MOBILE REAL-TIME AROUSAL DETECTION

Vasileios Alexandratos ⋆ *Murtaza Bulut*† *Radu Jasinschi*†

⋆ Department of Embedded Systems, TU Delft, Delft, the Netherlands
vassilios.alexandratos@gmail.com
†Department of Smart Sensing & Analysis, Philips Research, Eindhoven, The Netherlands
murtaza.bulut@philips.com, radu.jasinschi@philips.com

ABSTRACT

We introduce a mobile system that is able to detect arousal in real-time based on electrocardiogram and electrodermal activity. The system is using an Android smartphone and wearable sensors, which include a smart watch and a heart rate belt that gather skin conductance and heart rate data, respectively. Algorithms for processing the skin conductance and heart rate data, as well as an automated method for labeling the collected 'arousal' and 'non-arousal' experimental data are developed. Small-scale user tests show 84% 10-fold, 83% between-subject, and 68% new-subject arousal detection accuracy.

Index Terms— arousal, skin conductance, heart rate variability.

1. INTRODUCTION

For a healthy lifestyle it is important that people are aware of arousing and potentially stressful situations, so they can take the necessary actions to cope with them. In this paper, we describe a smartphone-based system which detects arousal in real-time using Electrocardiogram (ECG) and Electrodermal Activity (EDA) data, which consists of R-R peaks and skin conductance, respectively. Algorithms for automatic labeling of the collected data, and for real-time feature extraction from heart rate and skin conductance are presented.

The system uses the Bluetooth communication capability of the smartphone to transfer the arousal and non-arousal classification results to the nearby connected devices. Informing selected contacts of the user in real-time about the current arousal state of the user can be valuable for applications targeting autistic children, elderly with dementia and their caregivers [1].

Arousal or stress detection based on physiological signals has been explored in a multitude of studies, using physiological parameters such as ECG, EDA, skin temperature (ST) and pupil diameter in tasks such as the Stroop color-word interference and the Montreal Imaging Stress Task (MIST). The achieved stress vs. no-stress classification performances range from 82.8% to 99.5% [2–5]. In these studies, all analyses were done offline and the built stress detection systems did not have any real-time capabilities.

A real-time mobile stress recognition system that uses ECG data is presented in [6]. In this system, the stress recognition is based on a personal stress threshold established for each subject in a lab setting using a 'protocol intended to alternately evoke sympathetic and parasympathetic responses'. When stress is detected, the application prompts the user to perform certain breathing exercises to alleviate stress. This study is mainly a 'qualitative exploration of how people adopt mobile therapies', and no formal evaluation of the efficacy of the system is available.

In the real-time mobile stress detection system presented in [7], the stress detection is based on activity information, ECG, ST and breathing rate. The system is implemented on a smartphone and can discriminate between five stress levels (from no stress to very high stress). This system is trained with fifteen subjects, and its offline classification accuracy reaches 90.4%. The online classification accuracy reaches only 39.7%. The reason for this big drop in performance was identified as overfitting of the input data.

The systems in [6, 7] are similar to ours in terms of their real-time functionality, but they do not consider correcting the errors in the heart beat signal. In addition, for the reduction of the interpersonal variability in the recorded bio-signals, these studies rely on subject-specific thresholds, while in our system a subject-independent normalization method is used instead.

2. AROUSAL DETECTION SYSTEM

The main units of the arousal detection system are a smartphone and two sensors that measure EDA and ECG activity. As a smart phone we used the Google Nexus 4 (1.5 GHz CPU, 2 GB RAM), running Android v4.2 (Jelly Bean). The EDA data is obtained from a wrist-worn Philips DTI-2 watch [8], which can stream (via Bluetooth) the raw skin conductance data in real-time. The ECG data is obtained using the Zephyr HxM BT heart rate measurement chest belt. The R-R peaks of the ECG are detected automatically by the device and only the timestamps of the peaks are transmitted (via Bluetooth).

The software of the system, running on the smartphone, consists of two parts which are responsible for arousal detection and data storing. The part responsible for the online arousal detection (i) receives the sensory data, (ii) extracts the parameters relevant to the arousal detection from it (as soon as one minute of data has been collected), (iii) identifies the arousal of the user based on these parameters, (iv) notifies the user about his/her arousal state, (v) transmits this information to the connected device(s) and (vi) discards the oldest 30 seconds of data (this way, a sliding window with 50% overlap is realized). The part responsible for storing the arousal data for offline reviewing (i) stores all sensory data to the phone's memory, (ii) records and stores audio clips relevant to the arousal response, and (iii) uploads all arousal-related data to a personal account in Dropbox. A flowchart showing the functionality of the arousal detection system is shown in Fig. 1.

The system uses the smartphone's microphone to store audio data relevant to the arousing stimuli or event. Audio clips are

This work was conducted at Philips Research as a part of Vasileios Alexandratos' Master's Thesis at TU Delft.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

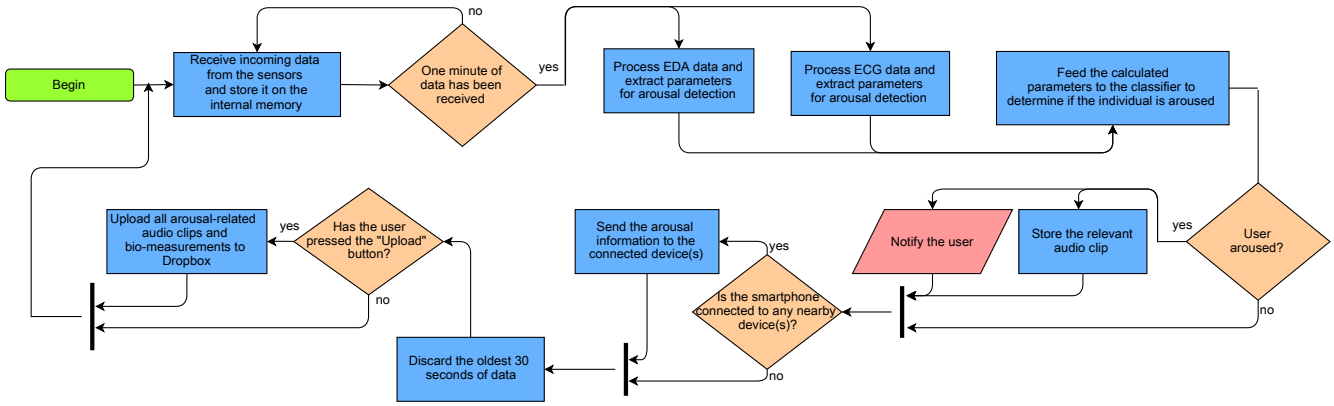

Fig. 1. Flowchart representing the main functionalities of the arousal detection system.

recorded continuously and are stored in memory using a 'circular buffer' approach (i.e. older audio data is overwritten by newer data), and whenever arousal is detected, the corresponding audio clip is marked as 'relevant' and is stored along with its timestamp.

2.1. Skin conductance and heart rate feature extraction and normalization

The first step for processing of the raw skin conductance (SC) signal includes a five-point moving average filtering. The developed peak-detection algorithm, which works in two phases, is then applied to the smoothed signal. First, it detects all local minima (onsets) using the first derivative test. Using this test, a local minimum is identified when the first derivative of the SC signal switches from negative to positive. Second, it searches for the corresponding peak of every onset identified in the first phase. For this, starting from every onset, the algorithm uses the first derivative test to find the closest following local maximum. Using this test, a local maximum is identified when the first derivative of the SC signal switches from positive to negative. The maximum duration of a typical skin conductance response (SCR) is ten seconds [9], so the algorithm searches for a corresponding peak only in a ten-second interval following an onset. Using this algorithm, the peak detection accuracy is approximately 90%, compared to visual inspection.

The features calculated from each one-minute SC frame are the mean skin conductance level (SCL), the standard deviation of the SC signal (STD), the slope of the least squares regression line fitted to the raw SC signal (Slope), the mean amplitude of all peaks in the frame (Mean Amplitude) and the mean rise time of all peaks in the frame (Mean Rise Time).

The Heart Rate Variability (HRV) calculations involve an automatic Inter-Beat Interval (IBI) signal correction before the feature extraction, since beat mis-detection on an ECG has a big impact on the HRV analysis [10]. We implemented a version of the algorithm described in [11], which not only detects errors in the IBI signal, but also applies the appropriate corrections. An error in an IBI value is detected via an adaptive threshold calculated from the five preceding IBI values. Whenever an error is detected on an IBI value, the algorithm corrects it using a set of rules, which emulate how a human expert would manually correct the IBI data, e.g. splitting the erroneous IBI value in two, in cases of a missed heartbeat, and adding two erroneous IBIs together, in cases of a falsely identified heartbeat.

The Power Spectral Density (PSD) of the corrected IBI signal (re-sampled at 4 Hz using linear interpolation) is computed using the Welch method. The HRV features are then calculated from the PSD signal. These features are the power in Low Frequency (LF, 0.04-0.15 Hz), the power in High Frequency (HF, 0.15-0.4 Hz), the normalized LF (LFnorm) power, the normalized HF (HFnorm) power and LF/HF ratio. LFnorm and HFnorm are calculated as LF/(LF+HF) and HF/(LF+HF), respectively, as proposed in [12]. The HRV time-domain features include the mean IBI, SDNN (standard deviation of the N-N intervals), RMSSD (square root of the mean squared differences of successive N-N intervals), and pNN50 (ratio of the number of pairs of successive N-N intervals that differ by more than 50 ms divided by total number of N-N intervals).

Both heart rate and skin conductance exhibit large interpersonal variations. In related studies, the method to minimize these variations includes normalization of the features calculated during stress using the corresponding features during baseline for each subject, or the scaling of the signal amplitudes from each participant between zero and one. However, these methods are not applicable to an online stress detection system, such as ours, which has to deal with unseen physiological data. For the minimization of the interpersonal variations in our system, the normalization method is subject-independent and is performed as follows: CoV(SC)=$\frac{(STD)}{(MeanSCL)}$, MeanAmplNorm=$\frac{(MeanAmplitude)}{(MeanSCL)}$, CoV(NN)=$\frac{(SDNN)}{(MeanIBI)}$, RMSSDNorm=$\frac{(RMSSD)}{(MeanIBI)}$. The remaining features, Number of Peaks, Slope, Mean Rise Time (sec.), pNN50, HFnorm, LFnorm and LF/HF are used as they are without any further normalization.

3. DATA COLLECTION

The training data is collected using the MIST procedure [13], which is commonly employed for inducing stress/arousal. The data collection experiment consisted of three conditions: baseline, control and arousal, which were completed in a random order with one minute pause between them.

During the baseline condition, the participants were asked to remain sitting still and silent in an upright position for approximately three minutes. During the control condition, the participants were asked to read aloud a text passage (a children's story) written on paper for approximately five minutes. For the arousal condition, the mathematical subtraction test described in [14] was used. This test

	Classification Algorithm		
	SVM	J48	RF
10-fold CV	65%	67%	70%
LOSO CV	56%	58%	61%

Table 1. Initial classification results obtained using the dataset collected from the twelve subjects.

required participants to continuously subtract 13 starting from 1079 at their maximum speed while uttering the subtraction procedure (i.e. 1079 minus 13 is 1066, 1066 minus 13 is 1053, and so on). If they made an error, the experimenter notified them to correct themselves and then continue with the subtraction. The duration of the arousal condition was approximately five minutes (or less for participants that reached zero in a shorter time). After the final condition, participants were asked how they felt, but no formal (i.e. questionnaire-based) evaluation of their affective state was performed.

Sixteen (twelve males) healthy volunteers, between 23 and 39 years of age (mean age 29.6 $\pm$ 6.3), participated in this first round of experiments. All participants had an engineering background and were comfortable with mathematics. The tests were carried out in a laboratory environment, using the equipment described in Section 2. The participants were in comfortable sitting positions during the experiment and the experimenter was always present in the test room. The experimenter only interfered during the arousal session if participants made subtraction errors but not during the other two sessions.

For four of the participants, the skin conductance data was not recorded due to technical reasons, so all their data was excluded from the analysis. For the remaining twelve subjects (ten male), the collected dataset size was 299 frames, where each frame is one minute long and overlaps 50% with the preceding frame. Given the varying duration of the experiment for each subject, 'arousal' (i.e. data from the arousal session) and 'non-arousal' (i.e. data from the baseline and control conditions) datasets were not of equal size. In particular, 59.8% of the collected data was from the non-arousal sessions.

4. CLASSIFICATION RESULTS

The most widely used algorithms for similar arousal/stress detection tasks include the Support Vector Machines (SVM) and the J48 Decision Tree. These single-classifier algorithms were used, as well as one classifier ensemble, the Random Forest (RF), which offers good prediction performance. All classifiers are implemented using the WEKA software [15] and the classification accuracy is measured in terms of the 10-fold and the Leave-One-Subject-Out (LOSO) Cross Validation (CV). Table 1 summarizes the arousal vs. non-arousal classification accuracy obtained using the data collected from the twelve subjects. The observed initial classification accuracy is low compared to the related studies. This is mainly due to confusion between the control and arousal conditions, as can be seen from the 10-fold CV results reported in Table 2.

4.1. Automatic data labeling

To deal with the fact that control and arousal conditions were somewhat similar in terms of their arousing effects in the collected data, only selected data from these sessions are used. To determine which segments of data (i.e. frames of one min. length, with an overlap of 50%) can be used for training the classification algorithm, we developed a threshold-based automatic data labeling method.

	Classification Algorithm		
	SVM	J48	RF
Baseline vs Control	79%	77%	78%
Baseline vs Arousal	83%	76%	81%
Control vs Arousal	68%	58%	69%

Table 2. Baseline vs Control, Baseline vs Arousal and Control vs Arousal models evaluation results.

	Classification Algorithm		
	SVM	J48	RF
10-fold CV	80%	73%	72%
LOSO CV	67%	56%	56%

Table 3. Classification results using the re-labeled input dataset containing data from all twelve subjects.

The method developed for labeling the frames is based on the LF/HF ratio. This feature is selected because of its reliability for stress detection: it significantly increases when individuals experience stress [16]. The first step of this method is to calculate the median value of the LF/HF ratio in each subject's data for all experimental conditions combined, to get a subject-specific threshold value. The second step involves the comparison of the LF/HF ratio in every frame with the calculated threshold. Based on this comparison, frames recorded during the baseline and control conditions with an LF/HF ratio larger than the threshold are discarded. Similarly, the frames recorded during the arousing condition are discarded if their LF/HF ratio is lower than the calculated threshold.

After applying the described labeling method, the input dataset length decreased from 299 to 170 frames, of which 58.2% corresponds to 'non-arousal' frames. Since the thresholding method was based on the LF/HF ratio, this feature was excluded from the feature vector, along with all other features correlated (defined as absolute correlation equal to or larger than 0.4) with it. The Pearson correlation coefficient (PCC) was used to calculate the correlations, and the PCC value between the LF/HF ratio and HFnorm, LFnorm, Number of Peaks, CoV(NN), pNN50, Mean Rise Time, RMSSDNorm, Slope, MeanAmplNorm, CoV(SC) was found to be -0.70, +0.70, +0.32, +0.31, -0.27, +0.23, -0.17, +0.05, -0.04 and +0.03, respectively. Since LF/HF ratio was correlated (abs(PCC)>0.4) with the LFnorm and HFnorm, they were excluded from the features vector. With these exclusions, the remaining features are three ECG (CoV(NN), pNN50, RMSSDNorm) and five EDA (Number of Peaks, Mean Rise Time, Slope, MeanAmplNorm, CoV(SC)) features.

The classification results obtained using the reduced feature vector are shown in Table 3. The classification accuracy increased compared to the initial results (Table 1), but the LOSO accuracy is still low, because for three of the subjects the arousal detection accuracy was below the chance level (44%, 44%, 41%). To make the arousal classifier more specific and stable, the data from these subjects was discarded. Table 4 summarizes the classification results obtained using the dataset from the nine remaining subjects (seven males).

Based on the final tests, the classification algorithm implemented on the developed arousal detection system is the SVM, trained with the three ECG (CoV(NN), pNN50, RMSSDNorm) and

	Classification Algorithm		
	SVM	J48	RF
10-fold CV	84%	76%	80%
LOSO CV	83%	63%	70%

Table 4. Classification results using the re-labeled input dataset containing data from nine subjects.

five EDA (Number of Peaks, Mean Rise Time, Slope, MeanAmplNorm, CoV(SC)) features from nine subjects.

Furthermore, the classification accuracy that can be achieved using only one of the two bio-signals was explored. The 10-fold classification accuracy obtained with the SVM algorithm using only the ECG features was 68%, while the accuracy using only the EDA features was 80%.

4.2. Verification tests

Four new subjects (all male) that did not participate in the previous tests and were naïve to the system and the experiment participated in the verification tests. These subjects were similar to the first set of subjects, in the sense that both groups consisted of participants in age range from 25 to 40 years, had an engineering background and were comfortable with mathematics.

The verification experiments were performed using the system trained using the data and features collected from nine subjects, as described earlier. The test procedure was similar to the first tests, with the differences that for the baseline session subjects were asked to examine four objects (a pen, the two sensors, and the smartphone) while they were seated, silent and still, and that after each session they were asked to complete two questionnaires, the GVA (visual analogue scale for Global Vigor (GV) and Global Affect (GA) [17]) and the SACL (Stress Arousal Checklist) [18]. Note that the subjects were able to see their responses from the previous sessions.

In the GVA questionnaire, participants are asked to rate emotions on a continuous scale. Ten emotional labels are included: alert, sad, tense, effort to do something, happy, weary, calm, and sleepy. In the SACL test, participants are asked to rate 20 emotions on a five-point scale (from strongly agree to strongly disagree). Ten emotions related to stress and ten emotions related to arousal are included. These emotions are: calm, contented, active, vigorous, comfortable, lively, uneasy, tired, sleepy, worried, distressed, uptight, drowsy, tense, relaxed, passive, energetic, alert, bothered, and aroused.

The arousal vs. non-arousal classification results given by the system (arousal alert percentage) and the subjective GV, GA, stress and arousal values, derived from the questionnaire data, are shown in Fig. 2. The arousal alert percentage was calculated by counting the number of frames detected as arousing in each session. The results show that the arousal alert percentage that the system detects correlates well with the subjectively evaluated arousal by the users. In addition, we see that the stress increases and global affect decreases during subtraction, indicating that the valence of the subtraction session is negative. Note that there is a high between-subject variability as can be observed from the high standard errors. This is mainly due to the low number of subjects but also due to the type of the conducted experiments. It is expected that for a task that is more arousing and stressful than the subtraction this variability will be less.

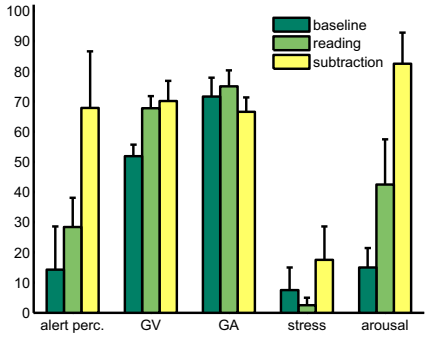

Fig. 2. Results obtained from the verification tests. Bars and error bars represent mean values and standard errors, respectively.

5. DISCUSSION

The main limitation of the current study is the lack of reference arousal information. All data collected during subtraction session was assumed to be arousing, and the rest (i.e. data from the baseline and control sessions) non-arousing. Clearly, this is a rough assumption, as the affective state of participants can be different than expected and can vary during each session. The method used to induce stress was based on a commonly employed method of performing challenging mental arithmetic tasks [14], but this method may have been less effective on our participants since they had engineering background. Informal interviews conducted after the experiment indeed showed that the subtraction task was rated from very challenging to less challenging. In order to account for this variation, a subject-specific threshold calculated from physiological data, as explained in Section 4.1, was used to select more representative arousal and non-arousal data; also, data from three subjects classified below chance level were discarded. Verification tests (Fig. 2) showed that the system produced an arousal alert for baseline, control, and arousing conditions in 14%, 28% and 68% of the frames, respectively.

Clearly there is room for significant improvement. As a first step, more data representative of arousal and non-arousal should be collected. Ideally, these should be substantiated by quantitative (electroencephalography (EEG), speech or face-based emotion analysis) or subjective (questionnaires, external observers) means. Second, there is need to test the system with more people.

The proposed system has novel features such as informing selected contacts about the current affective state of the user, storing of the corrected physiological data, and storing of an audio clip relevant to the arousal stimuli or event for an offline review. Exploring how these functionalities can be useful is a work in progress. We believe they can be beneficial for assisting autistic children and demented elderly, and their caregivers.

6. CONCLUSION

A mobile system that identifies arousal, and potentially stress, in real-time based on physiological data was built and tested. Algorithms for processing skin conductance and heart rate data, as well as an automated method for labeling 'arousal' and 'non-arousal' data were developed. The presented system can be used in applications related to stress coping and lifestyle improvement.

7. REFERENCES

[1] L. Bellodi, R. Jasinschi, G. De Haan, and M. Bulut, "Dialogue support for memory impaired people," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific*, 2012.

[2] F. T. Sun, C. Kuo, H. T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *Mobile Computing, Applications, and Services*, pp. 211–230. Springer, 2012.

[3] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 1355–1358, 2006.

[4] A. de Santos Sierra, C. Sánchez Ávila, J. G. Casanova, and G. B. del Pozo, "A stress-detection system based on physiological signals and fuzzy logic," *Industrial Electronics, IEEE Transactions on*, vol. 58, no. 10, pp. 4857–4865, 2011.

[5] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 410–417, 2010.

[6] M. Morris and F. Guilak, "Mobile heart health: Project highlight," *Pervasive Computing, IEEE*, vol. 8, no. 2, pp. 57–61, 2009.

[7] K. Frank, P. Robertson, M. Gross, and K. Wiesner, "Sensor-based identification of human stress levels," *Respiration*, vol. 10, no. 11, pp. 13, 2013.

[8] J. H. D. M. Westerink, M. Ouwerkerk, G. J. de Vries, S. de Waele, J. van den Eerenbeemd, and M. van Boven, "Emotion measurement platform for daily life situations," in *Proc. Third Intl Conf. Affective Computing and Intelligent Interaction and Workshop*, pp. 217–223, 2009.

[9] M. E. Dawson, A. M. Schell, and D. L. Filion, "The Electrodermal System," *Handbook of psychophysiology*, p. 159, 2007.

[10] G. G. Berntson and J. R. Stowell, "ECG artifacts and heart period variability: Don't miss a beat!," *Psychophysiology*, vol. 35, no. 1, pp. 127–132, 1998.

[11] J. Rand, A. Hoover, S. Fishel, J. Moss, J. Pappas, and E. Muth, "Real-time correction of heart interbeat intervals," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 5, pp. 946–950, 2007.

[12] A.J. Camm, M. Malik, J.T. Bigger, G. Breithardt, S. Cerutti, R.J. Cohen, P. Coumel, E.L. Fallen, H.L. Kennedy, R.E. Kleiger, et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[13] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, and J. C. Pruessner, "The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," *Journal of Psychiatry and Neuroscience*, vol. 30, no. 5, pp. 319, 2005.

[14] J. Ogorevc, A. Podlesek, G. Gersak, and J. Drnovsek, "The effect of mental stress on psychophysiological parameters," in *Medical Measurements and Applications Proceedings (MeMeA), IEEE International Workshop on*, pp. 294–299, 2011.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[16] N. Hjortskov, D. Rissen, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work," *European journal of applied physiology*, vol. 92, no. 1-2, pp. 84–89, 2004.

[17] T. H. Monk, "A visual analogue scale technique to measure global vigor and affect," *Psychiatry Research*, vol. 27, no. 1, pp. 89–99, 1989.

[18] C. Mackay, T. Cox, G. Burrows, and T. Lazzerini, "An inventory for the measurement of self-reported stress and arousal," *British Journal of Social & Clinical Psychology*, 1978.