features***	***Day 3 Features***	***Day 4 Features***	***Day 5 Features***
43 SC-2Diff-minRatio	23 ZYG-EMG-1Diff-minRatio	10 SC-1Diff-minRatio	16 ZYG-EMG-max	28 ZYG-EMG-2Diff-minRatio
43 SC-2Diff-maxRatio	23 ZYG-EMG-1Diff-maxRatio	10 RSP-Ampl-1Diff-max	10 ECG-QS-min	26 SC-2Diff-maxRatio
23 ZYG-EMG-1Diff-maxRatio	14 RSP-2Diff-range	10 SC-1Diff-maxRatio	9 ECG-QS-range	24 SC-2Diff-minRatio
23 ZYG-EMG-1Diff-minRatio	13 RSP-2Diff-min	6 RSP-Ampl2Diff-maxRatio	8 ECG-HrvDistr-mean	23 ZYG-EMG-2Diff-maxRatio
23 RSP-Pulse-max	12 ZYG-EMG-2Diff-mean	5 ECG-HrvDistr-mean	6 RSP-Pulse1Diff-maxRatio	12 RSP-Ampl-mean

[a.] ZYG: Zygomatic facial muscle, Amp: Amplitude, min: Minimum, max, Maximum, HRV: Heart rate variability, 1Diff: First Difference., 2Diff: Second Difference.

TABLE III. AVERAGE CLASSIFICATION ACCURACY FOR THREE TRAINING STRATEGIES USING MIXED FEATURES FROM ALL PHYSIOLOGICAL CHANNELS (%)

Subject ID	D-CV	W-PF	W-SF	D-CV	W-PF	W-SF
	Valence (IAPS)			*Arousal (IAPS)*		
S1	59	74	64	52	72	56
S2	54	63	61	51	69	53
S3	50	74	61	52	75	49
S4	50	76	54	48	73	59
Avergae	53.25	71.75	60.00	50.75	72.25	54.25
	Valence (Self)			*Arousal (Self)*		
S1	52	73	68	50	63	62
S2	53	62	59	49	65	48
S3	51	79	64	51	76	60
S4	50	70	54	52	69	69
Average	51.50	71.00	61.25	50.50	68.25	59.75

[b.] D-CV Day Cross-Validation, W-PF: Winnow with Pooled Features, W-SF: Winnow with Day-Specific Features

features, which are features selected from four days data combined, using the same training procedure. As a baseline, we used a day cross-validation procedure, where a single SMO classification model was constructed from pooled data of four days, and testing was done on the remaining day data. This process was repeated five times in order to test on all available data. The baseline procedure represents a static classification approach without an update mechanism. This process was repeated for the four categories (Valence-IAPS , Arousal-IAPS , Valence-self, and Arousal-self). The results in Table III were obtained from a mixed feature set, in which the top five features were selected from all physiological channels.

Results showed that accuracy scores using day-specific and pooled features are higher than day cross-validation baseline accuracy. This indicates that we were able to leverage the dynamicity of winnow algorithm to enhance classification accuracy for this type of data. It can also be seen from Table III that winnow with pooled features on average outperformed winnow with day-specific features. We were expecting that day specific features could provide higher performance. The explanation to this lower performance of day-specific features could be due to the fact that day data tends to have higher clustering cohesion and tightness compared to data for the same emotion category across multiple days. In order to test this effect further, in the next section we present classification results using single channel's data using the same procedures described earlier for day cross-validation, pooled features, and day-specific features. This also will allow us to shed light on the performance and reliability of these individual channels for affect detection over multiple sessions.

D. Channels Effect on Affect Detection Accuracy

The same training procedures described above were performed on each individual physiological channel's data. Results are not shown here but will be outlined in the Analysis of Variance (ANOVA) analysis described next.

In order to examine the effect of three training strategies on affect detection accuracy, an ANOVA was conducted on all accuracy scores combined. This is a one-way repeated measure ANOVA with accuracy as the dependent variable and training strategy as the independent variable. A significant main effect was found for training strategy ($F(2, 285) = 57.67$, $p < 0.05$). Bonferroni posthoc tests revealed that accuracy scores for winnow with pooled features ($M = 65.14$) were higher than those for winnow with day-specific features ($M = 57.81$) and day cross-validation ($M = 55.55$). These results suggest that using winnow with pooled features is more suitable for building predictive models of affect than the other two training strategies. Pooled features showed that it had the capacity to describe someone's overall affective states with significantly higher accuracy compared to the more discrete day specific-features. On the other hand, using winnow ensemble achieved what was expected by outperforming the single model approach represented by the day cross-validation approach.

We tested the effect of both physiological channel and emotion on affect detection accuracy using two-way repeated measures ANOVA. This analysis was done using winnow accuracy scores with pooled features only. We found significant main effect for channel ($F(5, 72) = 12.60$, $p < 0.05$), indicating that physiological channels vary in their usefulness for affect detection. Posthoc tests revealed that accuracy scores for EMG-cur ($M = 69.87$) and mixed features ($M = 72.25$) were significantly higher than other channels ECG ($M = 62.37$), EMG-zyg ($M = 58.5$), RSP ($M = 61.62$), and SC ($M = 61.62$). We did not find significant effect for emotion category ($F(3, 72) = 1.14$, $p = 0.34$), which indicates that there are no significant differences in the accuracy at which valence and arousal are detected given a particular channel. However, when the levels of emotional categories were decreased to two rather than four categories, the effect of emotion was only marginally insignificant ($F(1, 84) = 3.5$, $p = 0.065$). A significant effect for the interaction between channel and emotion was found ($F(5, 84) = 2.36$, $p < 0.05$). This indicates that some channels have stronger influence on one of the two affective components (valence, arousal) over the other.

The interaction effect was further explored by conducting simple effects tests. The tests revealed that both EMG channels were more useful for detecting valence than arousal; EMG-cur ($F(1, 84) = 4$, $p < 0.05$), and EMG-zyg ($F(1, 84) = 10.44$, $p < 0.05$). Other channels where equally likely to detect both valence and arousal with the same accuracy, ECG ($F(1, 84) = 43$, $p = 0.62$), mixed features ($F(1, 84) = 70$, $p = 0.15$), RSP ($F(1, 84) = 0.07$, $p = 0.79$), and SC ($F(1, 84) = 0$, $p = 0.99$). Fig. 1 shows the interaction effect between channel and emotion. Our findings came in accordance with literature in regard to both EMG channels. The corrugator and zygomatic EMG have always shown consistent changes with the valence component of emotion [16]. On the other hand, previous research has always considered SC as an index of arousal [17]. Taking this in regard, our results might have been affected by using IAPS images as stimulus.

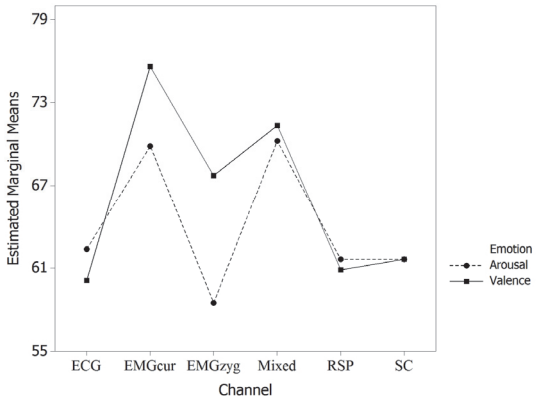

Fig. 1. Interaction effect between channel and emotion

The results also showed that detecting arousal with acceptable accuracy required more physiological markers in comparison to valence (see Fig. 1). This probably the reason that some previous research has outlined that the detection of arousal is harder than valence [1]. The literature is somehow inconsistent in this regard, with studies reporting higher detection rates for arousal than valence [4] and the contrary [1]. However, an interesting study conducted by Gomez, et al. [18] found that induced physiological changes of subjects' valence lasted longer than those of arousal in which they dissipate quickly. This might explain the higher detection rates of valence compared to arousal. However, it should also be noted that other researchers believe that valence detection can be more difficult to detect compared to arousal as valence information is conveyed more subtly [19].

IV. Conclusions, Limitations and Future Work

We have shown that diagnostic physiological features of affect exhibit day variations. This is a challenging issue for building effective affect detection systems. Using day-specific features did not yield improved affect detection over that of using pooled feature set. Both facial EMGs were more predictive of valence than arousal compared to ECG, RSP and SC. This has implications if designers of affect detection systems were more interested in detecting valence than arousal. This also suggests that facial EMG is more reliable than other measures when considering affect detection over multiple sessions. Additionally, EMG-cur and a fusion of features from all channels yielded the highest detection rates for both valence and arousal. There are two primary limitations with the present study. One limitation of our work is the relatively small sample size, so replication with a larger sample is warranted. The second limitation is that emotions were artificially induced rather than spontaneously experienced. This approach was adopted because strict laboratory control was desired in the present experiment. Replicating this research in more naturalistic contexts is an important step for future work.

Acknowledgment

Omar AlZoubi and Davide Fossati are supported by award NPRP 5-939-1-115 from the Qatar National Research Fund. Sidney D'Mello is supported by NSF grants (HCC 0834847, and DRL 1235958).

References

[1] J. Kim and E. Andre, "Emotion Recognition Based on Physiological Changes in Music Listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, pp. 2067-2083, 2008.

[2] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, pp. 1175-1191, 2001.

[3] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting Naturalistic Expressions of Nonbasic Affect using Physiological Signals," IEEE Transactions on Affective Computing, vol. 3, pp. 298-310, 2012.

[4] A. Lichtenstein, A. Oehme, S. Kupschick, and T. Jürgensohn, "Comparing Two Emotion Models for Deriving Affective States from Physiological Data," in Affect and Emotion in Human-Computer Interaction. vol. 4868, C. Peter and R. Beale, Eds., ed: Springer Berlin / Heidelberg, 2008, pp. 35-50.

[5] O. Alzoubi, M. S. Hussain, S. D'Mello, and R. A. Calvo, "Affective modeling from multichannel physiology: analysis of day differences," in Proceedings of the 4th international conference on Affective computing and intelligent interaction-Volume Part I, 2011, pp. 4-13.

[6] L. I. Kuncheva, "Classifier Ensembles for Changing Environments," in Multiple Classifier Systems, ed, 2004, pp. 1-15.

[7] P. Ekman, "An argument for basic emotions," Cognition & Emotion, vol. 6, pp. 169-200, 1992.

[8] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, pp. 1161-1178, 1980.

[9] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on, 2005, pp. 331-336.

[10] D. R. Lowne, S. J. Roberts, and R. Garnett, "Sequential non-stationary dynamic classification with sparse feedback," Pattern Recognition, vol. 43, pp. 897-905, Mar 2010.

[11] M. Bradley and P. J. Lang, "The international affective picture system (iaps) in the study of emotion and attention," in Handbook of Emotion Elicitation and Assessment, J. A. Coan and J. J. B. Allen, Eds., ed New York:: Oxford University Press, 2007, pp. 29-46.

[12] J. Wagner. (October, 2009). Augsburg Biosignal Toolbox (AuBT). Available: http://mm-werkstatt.informatik.uni-augsburg.de/project_details.php?id=%2033

[13] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," Biological Psychology, vol. 84, pp. 394-421, 2010.

[14] F. v. d. Heijden, R. P. Duin, D. d. Ridder, and D. M. Tax, Classification, parameter estimation and state estimation - an engineering approach using Matlab. Chichester: John Wiley & Sons, 2004.

[15] A. K. Jain, R. P. W. Duin, and M. Jianchang, "Statistical pattern recognition: a review," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, pp. 4-37, 2000.

[16] A. O. Hamm, H. T. Schupp, and A. I. Weike, "Motivational organization of emotions: Autonomic changes, cortical responses, and reflex modulation," Handbook of affective sciences, pp. 187-211, 2003.

[17] R. W. Levenson, "Autonomic Nervous System Differences among Emotions," Psychological Science, vol. 3, pp. 23-27, 1992.

[18] P. Gomez, P. G. Zimmermann, S. Guttormsen Schär, and B. Danuser, "Valence lasts longer than arousal: Persistence of induced moods as assessed by psychophysiological measures," Journal of Psychophysiology, vol. 23, pp. 7-17, 2009.

[19] R. W. Picard, Affective Computing, second ed. Cambridge, Massachusetts: The MIT Press, 1997.