I²MTC 2008 – IEEE International Instrumentation and
Measurement Technology Conference
Victoria, Vancouver Island, Canada, May 12–15, 2008

Relating GSR Signals to traditional Usability Metrics: Case Study with an anthropomorphic Web Assistant

Pierfrancesco Foglia[1], Cosimo Antonio Prete[1], Michele Zanda[2]
[1]Dipartimento Ingegneria dell'Informazione, Università di Pisa – Via Diotisalvi, 2 – 56126 Pisa - Italy
{foglia, prete}@iet.unipi.it
[2]Computer Science and Engineering, IMT Lucca Advanced Studies – P.zza S. Ponziano, 6 - 55100 Lucca – Italy
m.zanda@imtlucca.it

Abstract - *This paper shows how GSR signals can be related to traditional usability metrics. The test case is an e-government Website, enriched with an animated face to help users' navigation. First, we present the results on the effects of the talking face, which have been assessed with traditional usability metrics in a within-group experiment. Then, we show how such traditional analyses have correspondence in GSR signals. As the GSR signal is very subjective, its log data must be analyzed with care, in order to have significant and robust results. Applying a differential analysis, the GSR traces supported some findings obtained with traditional usability metrics.*

Keywords - *physiology, galvanic skin response, usability, Web, animated face.*

I. INTRODUCTION

The objective of our research is inferring useful information from human physiological signals, to be adopted by computer systems as input variables. We adopted the Galvanic Skin Response (GSR) sensing to assess if it can measure emotions, usually identified in usability tests with subjective measurements (surveys, questionnaires), in a more objective way.

We describe at large how physiological signals behave and how they can be treated in order to obtain robust information with many users.

To investigate how traditional usability metrics relate to GSR physiological sensing, we developed a prototype Website, with and without an Animated Face (AF) to support users' navigation. In a previous study [4], we focused on traditional usability metrics (completion times, visited pages, questionnaire answers ...) to assess the effectiveness and the effects of the AF. In this paper, we assess the feasibility of relating the AF effects, measured with usability metrics, to GSR signal variations. After presenting in sum the beneficial effects of the AF on users' performances, we show the effects of visiting avatar-enriched Web pages on GSR physiological measures. The GSR sensor provides very subjective metrics. For this reason, in order to increase the robustness of our physiological analyses, we compared GSR signals between users with novel differential analyses.

We observe that skin conductivity must be supported by traditional usability metrics if precise user emotions must be identified.

II. GSR ISSUES

The GSR is an indicator of human skin conductivity, and is measured via two electrodes. The GSR signal is affected when the sympathetic nervous system is active ([1], [12]), and its use in psychiatric evaluations was early documented by Jung [9]. We chose the GSR because it can be measured non-invasively, is a good measure of arousal, and its relationship to usability aspects were recently described by Lin et al. [10], while its relationship to the cognitive workload was shown by Shi et al. [14]. We investigated whether the GSR is effective in identifying specific emotions in users interacting with the AF, and if skin conductivity measurements can enrich or support the traditional usability metrics. Previous studies ([7], [10], [12]) indicated that the GSR can be an indicator of arousal/emotional response, but also an indicator of stress/anxiety. Discriminating among the former two emotions and the latter two can be quite hard.

In our experiment, we attached the remote electrodes of the GSR sensor to the users' fingers with velcro stripes, not to interfere with users' freedom of movement. In our usability test, we have been very careful in using the GSR sensor and its measurements. Since the GSR is a very subjective signal, it makes little sense to compare two different traces from two different users. Indeed, the raw conductivity value (provided by the sensor) is a very subjective measure, each user has its basis threshold and peaks (fig. 1). GSR signal traces are hard to compare for two main reasons: each user has its own skin conductivity; some users can have very flat conductivity traces (fig. 2).

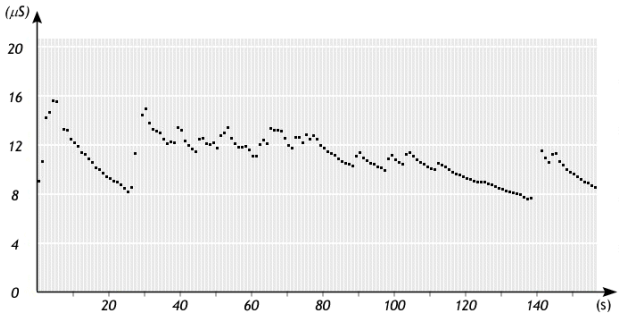

Fig. 1. Typical GSR signal trace. Peaks are related to precise events during task execution.

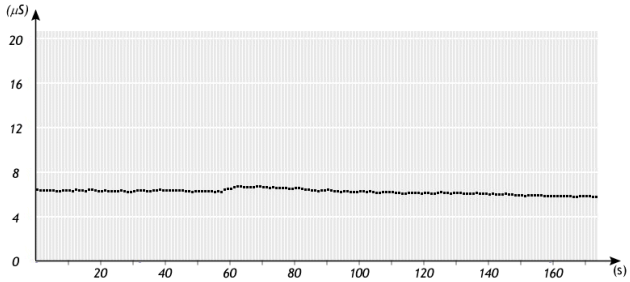

Fig. 2. The GSR signal of a flat trace user.

As observed by Picard [12], emotionally introverted users have peaky GSR traces, while highly extroverted users can generate very flat traces. Physiological signals are quite hard to cope with, they have full meaning only individually and comparisons must be very cautious. GSR traces always follow similar patterns [12]:

I. learning effect: an experienced event has less effect than a novel one,
II. summation effect: one single big event can be less powerful than many smaller events,
III. time variant effect: same event can cause different effect on same user at different times,
IV. recurrent pattern for emotions: an emotion typically causes a steep increase in the physiological signal (peak) followed by a smooth decrease,
V. subjective effect: same event can cause different effect on different users,
VI. relaxation pattern: a relaxed user shows a trace that decreases gracefully and continuously.

III. EXPERIMENT DESIGN

A. Motivation

We tested the AF because neurophysiologic studies have proven that faces stimulate human attention ([3], [5], [6]), and anthropomorphic agents increase users' perception of *social presence*, *telepresence* and *flow* [13]. *Social presence* refers to the feeling of "being with another" [2], *telepresence* is the sensation of "being there" [8], and *flow* is a construct depicting a user's interaction as playful and exploratory [15]. Consistent with such multidisciplinary physiological and computer science research on face perception, we conducted user tests to quantify how the AF influences users, and how users' skin conductivity relates to users' attitudes and behaviours during the usability experiment.

B. Test Case Website

Our objective is to measure with physiological analyses, effects observed with traditional Web usability metrics. We chose as our test case, a Public Administration prototype Website. The prototype Web site had two fully developed versions, one without and one with the AF (fig. 3). All activities and GSR signals were logged in a database.

Fig. 3. The home page of the e-gov Web site prototype (version with the AF). "*Help*" feature on top of the page.

C. Experiment Procedure

43 participants were recruited. The group was gender balanced, with age range 22 ÷ 64 (mean age 37). Out of the 43 recruited people, only one was not familiar even with logins and passwords: her test was only partially completed, and relative log data have not been considered in our analyses.

In order to further improve between-user comparisons, at the beginning of the test session each participant watched a short clip, which we used to make users relax and better calibrate the GSR.

We designed a within-groups experiment, in order to evaluate the effects of the AF in all users (Figure 4). Half of the participants (group1) completed the first task without the AF but completed the second task with the AF; the other half (group2) encountered the opposite situation (fig. 4 depicts this). The two given tasks were chosen to make users focus during the whole experiment. In the first task, users had to fill out the registration form with their personal details. Once participants completed the registration task, we provided them with a printed fake parking fine ticket with their personal details. In this way participants completed also the second task with full attention. Given that tasks were different, we minimized learning effects. At the end of each task, each user was asked to rate the mental effort that they perceived in completing the task. Users rated the required mental effort on the Subjective Mental Effort Questionnaire scale (SMEQ). The SMEQ scale is a single rating scale ranging from 0 to 150 [17]. In the end, participants were asked to fill out a questionnaire: users answered questions on 5-point Likert scales.

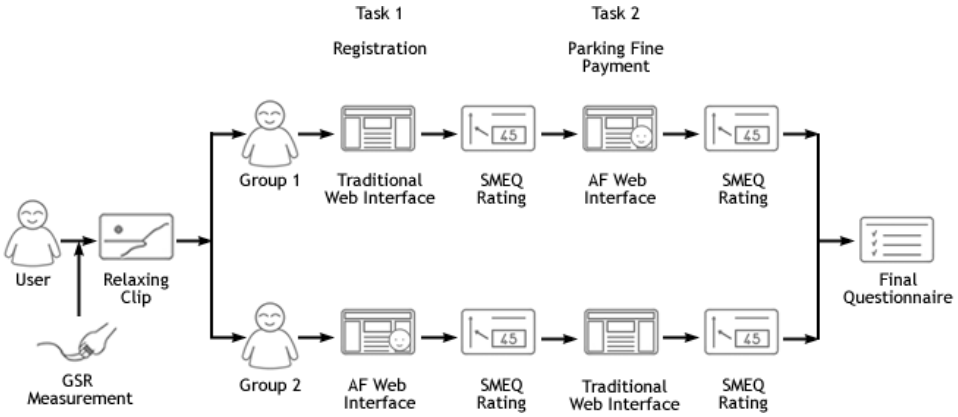

Fig. 4. Experiment procedure with the $task_1$ (registration) and $task_2$ (parking fine payment).

IV. RESULTS

A. Non-Physiological Analyses

Participants entering the AF-enriched home page used the "*help*" feature more frequently: 15 times out of 21 users with the AF, versus 2 out of 21 with respect to usual textual help ($t(1,42):p<0.01$).

The AF presence significantly reduced the number of visited pages (table 1). Our hypothesis is that fewer visited pages are related to a better user experience, in line with Zhu et al. [16]. Such reduction of visited pages is caused by the further information given by the AF, and by the increased use of the "*help*" feature by the users.

Conversely, the task completion time was not significantly altered (table 1). This result is easy to interpret: participants listened to the AF, and as a consequence their interaction with the website was not faster.

Participants, answering the questionnaire, positively rated the AF. On average, the effectiveness and the pleasantness were both rated above 3.6 on a 5-point Likert scale. Furthermore, only one participant switched off the AF.

B. Usual Physiological Analyses

Physiological signals are very subjective and influenced by many aspects, thus for each user we calculated a set of metrics in order to obtain more objective data:

I. Average galvanic skin conductivity on single task,
II. Average galvanic skin conductivity on single task normalized with the minimum,
III. GSR ratio on single task: $GSRratio = \frac{GSR_{\max} - GSR_{\min}}{GSR_{\min}}$
IV. GSR peaks steepness: percentage increase in skin conductivity.

Such metrics are common physiological measurements adopted by similar studies ([10], [11], [12], [14]). The first two metrics, average on single task and the normalized average, proved to be very subjective and thus useless: each user has own GSR conductivity. Even if the latter two metrics (*GSRratio* and *peaks steepness*) seemed to be more robust than the former ones, we observed no significant effect of the AF. Comparing $group_1$ and $group_2$ members under each single task with these metrics gave no significant results: GSR signals are very subjective, and comparing them directly makes little sense. Also concerning the SMEQ ratings (table 1 – first row) in the two tasks, at this first stage we observed no significant effects of the AF.

Finally, we analyzed the first impact in both groups with the AF: in $group_1$ at the beginning of the second task, in $group_2$ at the beginning of the first task. In both groups, the AF presence significantly increased the peaks observed ($t(1,40)$: $p=0.02$). This means that the AF affected for sure users' skin conductivity.

Table 1. Hypotheses showing significant results according to the Wilcoxon test. Each hypothesis is tested for both tasks. Roughly, *p*-values estimate the likelihood that observed differences are due to chance.

Hypothesis		means (or data)	*p*-value	Significance
a task is rated on the SMEQ harder than the other one	$task_1$	18.62 vs 22.8	>0.05	Not signif.
	$task_2$	20.35 vs 23.9	>0.05	Not signif.
AF-enriched web pages provide higher task completion rates than usual ones	$task_1$	(20 out of 21)	> 0.05	Not signif.
	$task_2$	(20 out of 21)	> 0.05	Not signif.
AF-enriched web pages provide fewer visited pages than usual ones	$task_1$	5.9 vs 6.3	0.042	Significant
	$task_2$	8.5 vs 10.2	0.039	Significant
AF-enriched web pages provide shorter task completion times than usual ones	$task_1$	305.6 s vs 230.6 s	> 0.05	Not signif.
	$task_2$	193 s vs 186.3 s	> 0.05	Not signif.

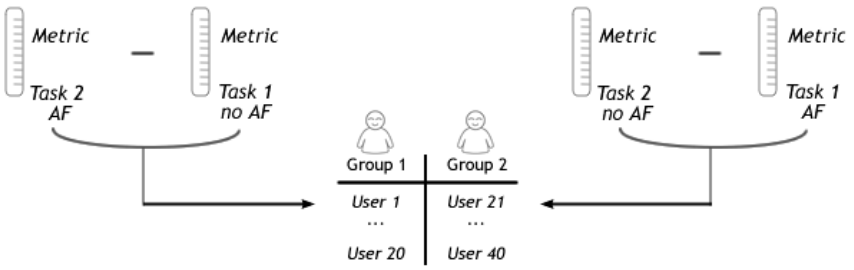

Fig. 5. Procedure to compute the Differential Analysis (DA), and analyze the results with statistics. Adopted metrics are described in Table 2.

Table 2. SMEQ, and GSR, with Differential Analysis (DA). Results are validated with t-Student tests.

Hypothesis	Metric for each group member	*p* -value	Significance
$Group_1$ users have a lower DA difference in SMEQ than $group_2$ users	*SMEQ* $task_2$ – *SMEQ* $task_1$	0.027	Significant
$Group_1$ users have a higher DA increase in their GSR maximum than $group_2$ users	*Max GSR value* $task_2$ – *Max GSR value* $task_1$	0.049	Significant
$Group_1$ users have a smaller DA decrease in the GSR ratio than $group_2$ users	*GSRratio* $task_2$ – *GSRratio* $task_1$	0.037	Significant
$Group_1$ users have a higher DA increase in peaks steepness than $group_2$ users	1^{st} *GSR peak value* $task_2$ – 1^{st} *GSR peak value* $task_1$	0.027	Significant

C. GSR Differential Analysis

We investigated very subjective metrics, such as SMEQ ratings and GSR signals, with differential analysis (DA). We designed the analysis in order to differentially compare the AF effect on users' metrics. Each user encountered the AF only in one task. Thanks to the within-group experiment, for each user we could calculate in the two tasks the metrics variations caused by the AF presence: (*metric* in $task_2$) – (*metric* in $task_1$). The results from the DA are tested for significance comparing metrics from members of $group_1$ with metrics from members of $group_2$ (fig. 5).

Considering results in table 2, $group_1$ users perceive a reduction of the mental effort, from $task_2$ to $task_1$ (on average -1.16 SMEQ points), while $group_2$ users perceive an increase (on average +8.9 SMEQ points). $Group_1$ users had AF in $task_2$ and they had a reduction in SMEQ ratings from $task_2$ to $task_1$. The opposite happened for $group_2$ users, who had the AF in $task_1$.

Thus, the AF presence significantly reduces the perceived mental effort in both groups.

Other interesting results arise from the DA of the GSR signals. $Group_1$ users (AF in $task_2$) have higher GSR signals (increased max values, smaller decrease of GSR ratios) than $group_2$ users (AF in $task_1$).

The last row in table 2 is related to the first time participants interact with the AF. Analyzing GSR peaks at the beginning of tasks with DA, the AF presence provides significantly steeper peaks.

Thus, AF-enriched web pages provide higher GSR values and peaks than usual web pages.

As we reported in paragraph II, identifying the exact emotion that drives the skin conductivity is not easy. In our test, considering the SMEQ ratings, the questionnaire answers and the users' attitude towards the AF, the GSR has mainly been an indicator of arousal and emotional response. Questionnaire answers rated the AF (section IV.*A*) positively. We clearly observed that participants were more relaxed and engaged while visiting an AF-enriched webpage than a usual one. This finding is in line with face perception studies and avatar mediated interactions, presented in section II.*A*. However, we have to report that one user, who was feeling very uncomfortable with the AF, had very high GSR traces, almost beyond the linear working zone of the sensor. For this single user, the GSR signal indicated stress, and she negatively rated the AF in the final questionnaire.

V. LESSONS LEARNED

In line with previous studies on facial interactions (section III.*A*), participants positively rated the AF, preferring to have it on the Website. The AF presence, together with the way we designed it, made the interaction more pleasant, as confirmed by the reduced number of visited pages, the SMEQ ratings and the questionnaire answers. We observed that participants were more relaxed and more emotionally involved with the assigned tasks when they interacted with AF-enriched web pages.

At the same time, the AF presence caused the GSR values to be higher, and caused steeper peaks as well (section IV.*C*). Peaks in skin conductivity are due to users experiencing emotions. The GSR can be adopted as a technique to quantify users' arousal/emotional response, or stress/anxiety. If the precise emotion is needed, it must be derived with traditional usability analyses, like completion times and questionnaire answers.

The GSR signal, as it is a physiological measurement (section II), must be analyzed with care. The first problem is

related to the nature of the emotions measured by the GSR. Future studies will have to discriminate between arousal and stress with orthogonal measures. For this reason, we collected feedback from the users (SMEQ ratings and questionnaire answers) to infer whether the GSR signal was mainly influenced by arousal or stress.

The second problem is that although the majority of users have similar GSR signal patterns (with steep peaks followed by smooth decreases), very extroverted users can have very flat traces. Measuring their skin conductivity is almost pointless.

The third issue is how GSR traces from different users can be compared to identify a common effect. Each GSR trace is closely related to a single user. In our test, we approached this issue with a within-groups experiment, designing and adopting a Differential Analysis. As a guideline to cope with physiological analyses in future usability tests, we suggest researchers to perform within-groups experiments in order to be able to compute a DA.

In sum, our research quantitatively shows that the AF reduces the visited pages and the mental effort, but it does not provide faster interactions. From a physiological point of view, it increases the GSR signal, which in our tests was linked to arousal and emotional response.

6. CONCLUSIONS

Our objective was relating physiological data to traditional usability metrics. We presented major features of physiological signals, and how they must be treated to obtain meaningful results. In a Web usability test, we measured users' skin conductivity (GSR). First, traditional metrics were adopted to analyze the effectiveness and the effects of enriching Websites with an Animated Face (AF). Then, we analyzed variations in users' skin conductivity to assess physiological effects of the AF. The Differential Analysis of the GSR signals, supported by traditional usability metrics, confirmed that participants increased their arousal and emotional response in the presence of the AF.

ACKNOWLEDGMENT

This work was supported in part by the "Fondazione Cassa di Risparmio di Pisa".

REFERENCES

[1] Abrams, S.. The polygraph in a psychiatric setting, *American Journal of Psychiatry*, 130(1):94-98, 1973.

[2] Biocca, F.. The cyborg's dilemma: progressive embodiment in virtual environment. *Journal of computer-mediated communication*, v.3(2), 1997. available at http://jcmc.indiana.edu/vol3/issue2/biocca2.html.

[3] Clark V.P., Maisog J.M. and Haxby J.V.. fMRI study of face perception and memory using random stimulus sequences, *Journal of Neurophysiology*, v.79(6):3257-3265, APS, 1998.

[4] Foglia P., Giuntoli F, Prete C.A. and Zanda M.. Assisting E-Government Users with Animated Talking Faces, *ACM Interactions*, 14(1):24-26, ACM Press, 2007.

[5] Haxby J.V., Gobbini M.I., Furey M.L., Ishai A., Schouten, J.L. and Pietrini P.. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, Sept. 28; 293(5539):2405-7, 2001.

[6] Haxby J.V., Hoffman E.A. and Gobbini M.I.. Human neural systems for face recognition and social communication, *Biological Psychiatry*, v. 51(1):59-67, Elsevier, 2002.

[7] Healey, J.A.. *Wearable and automotive system for affect recognition from physiology*. PhD Thesis, MIT, 2000.

[8] Heeter, C.. Being there: the subjective experience of presence, Presence: teleoperators and virtual environments, *available at http://commtechlab.msu.edu/randd/research/beingthere.html*, MIT Press, 1992.

[9] Jung, C.G.. On the psychophysical relations of the association experiment, *Journal of abnormal psychology*, 1, pp.:247-255, 1907.

[10] Lin T., Omata M., Hu W. and Imamiya A.. Do physiological data relate to traditional usability indexes? *Proceedings of OZCHI 2005*, pp:1-10, ACM Press, 2005.

[11] Moore M.M. and Dua U.. A galvanic skin response interface for people with severe motor disabilities, *ACM conference on computers and accessibility assets*, pp:48-54, 2004.

[12] Picard R.W.. Affective Computing, MIT Press, USA, 2000.

[13] Qiu L. and Benbasat I.. An Investigation into the effects of text-to-speech voice and 3D avatars on the perception of presence and flow of live help in electronic commerce, *Transactions on Computer Human Interaction (TOCHI)*, 12(4):329-355; ACM Press, USA, 2005.

[14] Shi Y., Choi E.H.C., Ruiz N., Chen F. and Taib R.. Galvanic Skin Response (GSR) as an index of cognitive workload, *ACM CHI Conference Work-in-progress*, 2007.

[15] Trevino L.K. and Webster J.. Flow in computer-mediated communication: electronic mail and voice mail evaluation and impacts, *Communication Research*, v. 19(5):539-573, SAGE, 1992.

[16] Zhu J., Hong J. and Hughes J.G.. PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation. *Transactions on Internet Technology (TOIT)*, 4(2):185-208; ACM Press, USA, 2004.

[17] Zijlstra R.. Efficiency in work behaviour. *A design approach for modern tools*. Delft, Delft University Press, Netherlands, 1993.