Usability Evaluation by Monitoring Physiological and Other Data Simultaneously with a Time-Resolution of Only a Few Seconds

Károly Hercegfi, Márton Pászti, Sarolta Tóvölgyi, and Lajos Izsó

Budapest University of Technology and Economics,
Department of Ergonomics and Psychology, Egry J. u. 1, 1111 Budapest, Hungary
{hercegfi,tim,ebediyen,izsolajos}@erg.bme.hu

Abstract. This paper outlines the *INTERFACE* methodology developed by researchers of our department. It is based on the simultaneous assessment of *Heart Period Variability (HPV), Skin Conductance (SC)*, and other data. The objective and significance of this paper are (1) showing its capability of identifying quality attributes of software elements with a *time-resolution of only a few seconds* and (2) presenting its *practical applicability* in the evaluation phase of a real software development process. The Department of Ergonomics and Psychology at the Budapest University of Technology and Economics carried out a contract-based applied research project for the Generali-Providencia Insurance Co. Ltd. The Company was in the process of further developing the software used in its customer centers, and our Department contracted to assess the user interface. Both analytical and empirical usability evaluation methods were applied. In this paper, we highlight the new experiences of the INTERFACE testing methodology.

Keywords: Usability testing and evaluation, empirical methods, case study, Heart Period Variability (HPV), Skin Conductance (SC).

1 The Frame of the Project

During the last decades, the Department of Ergonomics and Psychology at the Budapest University of Technology and Economics developed mutually useful relationships with some industrial partners successfully accomplishing various Research & Development (R&D) projects 4.

The Genarali-Providencia Insurance Co. Ltd. as a member of the Generali Group is owned by the Generali PPF Holding that has 9 million customers in Central and Eastern Europe. From these 9 million customers, 1.2 million are given by the Genarali-Providencia Insurance Co. Ltd.

Two kinds of different core systems are used by the Genarali-Providencia Insurance Co. Ltd. to keep their data: one for the life insurance (it is called SYNPAC), and another (it is called VIAS) for the non-life insurance, e.g. liability insurance, Casco insurance, etc. The management started to establish a new system that can visualize the data of both core systems. To carry it out, the IT department of the company started to develop a new frame system called Genesys.

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 59–68, 2009.

This software is used by approximately 1500 users. From these 1500 users, 300 users use it daily. The main part of the users work in personal customer centers. One smaller part of the users work as call center operators. It is a smaller group of the users, however, the time pressure in their job underlines the importance of the usability factors.

The Department of Ergonomics and Psychology at the Budapest University of Technology and Economics latched on to the software development to carry out the usability assessment.

The main goal of our project was to collect problems of the user interface (UI). From these misses we made a list that we gave it to the developers. Developing solutions for the found problems can be a next project.

Other goal of the Company was supporting a scientifically interesting research project transferring results of foundational researches to really applicable methodology in real software development process.

2 Applied Methods

The preparation of the project started in November 2007. The contract was signed in March 2008. We got in touch with the developers at the beginning of April 2008.

We started our project with studying the software and collecting data and facts. While we were studying it, the system was updated, so sometimes it was difficult to be up to date.

In the 2nd part of April we started to *observe* the users of the software, and we also made *interviews* with them. We performed the observation and interviews at 3 different locations: in the biggest customer service of the firm, in another type, smaller customer service, and in a call center.

We obtained some *logfile data* to establish objective initial data for the subsequent assessment.

We applied analytical methods (usability inspection methods) in May and June 2008. The backbone of the analytical evaluation was a Guideline Review, supported by Cognitive Walkthrough elements. Also GOMS model-based analysis was carried out.

The main part of the assessment was a carefully planned, deep empirical series of experiments applying the INTERFACE methodology described in the following section. The series of experiment were carried out in July 2008; the analysis of the collected records was performed in August 2008.

3 Description of Our Main Methodology: The INTERFACE

Figure 1 shows the conceptual arrangement of the *INTERFACE (INTegrated Evaluation and Research Facilities for Assessing Computer-users' Efficiency)* workstation.

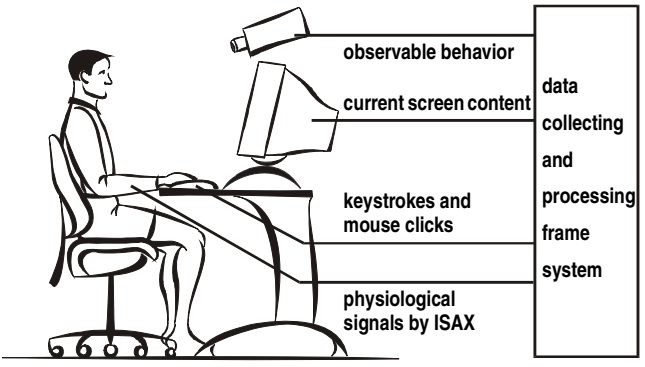

Fig. 1. Conceptual arrangement of the INTERFACE user interface testing workstation

The advantage of the methodology applied in our study lies in its capability of recording continuous on-line data characterizing the user's current mental effort derived from *Heart Period Variability (HPV)* and the user's emotional state indicated by *Skin Conductance (SC)* parameters simultaneously and synchronized with other characteristics of Human-Computer Interaction (HCI). This way, a very detailed picture can be obtained which serves as a reliable basis for the deeper understanding and interpretation of psychological mechanisms underlying HCI.

Elementary steps of HCI, like the different mental actions of users followed by a series of keystrokes and mouse-clicks, are the basic and usually critical components of using software. These steps can be modeled and analyzed by experts, but empirical studies of real users' interactions often highlight new HCI issues or give more objective results than expert analyses. One of the key aspects of the empirical methods is measuring *mental effort* as it is laid down e.g. in the earlier international standard of software product evaluation (ISO/IEC 9126:1991). Hence we need methods capable of monitoring users' current mental effort during these *elementary* steps.

To attain the above, a complex methodology was developed earlier at the Budapest University of Technology and Economics, by Prof. Lajos Izsó and his team [3, 4, 5]. This study presents an improved methodology and a new case study.

The INTERFACE simultaneously investigates the following:

- Users' observable actions and behavior
  - keystroke and mouse events;
  - video record of the current screen content;
  - video records of users' behavior: (1) mimics, (2) posture and gestures.
- Psycho-physiological parameters
  - Power spectrum of Heart Period Variability (HPV), regarded as an objective measure of current mental effort – we apply this signal successfully since more than 15 years [1, 2, 3, 4];
  - Skin Conductance (SC) parameters, indicating mainly the emotional reactions – recently integrated into our system.

A number of studies [1, 2, 3, 5, 7, 8, 9, 10] have shown that an increase in *mental effort* causes a decrease in the mid-frequency (MF) peak of the HPV power spectrum. The main advantage of the assessment method of the spectral components integrated into our system over the previously existing HPV-based methods is that the MF component of HPV shows changes in mental effort *in the time range of several seconds* (as opposed to the earlier methods with a resolution of tens of seconds at the best). This feature was achieved by an appropriate windowing data processing technique, and application of an all-pole auto-regressive model with built-in recursive Akaike's Final Prediction Error criteria and a modified Burg's algorithm.

We watch the Alternating Current (AC) component of the Skin Conductance (SC) responses focusing mainly on the *emotional* aspects of the HCI, in addition to our well-tried approach of *mental effort*. An interesting series of experiments analyzing SC responses is finished by one of our colleagues [6]. It is a good example of the promising way to use data mining techniques in empirical usability studies.

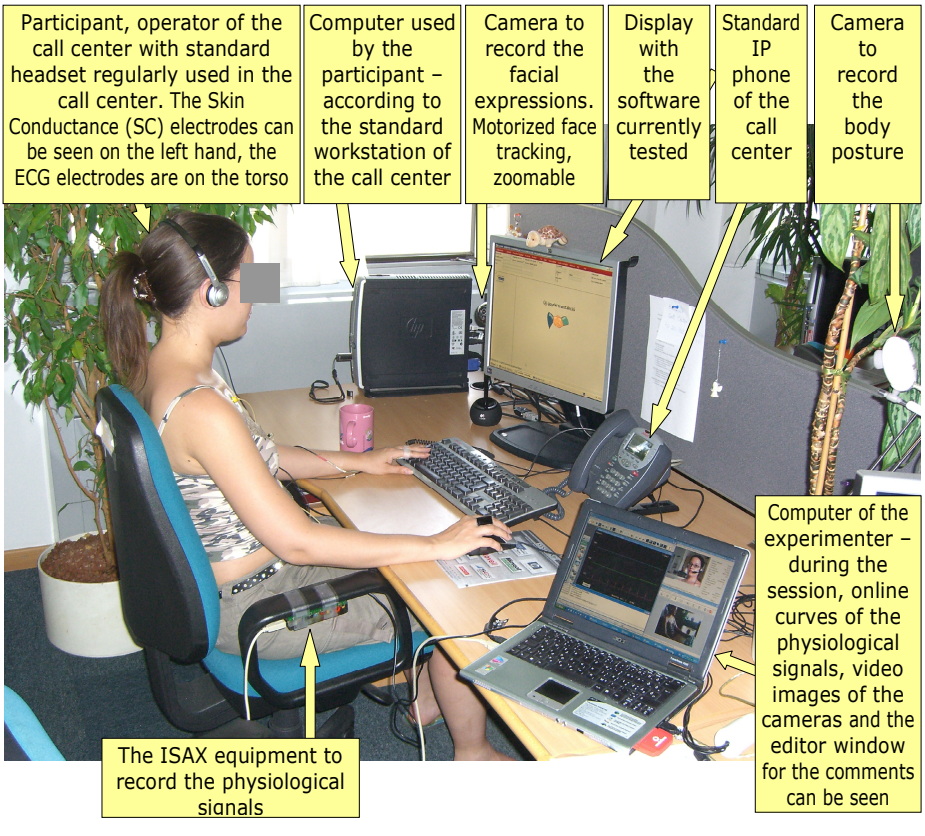

Fig. 2. The experimental arrangement applied during the sessions of the INTERFACE usability testing installed on a standard workstation of the call center

4 Applying INTERFACE in the Current Series of Experiments

The empirical sequence of experiments applying the previously introduced INTERFACE assessment was performed with the call center operators in the middle of July 2008.

Due to the impersonality caused by the phone calls, on one hand the simulation was more authentic than it could be in the personal customer centers. On the other hand, because of the specialty of the call center, we could count on concentrated, quick problem-solving. What more, the usability problems are the most critical in the call centers due to the time pressure.

The tasks occurred in the call center are usually not solved only by means of the Genesys, but these other software are not being examined in this project. However,

Fig. 3. The INTERFACE Viewer screen with a record of the empirical test of the Genesys software. As it can clearly be seen, currently the user makes significant mental effort – it is shown by the facial expression and gesture, and the low value of the last green profile curve of the Mid-Frequency (MF) power of the Heart Rate Variability (HRV) at the cross-hair.

we had to give controlled, simulated tasks to the users in order that they can solve them in the Genesys system. These simulated tasks helped us to be able to compare the 12 sessions. We used the version of the Genesys from the test server that is substantially equal to the version actually used in real.

12 real operators were involved as participants, each of whom we recorded a one-hour-long session with. The quantity of data gained from these sessions is really significant according to the depth of the enquiry. Due to the real life situation the users participating the series of experiments, were more or less disturbed by their colleagues' calls or talks. Since these are the employees' real conditions, they have got used to them, and so they could work on typical, real workstations of their workplace. These advantages gave reason not to hold this usability testing in laboratories.

Nevertheless, we chose a workstation that was located in the corner of the operators' room, in order not to disturb the others. Behind this workstation, the team leader's glass wall can be found, and so our staff could sit and make simulated phone calls from behind this pane. As mentioned before, during the recorded sessions, the users were given tasks from the real life, with quasi-real data, names, problems, questions, etc.; the main difference was that the customers were not real customers, but from our staff.

Three ECG electrodes were put on the users' torso and one on the left hand (in case of a right-handed person) for measuring Skin Conductance. After that the users put on their headsets and adjusted their seats.

Figure 2 shows the experimental arrangement applied during the sessions of the INTERFACE usability testing installed on a standard workstation of the call center.

Figure 3 shows the INTERFACE Viewer screen with a record of the empirical test of the Genesys software.

At the beginning of the session, we asked the users to relax for two minutes. We told them the aim and the details of the assessment. We always emphasized that we are not willing to assess they themselves, but the Genesys software by means of their help.

Relaxation was followed by two-minute mental effort: mental arithmetic. The result of the counting was not important; we only wanted to generate mental effort. These periods were planned for "calibrating" the physiological curves.

After that, the hard work started. The first customer (from our staff) rang the phone and asked some questions. To answer these questions the operator had to use the software under testing. Later 4 more similar calls came. Each call contained 2 to 4 questions. One part of the questions were just to "warm up", others were really difficult. The subtasks were based on the interviews, observations and expert analyses performed earlier.

The last part of the INTERFACE session was an interview.

5 Validation

As it was mentioned, the periods of relaxation and mental arithmetic were planned for "calibrating" the physiological curves.

The curves shown in the upper part of the Figure 4 were recorded during session #11, the ones shown at the bottom were recorded during session #10.

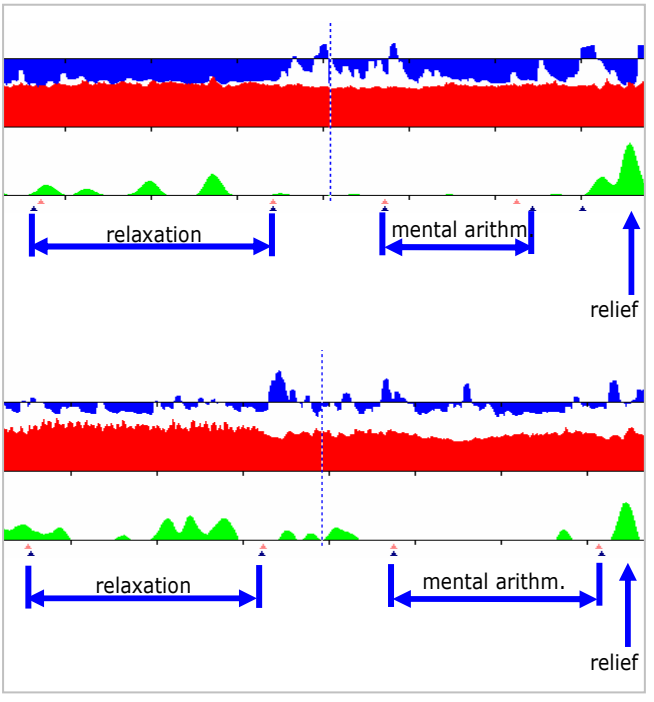

Fig. 4. The typical pattern of the relaxation and mental arithmetic in cases of two participants

In both cases, the three curves are the blue curves of the AC of Skin Conductance (SC), the red RR curves (Heart Periods), and the green profile curves of the Mid-Frequency (MF) power of Heart Rate Variability (HPV).

The blue curve of AC of SC is relatively smooth during both the relaxation and the mental arithmetic. During these sections, there are not any emotional peaks, and these two participants can be characterized as "stabile" type according to the typology of physiology. However, the beginnings and the ends of the sections are followed by peaks.

During relaxation, the MF component of the HPV increases, then the red RR curve has zigzags, and the green profile curve is relatively high. (In case of perfect relaxation, the profile curve should be consistently high. However, this is not expected in this experimental situation. The curve can be considered as high, especially in comparison with the next section.)

During mental arithmetic, the red RR curve gets smoother, and the green profile curve is significantly low.

After the "calibration" tasks, the participants really relieve. During this short period of relief, the participants get more relaxed than during the conscious, intended relaxation: the green curves have their highest peaks here.

These "calibration" tasks prove a validation of our method. The values of the MF power of the HPV were significantly higher during relaxation than during mental arithmetic. A non-parametric statistical method, the Wilcoxon Signed Ranks Test proves the difference (sig. 0.037 – Figure 5).

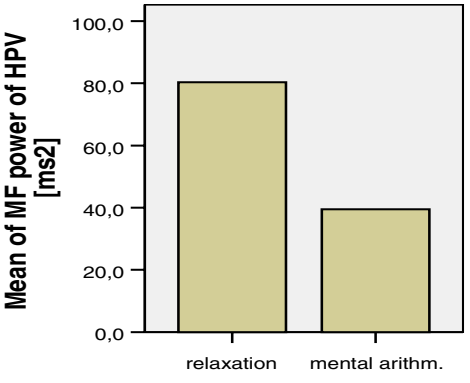

Fig. 5. Validation of measuring Mid-Frequency (MF) power of Heart Rate Variability (HPV) as an indicator of mental effort: the MF power of HPV was significantly higher during relaxation than during mental arithmetic (sig.0.037)

It is a significant difference, in spite of the non-perfect relaxation.

However, the mental arithmetic task works better: the significance of the difference between the values of MF power of HPV during mental arithmetic and in general, during the whole software usage section is better: the Wilcoxon test results sig. 0.002.

The values of the deviation of the AC component of Skin Conductance (SC) do not differ during the relaxation and the mental arithmetic significantly. As it was described earlier, this is the expected result.

However, the deviations of the AC of SC during the relaxation and the mental arithmetic are significantly lower than in general, during the whole software usage section: the Wilcoxon tests result sig. 0.009 and sig. 0.017.

After these results, we can say that the low value of the curve of the MF power of HPV really means mental effort, and the high deviation of AC of SC probably means higher emotions. Than, in the section of software usage, we look for moments with relatively high (and unwanted) mental efforts and high (and unwanted and not positive) emotions. This method gives us the key to find the problems of the UI.

6 A Sample UI Problem Identified by the INTERFACE Methodology

Commercial sensitivities prevent publication of the most of the details of the particular software problems found. However, Figure 6 gives an illustration.

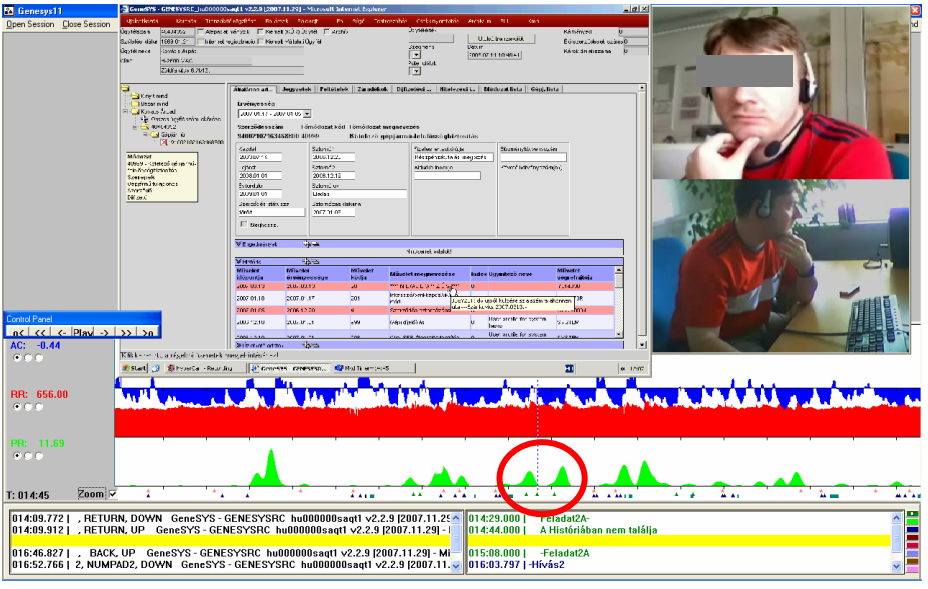

Fig. 6. The 11th participant during the first task of the second call. The mental effort can clearly be seen: it is shown by the facial expression and gesture, and the low value of the last green profile curve of the Mid-Frequency (MF) power of the Heart Rate Variability (HRV) at the cross-hair. In this case, the problem was caused by a bad design solution to ensure choosing a time period for a list view. This problem can also be found by analytical methods – but the INTERFACE highlighted this problem and gave objective evidence for it.

7 Conclusion

Based on the results presented here as well as in related papers, it can be stated that the INTERFACE methodology in its present form is capable of identifying the relative weak points of the HCI. By this methodology and the related workstation, it was possible to study events occurring during the HCI in such detail and objectivity that would not have been possible using other methods presently known to us. The sophisticated Heart Period Variance (HPV) profile function integrated into the INTERFACE system is a powerful tool for monitoring events in such a narrow time frame that it can practically be considered as a time-continuous recording of relevant elementary events. Measuring the Skin Conductance (SC) is a new opportunity to modulate the results.

Acknowledgements. The authors would like to thank Dr. Eszter Láng for the earlier developments, the Generali-Providencia Insurrance Co. Ltd. for the support of the deep research, and the participants of the series of experiments for their valuable contribution.

References

1. Chen, D., Vertegaal, R.: Using Mental Load for Managing Interruptions in Physiologically Attentive User Interfaces. In: Proc. CHI 2004, pp. 1513–1516. ACM Press, New York (2004)
2. Hercegfi, K., Kiss, O.E., Bali, K., Izsó, L.: INTERFACE: Assessment of Human-Computer Interaction by Monitoring Physiological and Other Data with a Time-Resolution of Only a Few Seconds. In: Proc. ECIS 2006, ECIS Standing Comm., pp. 2288–2299 (2006)
3. Izsó, L.: Developing Evaluation Methodologies for Human-computer Interaction. Delft University Press, Delft, The Netherlands (2001)
4. Izsó, L., Hercegfi, K.: HCI Group of the Department of Ergonomics and Psychology at the Budapest University of Technology and Economics. In: Ext. Abstracts CHI 2004, pp. 1077–1078. ACM Press, New York (2004)
5. Izsó, L., Láng, E.: Heart Period Variability as Mental Effort Monitor in Human Computer Interaction. Behaviour Information Technology 19(4), 297–306 (2000)
6. Laufer, L., Németh, B.: Predicting User Action from Skin Conductance. In: Proc. IUI 2008, pp. 357–360. ACM Press, New York (2008)
7. Lin, T., Imamiya, A.: Evaluating Usability Based on Multimodal Information: An Empirical Study. In: Proc. ICMI 2006, pp. 364–371. ACM Press, New York (2006)
8. Mulder, G., Mulder-Hajonides van der Meulen, W.R.E.H.: Mental Load and the Measurement of Heart Rate Variability. Ergonomics 16, 69–83 (1973)
9. Orsilia, R., Virtanen, M., Luukkaala, T., Tarvainen, M., Karjalainen, P., Viik, J., Savinainen, M., Nygard, C.-H.: Perceived Mental Stress and Reactions in Heart Rate Variability – A Pilot Study Among Employees of an Electronics Company. International Journal of Occupational Safety and Ergonomics (JOSE) 14(3), 275–283 (2008)
10. Rowe, D.W., Sibert, J., Irwin, D.: Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. In: Proc. CHI 1998, pp. 18–23. ACM Press, New York (1998)