Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to Enhance Universal Access

Christian Stickel[1], Martin Ebner[1], Silke Steinbach-Nordmann[2], Gig Searle[3], and Andreas Holzinger[3]

[1] Social Learning/Computing and Information Services, Graz University of Technology, Steyrergasse 30/I, A-8010 Graz, Austria
[2] Fraunhofer Institute for Experimental Software Engineering (IESE) 67663 Kaiserslautern, Germany
[3] Institute of Medical Informatics, Statistics and Documentation, Research Unit HCI4MED Medical University Graz, Auenbruggerplatz 2/5, A-8036 Graz, Austria
martin.ebner@tugraz.at, stickel@tugraz.at, Silke.Steinbach-Nordmann@iese.fraunhofer.de, gig.searle@meduni-graz.at, andreas.holzinger@meduni-graz.at

Abstract. Emotion is an important mental and physiological state, influencing cognition, perception, learning, communication, decision making, etc. It is considered as a definitive important aspect of user experience (UX), although at least well developed and most of all lacking experimental evidence. This paper deals with an application for emotion detection in usability testing of software. It describes the approach to utilize the valence arousal space for emotion modeling in a formal experiment. Our study revealed correlations between low performance and negative emotional states. Reliable emotion detection in usability tests will help to prevent negative emotions and attitudes in the final products. This can be a great advantage to enhance Universal Access.

Keywords: Biological Rapid Usability Testing, Valence, Arousal, Emotion.

1 Introduction

The principle of Universal Access [1] extends the definition of users to include people who would otherwise be excluded from information society by rapidly changing technology, e.g. the elderly and ageing [2]. Barriers are mostly not only physical but also mental [3]. The skill of adaptation decreases with age and slowly every new perception, which cannot be solved, is likely to create negative emotions, which in turn hinder the learning process and the motivation to adjust/change. However, mental barriers are not only a problem of the elderly but also for every non-expert end user.

The adaptation of a systems action to enable response to, even influence of, human emotions is a valuable goal for universal access. Nevertheless, if emotion detection is used as a usability tool in the software engineering lifecycle, major stressful issues could be eliminated before they even occur. However, the successful detection of

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2009, LNCS 5614, pp. 615–624, 2009.

emotions is not trivial and the question arises how to implement this in existing methods of usability engineering and how to interpret the changing of emotional states during a test, with regard to all possible influencing factors. An example from experimental psychology [4] shows that an unnoticed word previously associated with shock produces a galvanic skin response, even while subjects fail to notice its occurrence. So far, the measurement of minimal psycho physiological parameters during a users' test is sufficient to show changes in the autonomic nervous system (ANS). The use of biometrics in a usability context has been demonstrated by several previous work, e.g. [5], [6], [7], [8]. To meet the pragmatic and hedonic goals, which are related to the overall experience and results of human–computer interaction, we speak about User Experience (UX). [1] define this term as "A consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the specific context (and/or the environment) within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use, etc.)." Fun, or enjoyment, is a significantly important aspect of user experience with influence on overall satisfaction [10], [11]. The subjective experience of "joy of use" is a crucial factor, which, in the end, determines the success of a product. A software interface might be efficient and easy to use and thereby fulfilling the basic functional needs of the user, however if it creates bad experiences users won't like it, which will create an emotional barrier. Emotions have also an impact on cognitive processing during every interaction a user performs with an interface; For instance, joy allows unusual associations and improves creative problem solving [12]; on the contrary anxiety or stress constrains attention to features of the environment concerned with safety, danger and basically survival [13]. Every experience with the product increases or decreases the value a user gives it, which in turn influences the expectations and motivations to use the product. Motivation itself is a factor that influences learning behavior and tolerance towards errors, or as Norman states "Attractive things work better" [14]. A positive attitude on the part of the user towards the product will almost certainly overcome many barriers and the successful detection of emotions during user tests in the development process is crucial to this goal. However, a straightforward application of biometric methods alone will not reveal the dimension of emotions, which might be crucial to its adaption in the development process.

This paper covers an approach to utilizing the valence arousal space, which is a 2D model for emotion modeling [15] in a Usability performance test to extend and support standard usability methods [16].

2 Theoretical Background

Emotions are created every time a perception of important changes in the environment or in the physical body appears. Basically an emotion is a psychological state or process that functions in the management of maintaining the balance of information processes in the brain and the relevant goals. Every time an event is evaluated as relevant to a goal, an emotion is elicited. Positive emotions occur when the goal is advanced; while negative emotions occur when the goal is impeded. The core of an

emotion is the readiness to act in a certain way [4], so some goals and plans can be prioritized rather than others. An emotion can interrupt ongoing interactions; it tends to change the course of action, e.g. if the users' goal of browsing a certain website for information is continually impeded by the slow loading of the site and the relevance of this site to their goal is low, they will change their course of action and search for the information on another website. If they can only find the information on this site, then the relevance towards the goal is high, the frustration tolerance is higher and so the likelihood of leaving the site will be lower.

Knowing the process now, the emotions still need to be labeled. Throughout the last 150 years there have been several theories on basic emotions with changing "fundamental emotions", including anger, aversion, disgust, desire, happiness, interest, surprise sorrow, etc. The different theories also have different bases, such as facial expressions, hardwired, instinctive or tendencies in relation to action [17]. However, the present dominant theory of emotion in neuroscience research lists a discrete and limited set of basic emotions.

So far, the approach of Schlosberg [15] is still up-to-date; it categorizes all kind of emotions in a two-dimensional model, postulating that every emotion has two aspects: a cognitive and a physiological component. The two dimensions are called valence and arousal. There's also a broad agreement that at least one more dimension of emotion exists, however with unclear definitions. According to Huether [18] there are three basic emotions, namely: joy (valence), anxiety (arousal) and surprise. He states that all other emotions can be categorized in one of these dimensions. Hence, a third dimension could be surprise, however, it could also be time or intensity and was not used in the present study. The 2D model of emotion therefore is expedient, as it allows an a priori reduction of complexity and better application for usability testing.

2.1 Modeling and Classification of Emotions

Once the accordant data is gathered, it's quite straightforward to classify emotions in the 2D valence-arousal space. Arousal describes the physical activation and valence the pleasantness or hedonic value.

An emotion such as stress, for instance, is modeled as high arousal and low valence, while joy and elation would be high arousal and also high valence. The arousal component can be measured with psycho physiologic methods, such as Skin Conductance Level (SCL), Heart Rate Variability (HRV) or Electroencephalography (EEG). Changes in the sympathetic and parasympathetic nervous system allow conclusions on the physical activation to be drawn.

The valence component is more difficult, as it consists of cognitions. Valence can be determined by the right questions and questionnaires. There are, however, also approaches to measure and calculate the valence from physiologic data. Chanel [19] reports the successful use of pattern classification to distinguish between three specific areas of the valence-arousal from both peripheral and EEG signals. The original Schlosberg model was later enhanced by [20] (for a discussion see [21]) and called "circumplex model of affect" (fig. 1a): all affective states arise from two fundamental neurophysiologic systems related to valence and arousal. Specific emotions arise from activation patterns within these two systems, accompanied by cognitive interpretations and the labeling of the physiological experiences, e.g. in this model joy and

happiness are conceptualized as a combination of strong activation in the neural systems associated with pleasure and moderate activation in the neural systems associated with arousal. These facile ways of modeling emotions in two dimensions [15], [20] can be mapped nicely to the dimensions of User Experience as shown in fig. 1b) and thus provide a measure for the overall UX.

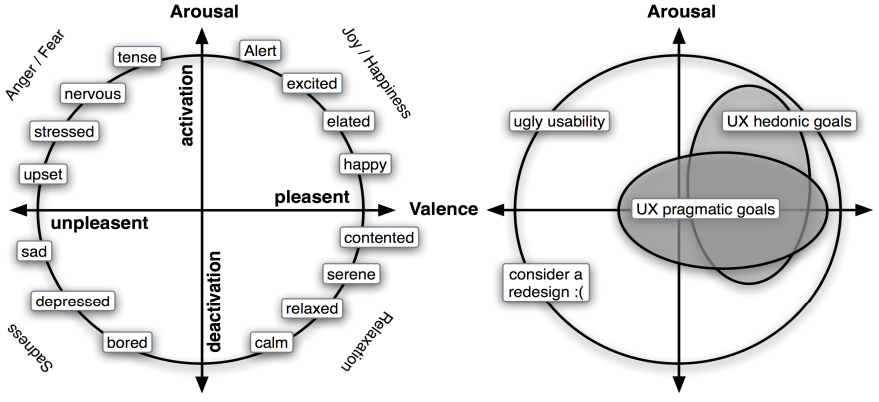

Fig. 1. a) The circumplex model of affect from Russel [20] b) applied to UX dimensions

Thereby the users' pragmatic and hedonic goals from UX correspond with certain areas of emotional states. Pragmatic goals concern functional metrics such as effectiveness and efficiency, whereby one can anticipate that some stress and unpleasantness will be accepted by the user in order to reach a goal. Hedonic goals, such as stimulation, identification, evocation, however, require a pleasant experience.

2.2 Measuring Emotions

Changes of physiological signals can be analyzed for stress arising during an interaction of a user with a product. These signals are regulated by the autonomous nervous system (ANS) and can be respiration, muscle tension, skin temperature or clamminess. More obvious signals are facial expressions, tone of voice, articulation, posture or gesture [22]. Each of these signals can be monitored or observed with some kind of sensor and method. The present study focuses on two these of these methods, which are measures of heart rate (HR) and skin conductance level (SCL). Skin conductivity depends on the activity of respiratory glands. This means the more sweat is produced the more electric current will be transported. SCL is generally a perfect indicator of the emotional state, as simple stress stimuli are followed by a rapid rise of skin conductivity within seconds. Muter et al. [23] found that especially SCL seems to be a good indicator for the overall usability of software, as they found a correlation between user-hostile systems and an increase of SCL. The heart rate (HR) is calculated as the number of contractions of the heart in a minute. It is therefore counted in "beats per minute" (bpm). The heart of an adult beats at about 70-75 bpm in resting state and 80-100 bpm during the day. However this is varies among people, depending on their physiology, age and stress levels, when the body is dealing with stress the heart rate

increases. From the HR the heart rate variability (HRV) can be calculated. High performance mental task is usually accompanied by an increase of heart rate (HR), blood pressure (BP) and a decrease of heart rate variability (HRV) [24]. Thereby the task complexity is positive correlated to the changes of the parameters, as the decrease of task complexity will lead to opposite changes.

2.3 Aspects of Emotion Detection in Practice

As the levels of arousal are subjective, it is also necessary to collect basic data from the test person in a relaxed state, which can then be compared to the data collected during the test situation. Further, it must be noted that the levels fluctuate during the day, especially when EEG is applied, so constant time windows for testing have to be defined. In order to build the valence-arousal space, the valence dimension has to be determined continually. Therefore it is sufficient to ask the user. This can be done with a six-point scale from UNPLEASENT to PLEASANT after every task.

The skin conductance level (SCL) is a method of observing event related changes of the ANS, as this parameter responds very fast. The place of the sensor should be chosen carefully, because pressure and movement influence it. The placement of the sensors at the fingertips, as it was done in the present study, might therefore be reliable, however it restricts user input on the keyboard. Using the heart-rate variability (HRV) means taking a delay into account, as the heart-rate (HR) does not change as fast as the SCL. HRV in combination with HR is a reliable stress detector, however bigger time windows are needed.

3 Methods and Materials

We used a modified NPL Performance Measurement method [25] and tested a learning management system, which was developed at Graz University of Technology. The approach included the combination of two formal usability methods, additional psycho physiological measures and valence/difficulty rating after every task, in order to assess the users' emotions. Two different parts of the system were examined, whereby two user groups with 20 subjects each accomplished five fundamental tasks for every system part.

3.1 Research Questions

Our study was targeted to investigate connections of performance metrics and emotions. It was anticipated that the tested system would have a good learnability. The first hypothesis (**H1**) was that the stress level of all test users decreases during the test, as they rapidly become accustomed to the interface. The second hypothesis (**H2**) assumed a positive correlation between user efficiency and the hedonic quality of the classified emotions. This implicates that those users who performed badly showed negative emotions, while users that performed well showed positive emotions. As different psycho physiologic methods had been used, the secondary question was how the results of these methods would correlate. Another question was the correlation of the emotion dimensions with the task difficulty rating and overall SUS rating.

3.2 Experimental Design

The test was split in a control condition (K1), the performance test (L1) and a short thinking aloud test (L2). Right after test, the System Usability Scale (SUS) questionnaire from Brooke [26] was used to derive User Satisfaction. During the whole test psycho physiological measures of EEG, SCL and HR were recorded. In L1 and L2 additional videos and screen recordings were made (examples can be seen in [8]. In the control condition, the test persons were asked to relax in order to get some base data. The relaxation process was supported by a Brainlight system (http://www.brainlight.com). It was used as a stimulation unit to induce relief by Steady State Visual Evoked Potentials (SSVEP), changing in a frequency range between 8 and 12 Hz for the duration of 10 minutes. The EEG recordings were done with an IBVA 3 electrode headband EEG from Psychiclabs Inc. (http://www.psychiclabs.net) at a rate of 512 Hz. The headband uses only 3 electrodes, so it can record an EEG of the frontal lobe only. For SCL and HR recordings a Lightstone from the wild divine project (http://www.wilddivine.com) was used. The HRV was later on calculated from the raw HR data. All data was recorded and synchronized using apple script on an Apple powerbook.

3.3 Analysis Procedure

For the control condition and the main tasks, the averages of SCL and HRV were calculated, per user and task. This data was then normalized in order to be comparable to the other datasets. The normalized values of all psycho physiological metrics from all test persons were summarized in a single database and combined with the outcome of the valence/difficulty task ratings. The data was then split into two groups (best/worst) according to the metric of user efficiency. Then the intersection of the arousal data per task, as well as the valence rating, was calculated over all users per group and normalized (see fig. 2), providing the necessary data for building the overall valence-arousal space. The space was built twice, first using HRV as arousal measure and then SCL. In a further approach the running difference measure of the normalized SCL (scaled by the maximum range of the data) was used to get an approximation for the instantaneous change in slope of the original signal. This signal was then thresholded at different levels, in order to get a sense of the number of peaks of the original SCL signal. Thereby, small thresholds should catch smaller peaks, while larger thresholds should catch only few very large peaks, which are usually correlated to high stress events.

3.4 Results

The analysis revealed a positive correlation between the user efficiency (performance) and the emotional state, determined in the valence arousal space. As can be seen in figure 2., the task results for the "worst" group show negative emotions, here less valence and a bigger variance in arousal, while the "best" group is distributed in an area of positive valence with less variance in arousal. Every datapoint in fig. 2 represents the average data of one task (for each group). The data is normalized on a (-2 / 2) scale. Fig 1 shows how to label the results with actual emotions, respectively mapping them to UX dimensions.

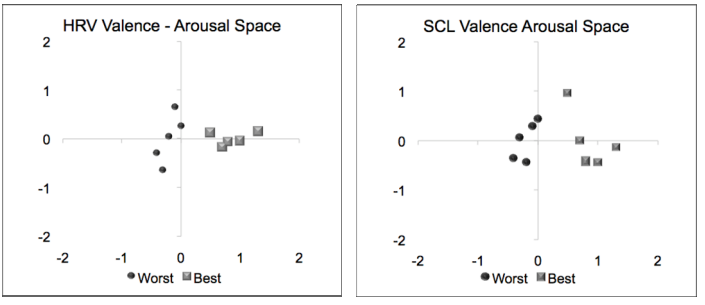

Fig. 2. a) Valence / Arousal (HRV) b) Valence / Arousal (SCL)

The comparison of both groups' averaged changes of HRV and SCL (fig. 3) over all tasks shows that HRV changes (fig. 3a) of the "worst" group range from stress to relaxation, while the "best" group shows no significant changes during the tasks. SCL changes (fig. 3b) shows an increasing trend for the "worst" and a decreasing trend for the "best". The success in Task 2 was very low (> 0. 5) and there the overall HRV decreased rapidly, which is a clear sign for stress. Figure 3 b) shows that the overall SCL of the "worst" increased during the whole test, beginning at task 2, showing stress.

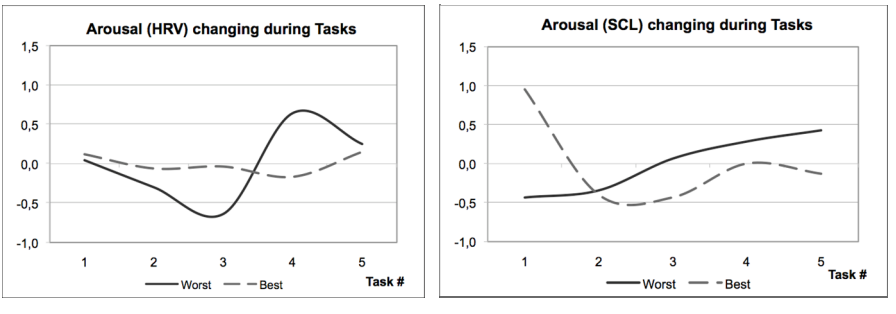

Fig. 3. a) HRV best vs. worst group b) SCL best vs. worst group

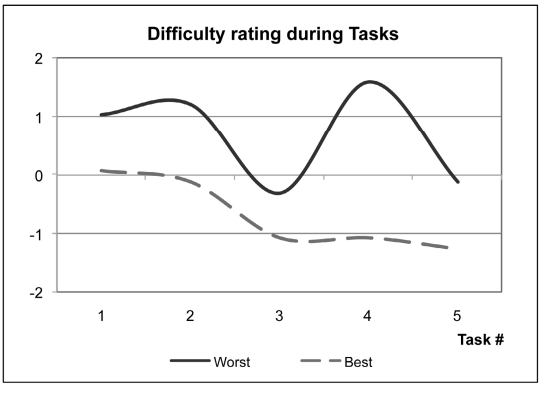

Fig. 4. Task Difficulty rating of best vs. worst group

The task difficulty rating displayed in fig. 4, shows that the "worst" group rated the overall difficulty higher and the variance of their rating is higher.

The user satisfaction metric, derived from the SUS, showed no significant difference between the two groups. The approach of counting large SCL peaks showed little, but no significant, differences between the groups.

4 Discussion

Hypothesis (H1) expected a decreasing stress level for all users, however this was only true for the successful group. Rather than the expected (H1) improvement in the worst group, a downward spiral was observed. Failure in a task appeared to increase the stress, which in turn appeared to increase the likelihood of failure in the following task. Hypothesis (H2) that expected a high user efficiency accompanied by positive emotions and vice versa low user efficiency with negative emotions was found to be true. The "Worst" group showed a high variance in the HRV values and a continuous increase of the SCL values, these are strong indicators for stress, as the valence-arousal space also shows. The high variance in the "worst" groups rating of the task difficulty in combination with the physiologic data is hard to explain. The 4th task was rated as very difficult although the HRV showed relaxation and the overall task success was moderate. The 2nd task was also rated as difficult and the HRV showed tension, however this task had an overall low task success. A hypothesis is that the relaxation reaction in task 4 was a counter reaction of the parasympathetic nervous system to the stress reaction, which can be seen in task 2 and 3 (low HRV Value). The valence-arousal space showed that this group tended to be frustrated. The "Best" group showed balanced values in HRV. The difficulty rating and SCL have a decreasing tendency. The balanced variance of the HRV might count for concentration. The decreasing SCL shows relief, which in combination with the valence can be seen as positive emotion. The approach of the peak extraction for SCL, for the overall test, showed no differences between the groups. So far, it is only reasonable to use this measure if the time of the task is invariant, particularly for small thresholds, because there will be more peaks in longer time windows. Further the number of peaks must be counted for every task in order to show any correlations between SCL peaks and task success or performance. So far, it can be concluded that the tested system elicited negative emotions for a group of users whose performance was low on the metric of user efficiency. These negative emotions are clues for a deeper analysis of the problems of this user group, as the ultimate goal should be a shaping of the user experience towards positive emotions. For users with a high efficiency positive emotions were detected. The area of the positive emotions can be mapped to the pragmatic goals of user experience, which means that the system fulfils the functional needs for these users. Emotions modulate almost all human interactions. If they are detected and synchronized to events or tasks the gained insight will help detecting causes of hidden irritation or frustration and provide a more complete picture of the overall user experience. Detecting issues when products are causing stress or aggravation will help developers to target areas for redesign.

Acknowledgements. This work has been partly funded by the European Commission under the project no. FP6-IST-2005-045056 EMERGE.

References

1. Stephanidis, C., Savidis, A.: Universal Access in the Information Society: Methods, Tools Interaction Technologies. Universal Access in the Information Society 1(1), 40–55 (2001)
2. Adams, R., Russell, C.: Lessons from ambient intelligence prototypes for universal access and the user experience. In: Stephanidis, C., Pieper, M. (eds.) ERCIM Ws UI4ALL 2006. LNCS, vol. 4397, pp. 229–243. Springer, Heidelberg (2007)
3. Holzinger, A., Searle, G., Nischelwitzer, A.: On some Aspects of Improving Mobile Applications for the Elderly. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4554, pp. 923–932. Springer, Heidelberg (2007)
4. Frijda, N.H.: The emotions. Cambridge University Press, Cambridge (1986)
5. Riseberg, J., Klein, J., Fernandez, R., Picard, R.W.: Frustrating the user on purpose: using biosignals in a pilot study to detect the user's emotional state. In: Conference on Human Factors in Computing Systems, pp. 227–228 (1998)
6. Ward, R.D., Marsden, P.H.: Physiological responses to different Web page designs. International Journal of Human-Computer Studies 59(1-2), 199–212 (2003)
7. Stickel, C., Fink, J., Holzinger, A.: Enhancing Universal Access – EEG based Learnability Assessment. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4556, pp. 813–822. Springer, Heidelberg (2007)
8. Stickel, C., Scerbakov, A., Kaufmann, T., Ebner, M.: Usability Metrics of Time and Stress - Biological Enhanced Performance Test of a University Wide Learning Management System. In: Holzinger, A. (ed.) 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian-Computer-Society, pp. 173–184. Springer, Berlin (2008)
9. Hassenzahl, M., Tractinsky, N.: User experience - a research agenda. Behaviour & Information Technology 25(2), 91–97 (2006)
10. Cockton, G.: Putting Value into E-valu-ation. In: Law, E.L.-C., Hvannberg, E.T., Cockton, G. (eds.) Maturing Usability: Quality in Software, Interaction and Value, pp. 287–317. Springer, Heidelberg (2007)
11. Ebner, M., Holzinger, A.: Successful Implementation of User-Centered Game Based Learning in Higher Education – an Example from Civil Engineering. Computers & Education 49(3), 873–890 (2007)
12. Isen, A.M., Daubman, K.A., Nowicki, G.P.: Positive affect facilitates creative problem solving. Journal of Personality and Social Psychology 52, 1122–1131 (1987)
13. Adams, R.: Decision and stress: cognition and e-accessibility in the information workplace. Springer Universal Access in the Information Society 5(4), 363–379 (2007)
14. Norman, D.A.: Emotional Design: Why we love or hate everyday things. Basic Books, New York (2004)
15. Schlosberg, H.: 3-Dimensions of Emotions. Psychological Review 61(2), 81–88 (1954)
16. Holzinger, A.: Usability Engineering for Software Developers. Communications of the ACM 48(1), 71–74 (2005)
17. Ortony, A., Turner, T.J.: What's basic about basic Emotions. Psychological Review 97(3), 315–331 (1990)
18. Hüther, G.: Biology of fear. Vandenhoeck & Ruprecht, Göttingen (1997)

19. Chanel, G., Ansari-Asl, K., Pun, T.: Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 375–380. IEEE, Los Alamitos (2007)
20. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
21. Remington, N.A., Fabrigar, L.R., Visser, P.S.: Reexamining the circumplex model of affect. Journal of Personality and Social Psychology 79(2), 286–300 (2000)
22. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001)
23. Muter, P., Furedy, J.J., Vincent, A., Pelcowitz, T.: User-Hostile Systems and Patterns of Psychophysiological Activity. Computers in Human Behavior 9(1), 105–111 (1993)
24. Schapkin, S.A., Freude, G., Erdmann, U., Ruediger, H.: Stress and managers performance: Age-related changes in psychophysiological reactions to cognitive load. In: Harris, D. (ed.) HCII 2007 and EPCE 2007. LNCS, vol. 4562, pp. 417–425. Springer, Heidelberg (2007)
25. Rengger, R., Macleod, M., Bowden, R., Drynan, A., Blayney, M.: MUSiC Performance Measurement Handbook, V2. NPL, DITC, Teddington (UK) (1993)
26. Brooke, J.: SUS: A “quick and dirty” usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry. Taylor & Francis, Abington (1996)