Analysis of Physiological Responses to a Social Situation in an Immersive Virtual Environment

Slater M.#, Guger C.*, Edlinger G.*, Leeb R.+, Pfurtscheller G.+,
Antley, A.#, Garau, M.#, Brogni A.#, Friedman D.#,

#Department of Computer Science, University College London, United Kingdom
*Guger Technologies OEG, Herbersteinstrasse 60, A-8020 Graz, Austria, www.gtec.at
+Institute of Human-Computer Interfaces, Graz University of Technology, Austria.

Abstract

An experiment was conducted in a Cave-like environment to explore the relationship between physiological responses and each of breaks in presence, and utterances by virtual characters towards the participants. Twenty people explored a virtual environment (VE) that depicted a virtual bar scenario. The experiment was divided into a training and an experimental phase. During the experimental phase breaks in presence (BIPs) in form of whiteouts of the VE scenario were induced for 2 seconds at four equally spaced times during the approximate 5 minutes in the bar scenario. Additionally 5 virtual characters addressed remarks at the subjects. Physiological measures including ECG and GSR were recorded throughout the whole experiment. The heart rate, the heart rate variability and the event-related heart rate changes were calculated from the acquired ECG data. The frequency response of the GSR signal was calculated with a wavelet analysis. The study shows that the heart rate and heart rate variability parameters vary significantly between the training and experimental phase. GSR parameters and event-related heart rate changes show the occurrence of breaks in presence. Event-related heart rate changes also signified the virtual character utterances. There were also differences in response observed between more and less social phobic participants.

Keywords: virtual environment, presence, breaks in presence, galvanic skin response, ECG, heart rate variability

1. Introduction

1.1 Measuring Presence

There are several strategies for the evaluation of the effectiveness of an immersive virtual environment (VE). A VE may be designed for a certain task, and performance measures specific to that task used for evaluation. The idea of 'presence' may also be used – this is more ubiquitous in the sense that it applies across many different applications, rather than being task specific. Often defined as the 'sense of being there' in the place depicted by the VE (Held and Durlach, 1992) it can be considered operationally as the extent to which observed behavior is similar to what it would be in a similar real-world situation (Meehan et al., 2002; Slater, 2004; Slater and Wilbur, 1997). In this approach 'behavior' should be considered as ranging from non-conscious low level physiological responses, through to high level volitional and conscious actions, including what people might say in interviews or in response to questionnaires. A critical discussion of the concept of presence can be found in (Draper et al., 1998).

Although people who experience immersive VEs such as head-mounted, head-tracked, stereo display systems, and Cave-like systems (Cruz-Neira et al., 1993) invariably and anecdotally report such a sense of presence, it is notoriously difficult to construct an accepted measure. The overwhelming response by researchers to this difficulty has been to develop presence questionnaires, for example (Singer and Witmer, 1999; Slater, 1999; Witmer and Singer, 1998). These are typically administered after participants have had a VE experience, and are scored on a series of Likert scales. Some method of combining (usually by averaging) the responses into an overall measure is then carried out, and this measure may then be related to other aspects of the environment. For example, a study that adopted this technique was by (Usoh et al., 1999).

It has been pointed out several times that questionnaires are problematic in the context of measuring presence: for example, they are unstable, in the sense of being very sensitive to prior experience (Freeman et al., 1999), they may not be able to distinguish reality from virtual reality (Usoh et al., 2000), and they can shed no light on whether 'presence' actually exists as a uniquely identifiable brain activity during the course of the experience to which it is meant to relate (Slater, 2004). Although answers to questionnaires can be invaluable, they are but the 'tip of the iceberg' of the totality of responses of the participant in the VE-human

overall system. As well as what people might say, we should also be interested in what they do – at as many levels of response as possible.

1.2 Behavioural Measures

Behavioral measures have been developed in an attempt to overcome, or at least supplement the use of questionnaires. Examples include gross behavioral responses such as looming, postural swaying (Freeman et al., 2000), and pointing behavior (Slater et al., 1995). From the idea that presence corresponds to behavior in the VE akin to what would have been observed in a similar real-world environment these behavioral measures are clearly appropriate. Their only problem is that they are application specific – for example, not every environment can include postural swaying or looming. Specific features would have to be designed into the VE solely for the purpose of evaluating presence, and such features may be inappropriate for the major purpose of the application.

One approach within the behavioral class has been to design environments that would lead to particular forms of response if sufficient presence were experienced. Applications concentrating on psychotherapy fall into this class – such as fear of spiders, heights, post traumatic stress disorder, and fear of public speaking – see (Rizzo and Buckwalter, 2001) for a recent example of this range of applications. These applications rely on presence, because if there were no presence, the corresponding anxiety necessary for successful therapy would not be induced, and therefore effective therapeutic intervention would be impossible.

Presence has been explicitly tackled within the context of a stress-inducing environment (Meehan et al., 2002). The purpose of this study was to examine whether expected low level physiological responses such as changes in GSR or heart rate would indeed occur as a result of a stressful experience within a VE – in this case being exposed to a precipice in the middle of a virtual room. The evidence strongly suggested that, in particular, heart rate measured as beats per minute, did increase as predicted when stress was induced. Perhaps this is the strongest evidence to date that the phenomenon of presence exists as an objective, measurable phenomenon, and that physiological time series can be exploited for measurement. This still leaves open the question of how presence can be measured in mundane, non-stressful environments.

1.3 Breaks in Presence

An alternative strategy for measurement of presence was proposed in (Slater and Steed, 2000). This work is premised on the idea of eliciting moments in time when 'breaks in presence' (BIPs) occur. A BIP is any perceived phenomenon during the VE exposure that launches the participant into awareness of the real-world setting of the experience, and therefore breaks their presence in the VE. Examples include gross events such as bumping into a Cave wall, getting wrapped in cables, through to more subtle effects such as revelations that come from seeing a tree as a texture map rather than a solid object. The original paper proposed a stochastic model that allowed the construction of a presence measure from knowledge of moments in time when participants reported such BIPs. This estimator was shown to be correlated with traditional questionnaire measures both in the original study, and also in a more recent study (Brogni et al., 2003). The main problem with this approach, however, is that it also relied on subjective reporting of BIP events, which therefore required prior training – thus potentially causing bias in the responses.

This BIP approach was taken further (Slater et al., 2003) where it was been shown that there is a likely heart rate and skin conductance response associated with a reported BIP. There was also some evidence that this physiological response was likely not caused by the very act of BIP-reporting. This approach provides an alternative strategy to the use of physiological measures. Rather than limiting the application of these to stress-inducing environments, it tries to find the physiological signature of a BIP, and since BIPs may occur in any environment it provides a more general approach – in particular the environments do not need to be stressful. Rather, BIPs themselves may be considered as the specific stress-inducers, and the goal is to try to automatically capture when (or if) these occur.

The research in this paper falls into the BIPs paradigm where we investigate whether such events are associated with a characteristic physiological signature. If they are, then they become objective experiential or perceptual events. Some other means, such as interviews, are necessary to assess the extent to which such events are actually experienced as breaks in presence. In addition we are interested in the inverse problem: if BIPs have a characteristic physiological signature, can the moments when these occur be predicted by analysis of the corresponding physiological time series? If this is the case then we would have an automatic method of detecting when such events occur, which would provide an overall measure of one aspect of the effectiveness of the VE.

In the first section the physiological recordings and measures are explained. Then the experimental procedure and the questionnaire are discussed. The next session compares the heart rate parameters in the training and experimental phase, compares social phobic and non-social phobic participants and shows the effect of BIPs on heart rate and GSR.

2. Physiological Measures

Participants were connected to a ProComp Infiniti by Thought Technology Ltd, which was fitted on a belt around the waist of the participants (see Figure 1). This was for recording physiological responses – specifically Galvanic Skin Response (GSR) and ECG.

Figure 1. A participant wearing the biosensors and the 3D glasses

2.1 Heart Rate and ECG Recordings

Several parameters can be extracted from ECG recordings in addition to the obvious one of heart rate (HR). For example, the heart-rate variability (HRV) can be used to describe the physiological behavior of the participant, and an event-related heart rate response may be useful to study the reaction of the subject to an event (e.g. such as a break in presence). The variability of HR is also influenced by the autonomous nervous system (ANS) activity. Statistical measures in time and frequency domains can be used for the quantification of the HRV. Recent studies show that the parasympathetic and sympathetic nervous activities influence the HRV at different frequencies (Task-Force, 1996).

In general there are several effects that influence the HRV:

1. Respiratory Sinus Arrhythmia (RSA) mediated by respiration is responsible for changes of the heart-rate in 2-5 seconds intervals and is controlled by parasympathetic activity. The sympathetic system is too slow to influence this frequency band;
2. Blood pressure regulation contributes to HRV in 10 second rhythms;
3. Changes with a periodic length above 20 seconds are mediated by the sympathetic system;
4. Changes in the range of minutes and hours are influenced by the neurohumoral oscillations in the circulating blood, by circadian rhythms or rapid eye movement phases during sleep.

Because the HRV analysis of the ECG signal in the time and frequency domains can be used for non-invasive investigation of autonomic cardiovascular regulation and sympathovagal interaction (Task-Force, 1996), each QRS complex (depolarization of the heart) is detected in the ECG signal and the distance from one to the previous is calculated and termed RR interval. Then the power spectrum of this time series can be estimated to describe the parasympathetic and sympathetic system (Bernardi et al., 1998, Guger et al., 2005, Task-Force, 1996). The power spectrum in the frequency range 0.04 – 0.15 Hz is normally referred as the low frequency (LF) component, and that in the range of 0.15 – 0.4 Hz as the high frequency (HF) component. The latter is mainly modulated by the parasympathetic system and the former by the parasympathetic and sympathetic systems. However, if the LF component is divided by the total power of the power spectrum then it can be seen to be mainly modulated by the sympathetic system (Task-Force, 1996). The ratio LF/HF describes the balanced behavior of the sympathetic and parasympathetic systems. Low frequency components (LF, 0.1 Hz) and high frequency components (HF, 0.15-0.4 Hz) indicate mental stress when the LF component is increased and the HF component is decreased. During dynamic exercise the heart rate changes but the HF component does not change significantly.

The ECG in this experiment was acquired as standard Einthoven I derivation (sampling frequency: 256 Hz) and the analysis was performed with the g.BSanalyze biosignal analysis software package (g.tec – Guger Technologies OEG, Graz, Austria).

2.2 Galvanic Skin Response

GSR, also sometimes called galvanic skin conductivity or Electro Dermal Activity (EDA), was also recorded in this experiment. This is measured by passing a small current through a

pair of electrodes placed on the surface of the skin and measuring the conductivity level. GSR was sampled at 32 Hz, and the signal obtained from electrodes on two fingers. Skin conductance is considered to be a function of the sweat gland activity and the skin's pore size. The real-time variation in conductance, which is the inverse of the resistance, is calculated. As a person becomes more or less stressed, the skin's conductance increases or decreases proportionally (Andreassi, 2000). GSR and other physiological measures have long been proposed for use in human-computer interaction, typically as methods for estimating the emotional state of users, and then adjusting computer response accordingly. A classic reference in this area is (Picard, 1997).

3. Experiment

3.1 Experimental Procedures

Twenty people were recruited by advertisement on the University Campus, and paid the equivalent of $10 for their participation. The participants were not involved in any way with the research group that carried out the experiment. They were presented with a disclaimer form that advised them of the potential risks involved in virtual reality displays, and advised that they could withdraw from the experiment at any time without giving reasons. They were read the procedures of the experiment from a standard script, and advised that they would be entering into a bar scenario, and that their task was to feel free to interact with the characters there and also pay attention to what might be happening there.

The virtual reality experience took part in a four-walled Cave-like system. This is an approximately 3 meter cubed area, with projection screens on the floor and three walls (but not the ceiling) providing a stereo view. We use the term 'Cave' in this paper to refer to the generic type of system described in (Cruz-Neira et al., 1993). The particular system used was a Trimension ReaCTor, together with Intersense IS900 head-tracking.

The scenario displayed to the participants was that of a 'bar' or 'pub'. This was approximately the same size as the physical Cave area, so that participants were free to physically walk around the virtual bar (hence no other tracking than head-tracking was needed).

There were 5 virtual characters in the bar, a barman and two couples, one pair standing near the bar, and the others sitting across the room. These virtual characters (2 women and 2 men)

would be ‘aware’ of the location of the participant (through head-tracking) and would often address remarks towards him/her. During the experience two songs played in the background one after the other, and in addition there was background chatter as might be heard in a real bar. The entire experience lasted approximately 5 minutes. The system was implemented using DIVE (Frecon et al., 2001; Steed et al., 2001) and the network interface between the various components achieved using VRPN[1]. The experiment was approved by the Ethics Committee.

Specifically, the virtual characters addressed remarks to the participant seemingly about a third person not present. This third person could be inferred to be a celebrity, who was to visit the bar. The virtual characters would wave and gesticulate towards the participant when they spoke. When the participant first entered the bar, some characters said ‘Hi’, ‘What’s up?’ and other greeting phrases. Then there were other phrases such as ‘Did you see the limousine yet?’, ‘Have you seen her photo in the newspaper?’, ‘If she doesn’t come soon I’m leaving’. The specific utterances of the virtual characters are not important for this paper – the most important point to note is that the environment was a ‘noisy’ one in both senses of the word. The virtual characters were designed to be more likely to say something and gesticulate to the participant under the condition of mutual gaze.

Each experimental session was divided into three phases:
The baseline phase: when participants entered the Cave they first stood for approximately 2 minutes with nothing being displayed. This was in order to check the physiological readings, and also to provide a baseline.
The training phase: next there was a short training session, where the VE depicted a room which had some large solid numbers on the floor, and participants were encouraged to move around from number to number, and also reach out to one of the numbers. We often find that participants must be explicitly told and shown that they can move around normally and bend down, or reach out within a VE, and this was the purpose here.
The experimental phase: the music started, and the bar scenario was initiated. Just before this black curtains were drawn across the opening to the Cave, so that the participants were isolated from the experimenters. Nevertheless the whole session was video taped, and the experimenters could watch the progress of the participants on a video monitor.

[1] http://www.cs.unc.edu/Research/vrpn

At four times at approximately evenly spaced intervals during the bar experience, the 4 projection walls became white, so that the bar and virtual people completely vanished visually. These 'whiteouts' lasted for 2 seconds each. These were the induced breaks in presence.

At the end of the second song the participants came out of the Cave, and were given an open interview which typically lasted about 15 minutes. This explored various aspects of their experience, in particular their reactions to the virtual characters, their presence, and their reactions to the whiteouts. Examples of the scenario are shown in Figure 2.

Figure 2. The bar scene

3.2 Questionnaires

Each participant completed a questionnaire prior to their immersion that gathered basic demographic information and also other background information regarding their use of computer games. Also each participant filled out a social phobia questionnaire (Watson and Friend, 1969) with 27 yes/no questions. Some examples are: (i) I feel relaxed even in unfamiliar social situations, (ii) I try to avoid situations which force me to be very sociable or (iii) I have no particular desire to avoid people.

After the experiment there was a further questionnaire that elicited their responses to the virtual characters followed by the interview. The results from the interview sessions are presented in (Garau et al., 2004) and are not considered in this paper.

4. Analysis

4.1 ECG Analysis Methodology

The first step in ECG analysis is to detect QRS (ventricular contraction) complexes in the ECG time series. The QRS complexes determine the distance in time from one heart contraction to the next one (RR interval). The term “NN interval” is used in the literature to indicate that only normal-to-normal beat distances are used for the calculations (non-normal beats like extra systoles are excluded). The QRS complexes in the ECG data were detected automatically based on a modified Pan-Tompkins algorithm (Pan and Tompkins, 1985). Then a visual inspection of the detected QRS complexes was performed to guarantee high data quality.

Changes in RR-intervals are referred to as HRV and can be described in time and frequency domains. The following are most important time domain measures:

MeanRR - mean RR interval [ms].
SDNN - standard deviation of NN intervals [ms].
MaxRR - maximum RR interval [ms].
MinRR - minimum RR interval [ms].
MinMaxRR - difference between MaxRR and MinRR [ms].
MeanHR - mean heart rate [bpm].
SDHR - standard deviation of the heart-rate [bpm].

The segmented measures divide the recorded ECG signal into equally long segments to calculate:

SDANN - standard deviation of the average NN interval calculated over short periods.
SDNNindex - mean of e.g. 1 min standard deviation of NN intervals calculated over total recording length.

The following measures yield differences between adjacent intervals:

SDSD - standard deviation of successive NN differences [ms].

RMSSD - square root of the mean squared difference of successive NN intervals [ms].

NN50 - number of intervals of successive NN intervals greater than 50 ms.

PNN50 - NN50 divided by the total number of NN intervals.

Frequency domain measures provide information on how power is distributed as a function of frequency. RR time series were resampled with a frequency of 2 Hz. Then the power spectrum of the resampled time series were estimated with the Burg method (Stoica and Moses, 1997) of order 15. The RR sequence was detrended and a Hanning window was applied prior to the spectrum estimation. The FFT length was 128 with an overlap of 64 and a sampling frequency of 2 Hz. Three main spectral components were distinguished: (i) very low frequency (VLF): <0.04 Hz, (ii) low frequency (LF): 0.04 – 0.15 Hz and (iii) high frequency (HF): 0.15 – 0.4 Hz. The unit of these parameters is ms^2. To express LF and HF in normalized units, each parameter is divided by the total power minus the VLF component. This minimizes the effect of the total power on LF and HF. The LF/HF ratio describes the balanced behavior of both components.

4.2 Comparison of Training and Experimental Phase

Table 1 shows the changes of HRV parameters in time domain averaged over all subjects for the two phases: (i) training and (ii) experimental phase. A sign test for paired samples was applied between the parameters of the training and experimental phase. The 4^{th} column gives the corresponding p-values.

Table 1: HRV time domain parameters

HRV	**Training**	**Experimental**	**p value**
MeanHR [bpm]	93.41	88.33	0.0072
RMSSD [ms]	22.51	27.08	0.0266
SDSD [ms]	22.61	27.21	0.0266
PNN50 [%]	4.94	8.26	0.0118
SDNNindex [ms]	37.46	40.86	0.0266

The analysis shows that there is a significant difference between the training and experimental phase in terms of the heart-rate (MeanHR): 93.41 bpm versus 88.33 bpm (p=0.0072), and also

in the heart rate variability (HRV) parameters. RMSSD, SDSD, PNN50 and the SDNNindex are smaller in the training phase than in the experimental phase.

Figures 3 shows the HRV frequency analysis results for one subject. During the training phase the LF component is increased to about 180 ms^2 and the HF component is decreased to about 5 ms^2. In the experimental phase LF is about 118 ms^2 and HF is 18 ms^2.

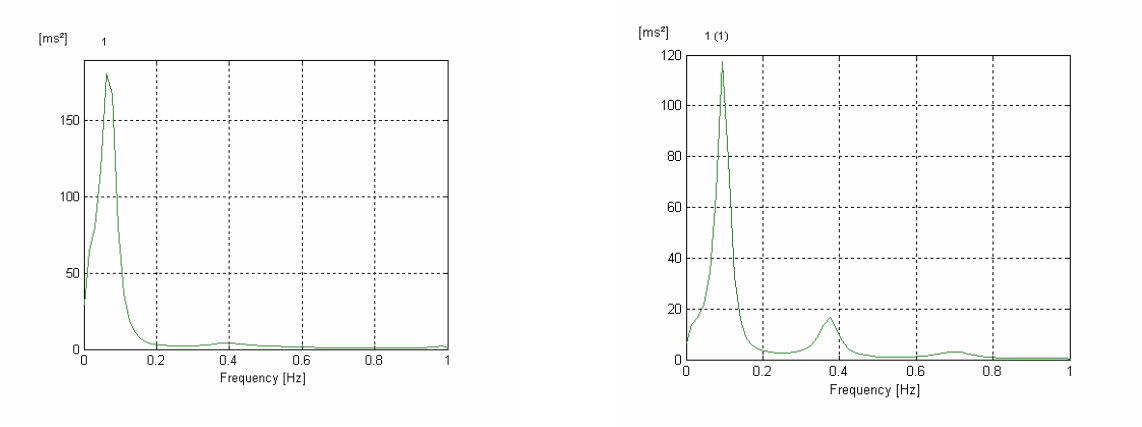


Figure 3: HRV in frequency domain of the training phase (left side) and experimental phase (right side).

The mean values for all subjects are represented in Table 2. Between training and experimental phase the normalized LF component (LFnorm) decreases from 74.5 to 64.9, the normalized HF component (HFnorm) increases from 15.1 to 22.3 and the LF/HF ratio decreases from 6.9 to 4.0. All three parameters have a p-value of 0.0266.

Table 2: HRV frequency domain parameters

HRV	**Training**	**Experimental**	**p value**
LFnorm [n.u.]	74.5	64.9	0.0266
HFnorm [n.u.]	15.1	22.3	0.0266
LF/HF [1]	6.9	4.0	0.0266

Figure 4 is a time frequency map of the HRV data. The map shows the evolution of the power spectrum over time for the baseline, training and experimental phase. An activated

parasympathetic system yields frequency components in the HRV map around 0.35 Hz (HF component) and the sympathetic system yields frequency components around 0.1 Hz (LF component). Basically the LF component is dominant throughout the whole experiment. The HF component in contrast varies between the baseline-, training- and experimental phase. The arrows indicate the changes between the different experimental phases. It can be seen that the activated HF component from the baseline phase becomes immediately smaller after the change to the training phase, but the amplitude increases towards the end of the training phase. With the change to the experimental phase the HF component disappears again and comes back after around 30 seconds. The HF component is present throughout the rest of the experimental phase.

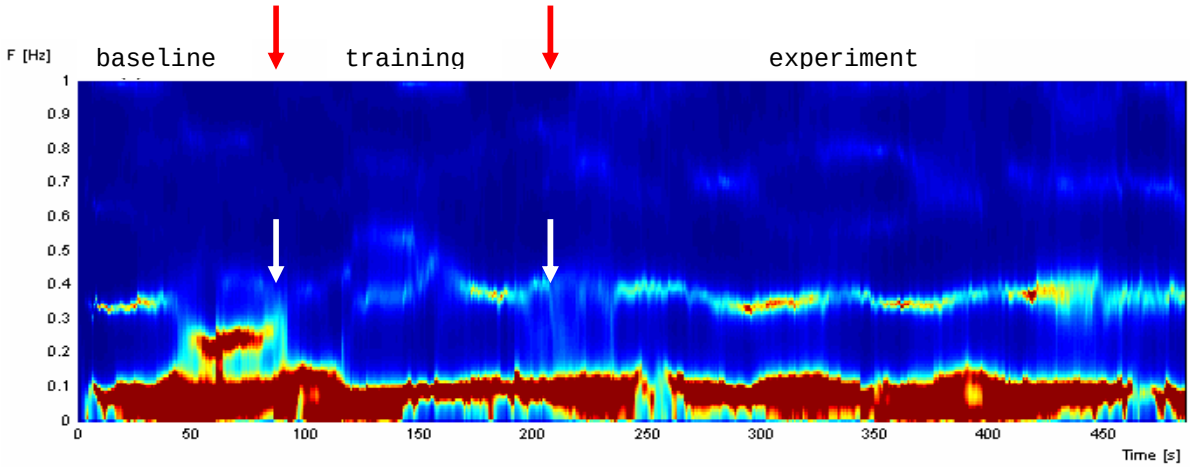


Figure 4: Time/frequency map shows the activation of parasympathetic and sympathetic systems.

4.3 Comparison of non-social phobic and social phobic participants

All participants filled out the social phobia questionnaire with 27 questions. A high score shows a more social phobic participant. The minimum observed score was 0 and the maximum was 23. The median was 6. For the following analysis the 9 subjects with the highest and the 9 subjects with the lowest score were selected. Other participants were not used for the comparison.

Figure 5 shows a boxplot of the heart rate and HRV parameter RMSSD. In social phobic participants the heart rate is 9.1 bpm higher than in non-social phobic. RMSSD is 14.2 ms lower. The p-value is in both cases below 0.05.

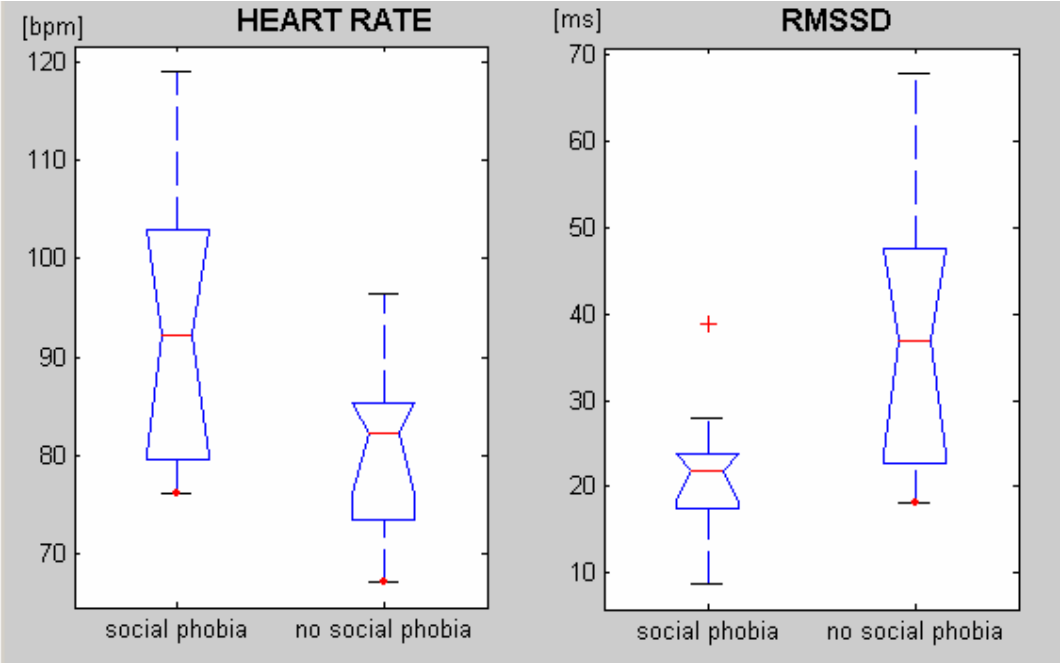


Figure 5: Heart rate and HRV difference depending on social phobic level of participants.

4.4 Event Related ECG: Effects of speaking avatars and BIPS

The change of the HR that was initiated by the speaking avatars and by the BIPS was analyzed with the event-related ECG analysis (ER-ECG). A specific number (n=2 to n=30) of RR intervals before the event and after the event (BIP, avatar speaking) were averaged. The difference shows the event-related change in HR. This procedure was repeated for all speaking events (about 30) and separately for all 4 BIPs. An important parameter is the number of RR intervals (n) used for the averaging. Therefore, the ER-ECG was calculated separately for n=2 up to n=30.

For the speaking events, the most significant change was found for n=3. The mean HR increase for all subjects was 0.7 bpm with a p-value of 0.04. The opposite was the case for the BIPs. The mean HR decrease for all subjects was 2.3 bpm because of the BIP (p=0.007). The most significant change was found with n=5.

4.5 The GSR Wave Form

In this section we consider whether the BIPs are signaled within the GSR data. Let x_{it} $(t = 1,..., n_i)$ be the GSR time series for the *i*th participant. An example is shown in Figure 6, for an arbitrarily selected participant.

Note that there is a tendency for GSR readings to drift upwards over time, due to accumulation of sweat on the surface of the electrodes. In order to overcome this problem we linearly de-trend all GSR data, using as break-points the start of the training and the start of the bar sequence, since these are entirely different experiences. We are only interested in local patterns of change rather than overall trend within the sequence. After de-trending the GSR data is also normalized to lie between 0 and 1, for convenience of plotting, and more importantly to allow averaging over the participants.

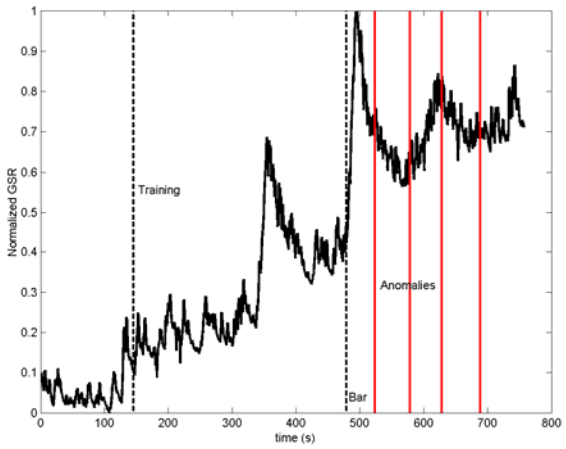


Figure 6. Event Sequence and GSR (participant 1)

Recall that there were 4 induced BIPs for each of 20 participants. The GSR wave-form is extracted for ±10s around each BIP point, and averaged over all BIPs over all participants. Each extracted GSR wave has its origin set at zero at the start of the sequence, so that only changes are averaged rather than absolute values. The result is shown in Figure 7 as the black curve. On the average the GSR curve rises to a peak about 3s after the onset of the anomaly.

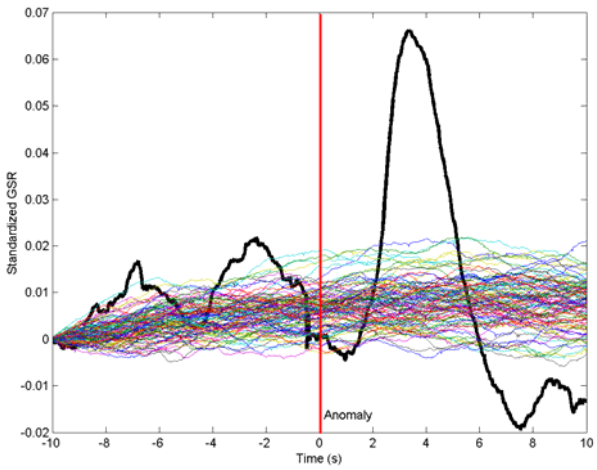


Figure 7. Average GSR wave forms – black curve true anomalies – coloured curves – randomized anomalies

It is possible to test the significance of this result by simulation. Suppose that we follow the same procedures as above, except that instead of using the true BIP times, we use 4 randomly generated times for each participant, within the valid period of the bar experience. If there is nothing special about the times when the true BIPs occurred, and our result is due to chance alone, then we should find a similar pattern for these randomly generated anomalies as we do for the true ones. If we repeat such simulated curve generation a large number of times and never or rarely see a pattern like the true curve, then this is a demonstration that it is unlikely that the characteristics of the true curve have occurred by chance. The colored curves in Figure 7 show 100 such simulated curves. It is clear that none of them approaches the shape of the true generated curve.

For each of the 100 simulated curves we compute two statistical characteristics – the mean and the maximum. For each of these we can therefore construct a sampling distribution. If we then compare these statistics for the true curve with the sampling distributions we have a quantitative measure of significance. These are shown in Figure 8. The mean for the true curve is 0.0097 and the maximum is 0.0661. These are so far into the right hand tail of the sampling distributions, that the significance level against the null hypothesis that the true curve shape was due to chance is zero for practical purposes.

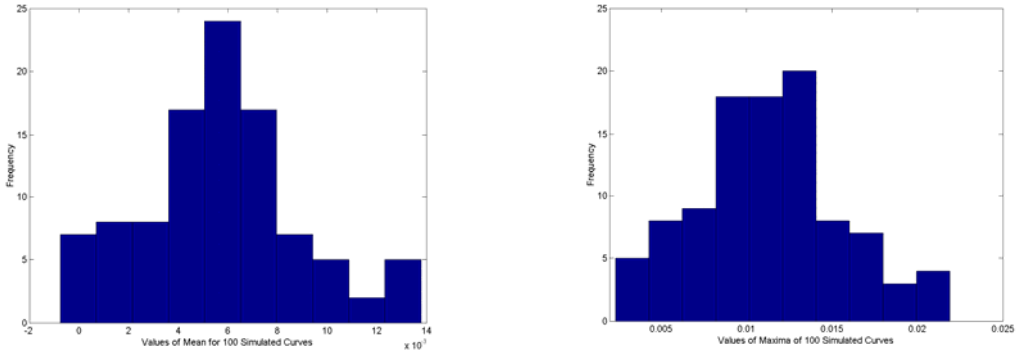


Figure 8. Sampling distributions for mean and maxima of the 100 simulated curves based

This section has shown that there is a characteristic GSR response to the induced BIPs. This adds further weight to the finding in (Slater et al., 2003) which was carried out for subjectively signaled BIPs. Here we have shown that unexpected anomalies cause a similar pattern of responses to the subjectively signaled BIPs. This result is encouraging, but not too surprising. Such BIPs, especially the rather extreme ones used here are likely to startle the participants, which is exactly what the GSR picks up. The more interesting question is the extent to which the BIPs can be identified by looking at the GSR sequences of individuals. We turn to this in the next section.

4.6 Using GSR to Infer BIPs

In this Section we consider the inverse relationship – from the GSR signal, can we 'predict' the locations where the BIPs occurred? It is clear from Figure 6 that this is not an obvious possibility. The most clear change in GSR happens around the start of the bar scenario – not surprisingly. Also the environment is very 'active' with many events happening all the time. Although in the previous section we showed that around the induced BIPs there is, on the average, a non-random response with a definite GSR spike, it would be difficult from the raw GSR data to identify such points in the case of GSR sequences of individual participants.

Instead we work with the hypothesis that the response to an anomaly might be embodied in the frequency domain and therefore not transparent in the original time series. In order to investigate this we use a continuous wavelet transform (Mallat, 1998) of the GSR signal.

$$C(scale, pos) = \int_{-\infty}^{\infty} f(t)\psi(scale, pos, t)dt \quad (1)$$

In Eq. (1) the original signal $f(t)$ is transformed by the continuous wavelet transformation ψ into the wavelet coefficient C. We use the Haar wavelet throughout (others were tested but with no difference in the results). Higher values of C indicate a greater similarity between the signal and the wavelet shape at the given scale and position. Scale is akin to frequency – small scale corresponds to rapidly changing parts of the signal (and therefore high frequency) whereas large scale corresponds to slowly changing parts of the signal (low frequency). As we shift and dilate the wavelet function over the signal we get an indication of its frequency properties at the various locations.

This is illustrated in Figure 9, which shows a plot of the wavelet transform coefficients for participant 1. The horizontal axis is time, and the vertical axis is scale. The scales chosen are 32 evenly spaced scales. Higher color intensity indicates higher absolute value of C.

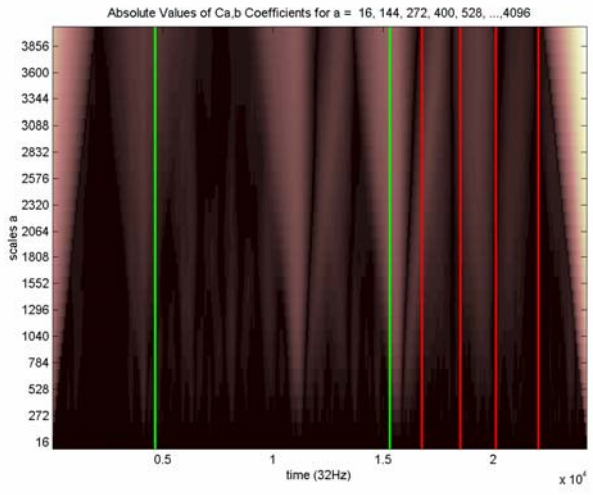


Figure 9. Wavelet transformation of complete GSR sequence for participant 1. Higher values of C indicated by greater color intensity. The first two lines indicate the start of the training and of the bar experiment, the last 4 lines indicate the anomalies.

Looking along the higher scales we see the slowly changing aspects of the curve, and at the lower scales the higher frequencies. It looks as if there is a change across a range of scales

around the start of each anomaly, and the start of the bar experience seems to correspond to a maximum across almost all the scale values.

We now derive a more systematic approach to exploring the relationship between the anomalies and the GSR time series. Consider the i^{th} participant (we drop the *i* for convenience) Let b_t be a variable such that $b_t = 1$ if time *t* is within 6 seconds of the start of a whiteout, and 0 otherwise. Hence for any participant *b* is a sequence of 0s with four sets of 192 1s (corresponding to the 2-second anomalies and sampling rate of 32Hz plus another 4 seconds to allow for latency in response). We treat *b* as a response variable and try to explain its variation by the variation in GSR. However, instead of using the GSR directly, we use its wavelet transformation coefficients. Therefore we use *k* explanatory variables $C_{tj}, (j = 1,...,k)$ where C_{tj} is the wavelet coefficient at time *t*, at the j^{th} scale being used. These are *k* scales distributed over some range of scales appropriate to the sequence. It is important to note that the time period $t = 1,\ldots,T$ is only over the time period of the bar scenario itself, and does not include the earlier resting and training times. *T* represents approximately 4.6 minutes in the actual bar scenario, and is therefore around 8900 for each person.

A normal linear regression relating b_t to the *k* scale coefficients $C_{t1}, C_{t2}, ..., C_{tk} (t = 1,...,T)$ would not be appropriate since predicted values from such an equation may fall outside of the range [0,1]. Instead we use standard logistic regression (McCullagh and Nelder, 1989) which restricts the fitted response variable to the correct range. This is shown in Eqs (2) and (3).

$$b_t = \frac{1}{1+\exp(-\eta_t)} \qquad (2)$$

with $\eta_t = \beta_0 + \beta_1 C_{t1} + \beta_2 C_{t2} + ... + \beta_k C_{tk}$ (3)

Whatever the value of the linear predictor (η) the predicted value of *b* will always lie between 0 and 1. The goal is to use the data to estimate the coefficients β_j. This is achieved by iteratively re-weighted least squares, and implemented by the function 'glmfit' in the MATLAB[2] statistics toolbox. The regression analysis was carried out for each of the 20 data

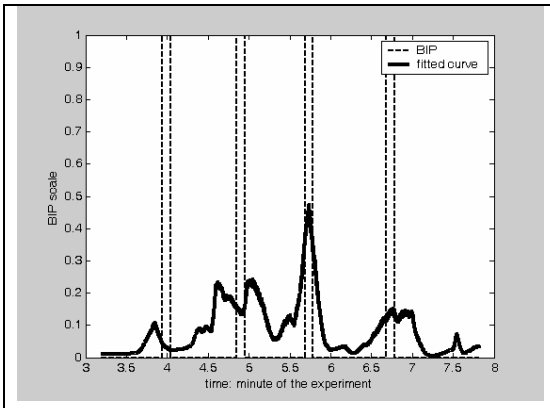
[2] http://www.mathworks.com/

sets, and for each Eq. (2)-(3) was used to compute the fitted values from the data (MATLAB function 'glmval').

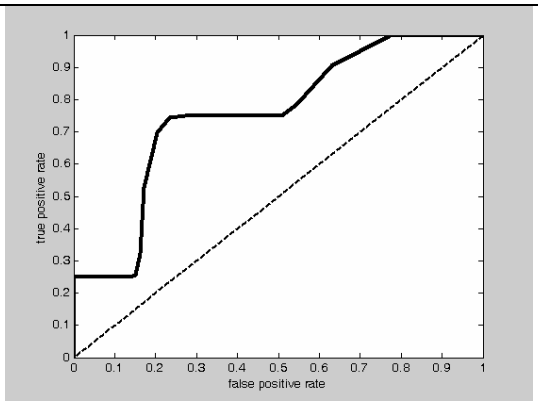
The question arises as to the set of scales that could be used. Too many explanatory variables results in over-fitting the data, and too few results in a poor fit. Moreover, using scales that are very low does not make much sense, since the very highest frequencies in GSR represent noise. The sequence for each participant could be analyzed individually, and the best set of explanatory (scale) variables determined appropriately for each case. However, here we carry out the same analysis across all the participants. Since $T > 8000$ we choose 8000 as the highest scale, and the lowest scale at 1000, with increments of 1000. This is equivalent to the 'largest' wavelet stretching over the entire sequence, and the smallest about 8 times less than the whole sequence. In this setup there are therefore just 8 explanatory variables. We fit the logistic regression model under this situation and therefore can obtain fitted values $\hat{b}$ from (2) and (3).

In order to examine goodness of fit between b and $\hat{b}$ we use a Receiver Operating Characteristic (ROC) analysis. This plots the true positive rate (vertical axis) against the false positive rate (horizontal axis) for each possible value of a diagnostic test. In other words, from the predicted values of the curve $\hat{b}(t)$ at time t we could construct the rule: there is an inferred BIP at time t if $\hat{b}(t) > c$. This result can be compared with the 'ground truth' (the actual value $b(t)$). If $b(t) = 1$ then the result is 'true positive', otherwise it is 'false positive'. The area under the curve (AUC) measures how well such a test discriminates between these two cases. An area of 0.5 means that the curve cannot discriminate at better than chance values, and the closer the area is to 1.0 the better the test. An interpretation of the AUC is that it is the probability that a randomly selected moment in time corresponding to a true BIP will be classified as a BIP than a randomly selected non-BIP moment (Mcneil and Hanley, 1984). The lowest AUC is 0.76±0.01, the median is 0.84±0.01 and the maximum is 0.96±0.005.

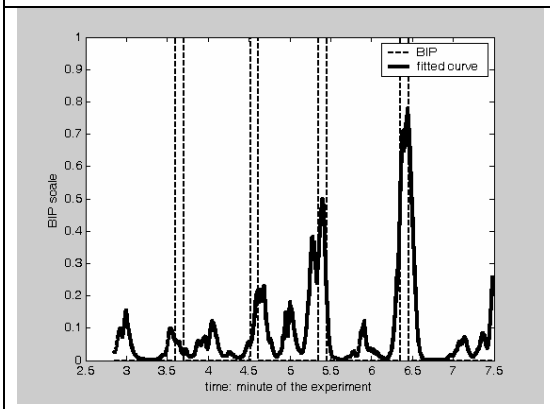
Figure 10 (a)-(c) shows these three cases – the plots of the true binary BIP curves and the fitted curves from Eqs. (2) and (3) against time, together with the corresponding ROC curve.



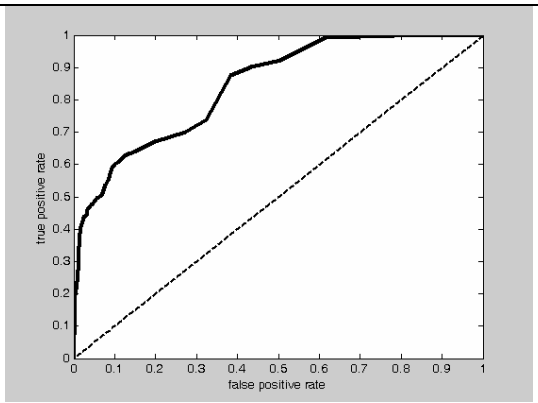
(a) Subject 3 (lowest fit)



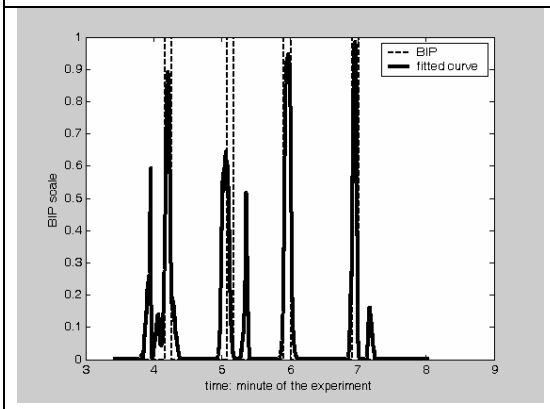
AUC = 0.76±0.01



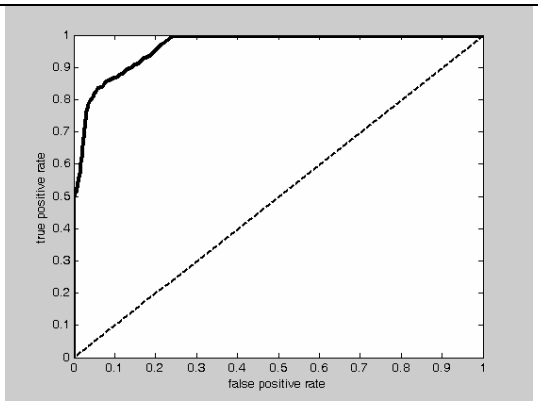
(b) Subject 2 (median fit)



AUC = 0.84±0.01



(c) Subject 11 (highest fit)



AUC = 0.96±0.005

Figure 10 : The left column shows the plot of $b(t)$ and $\hat{b}(t)$ against *t*, for the subjects with the lowest, median and highest AUCs. The corresponding ROC curves are shown in the right column. The time *t* is only during the time of the actual 'bar' experience.

This analysis has shown that it is possible to find an equation that models the relationship between the GSR sequence expressed in the form of wavelet coefficients at different scales

and the moments in a time sequence where anomalies occurred. This can be achieved with a very small number of parameters (in this case only 8).

5. Results

This paper has three main findings. The first is that the ECG parameters show significant changes of HR and HRV between the training and experimental phase. The second one is that HR and HRV parameters differ between participants who score higher on the social phobia scale than those who score lower. The third finding is that ECG and GSR analysis can be used to signify events such as speaking avatars and BIPs.

The time and frequency parameters display significant changes of HR and HRV between the training and experimental phase. The heart-rate decrease of 5.08 bpm could have 2 reasons:

1. during the training phase the subjects were specifically instructed to move around the VE to practice navigation, whereas in the experimental phase they were free to actively move around the bar or stay still if they preferred;
2. the subjects were more relaxed in the experimental phase when they were in the bar environment.

The HRV displays a high variability in the experimental bar phase. The time domain measures increased: RMSSD by 4.57 ms, SDSD by 4.6 ms, pNN50 by 3.3 % and the SDNNindex by 3.4 ms. In the case of frequency domain measures the LFnorm component decreased by 9.6, the HFnorm component increased by 7.2 and the LF/HF ratio is reduced by 2.9. It is well known that during dynamic exercise the HR is increased, but the HF component does not change significantly. In the experiment presented here the heart-rate changed and the HF component changed. Therefore it can be argued that the change was not initiated by dynamic exercise. Furthermore, an increased LF component and a decreased HF component normally indicate mental stress. This is precisely what happened in the training phases in our experiment, where subjects were more stressed than in the experimental phase. There are several possible explanations: the novelty of the experience of being in a VE, the unfamiliar equipment (including stereoscopic goggles and 3D mouse), and the novelty of entering the bar environment. Conversely, the positive relaxing effect of being in the bar could have caused the changes in HRV and HR.

The comparison of HR and HRV parameters of social phobic and non-social phobic participants showed that the HR is increased by 9.1 bpm in participants who are social phobic. Additionally the HRV parameters RMSSD and pNN50 were reduced. This shows that the participants who are social phobic were less relaxed in the bar environment. This is what would be expected in an equivalent real scenario, and hence is a sign of presence.

The most surprising result is the increase of the HR when an avatar speaks to the subjects. In two subjects the HR was increased by almost 3 % because of the speaking avatar.

In the case of the BIPs it is interesting that the HR decreased. This can be explained by the sudden 'whiteout' event, which surprised the subjects and changed their respiration rhythm. Furthermore it was shown that the GSR signal can be used to characterize the occurrence of this anomaly during the VE experience. This is in spite of the 'noisy' environment where there are many possible events that might cause people to become aroused or startled, thus triggering a GSR response. In addition the GSR responses to the induced anomalies can be modeled by a scale representation of the GSR signal, as obtained from a continuous wavelet transformation. Unlike the first result, which operates as an average over all anomalies and all participants, this result applies at the level of the individual, and is illustrated in Figure 10, where the anomaly occurrences of individuals are quite well-predicted by a logistic equation derived from a regression analysis. In other words the scale structure of the GSR signal can be used to model, for an individual, where the BIPs occurred. A reasonably good fit was achieved for most individuals with only 8 explanatory scale variables.

6. Conclusions

The primary contribution of this research is methodological. Apart from continuing our analysis of this particular experiment, examining the contributions of the other physiological data, the effects of interactions of the people with the virtual characters, the results of the in-depth interviews, and the implications for the theory of presence – we intend to continue to explore this paradigm with different types of BIP. For this first work, we have chosen a strong one – the whiteouts. But of course in a VE there are many different types of anomaly that may occur – caused by unwanted encounters with the physical equipment such as walls, cables and helmets, failures in graphics, tracking, sudden changes in frame rate or latency, and so on. Each one of these types of BIP needs to be investigated, and perhaps there may emerge one model that encompasses them all. In this paper we have talked about BIP as a particular

strong ‘failure’ in the VE system. The post-experimental interviews confirm that these events were considered as ‘breaks in presence’ by the participants. The analysis of the interview data also reveals a rich pattern of time variant presence responses that cannot be captured by the statistical analysis alone (Garau et al., 2004).

Acknowledgments

This project is funded by the European Union project PRESENCIA, IST-2001-37927.

References

ANDREASSI, J. L. (2000). *Psychophysiology human behavior and physiological response*. Mahwah, N.J.: L. Erlbaum, Publishers.

BROGNI, A., SLATER, M. & STEED, A. (2003). More Breaks Less Presence. In *The 6th Annual International Workshop on Presence*. Aalborg, Denmark, 2003.

CRUZ-NEIRA, C., SANDIN, D. J. & DEFANTI, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*: ACM Press, 1993, pp. 135-142.

DRAPER, J. V., KABER, D. B. & USHER, J. M. (1998). Telepresence. *Human Factors,* 40:354-375.

FRECON, E., SMITH, G., STEED, A., STENIUS, M. & STAHL, O. (2001). An overview of the COVEN platform. *Presence-Teleoperators and Virtual Environments,* 10:109-127.

FREEMAN, J., AVONS, S. E., MEDDIS, R., PEARSON, D. E. & IJSSELSTEIJN, W. I. (2000). Using behavioral realism to estimate presence: A study of the utility of postural responses to motion stimuli. *Presence-Teleoperators and Virtual Environments,* 9:149-164.

FREEMAN, J., AVONS, S. E., PEARSON, D. E. & IJSSELSTEIJN, W. A. (1999). Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence-Teleoperators and Virtual Environments,* 8:1-13.

GARAU, M., RITTER-WIDENFELD, H., ANTLEY, A., FRIEDMAN, D., BROGNI, A. & SLATER, M. (2004). Temporal and Spatial Variations in Presence: A Qualitative Analysis, 7th International Conference on Presence. In *Presence 2004*. Valencia, Spain, 2004, pp. 232-239.

HELD, R. M. & DURLACH, N. I. (1992). Telepresence. *Presence: Teleoper. Virtual Environ.,* 1:109-112.

MALLAT, S. G. (1998). *A wavelet tour of signal processing.* San Diego: Academic Press.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized linear models*. London ; New York: Chapman and Hall.

MCNEIL, B. J. & HANLEY, J. A. (1984). Statistical Approaches to the Analysis of Receiver Operating Characteristic (Roc) Curves. *Medical Decision Making,* 4:137-150.

MEEHAN, M., INSKO, B., WHITTON, M. & BROOKS, F. P. (2002). Physiological measures of presence in stressful virtual environments. *Acm Transactions on Graphics,* 21:645-652.

PAN, J. & TOMPKINS, W. J. (1985). A Real-Time Qrs Detection Algorithm. *Ieee Transactions on Biomedical Engineering,* 32:230-236.

PICARD, R. W. (1997). *Affective computing*. Cambridge, Mass.: MIT Press.

RIZZO, A. & BUCKWALTER, J. G. (2001). Special issue: Virtual reality applications in neuropsychology - Guest editors' introduction. *Presence-Teleoperators and Virtual Environments,* 10:Iii-V.

SINGER, M. J. & WITMER, B. G. (1999). On selecting the right yardstick. *Presence-Teleoperators and Virtual Environments,* 8:566-573.

SLATER, M. (1999). Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence-Teleoperators and Virtual Environments,* 8:560-565.

—. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence-Teleoperators and Virtual Environments*, 13:484-493.

SLATER, M., BROGNI, A. & STEED, A. (2003). Physiological Responses to Breaks in Presence: A Pilot Study. In *The 6th Annual International Workshop on Presence*. Aalborg, Denmark, 2003.

SLATER, M. & STEED, A. (2000). A virtual presence counter. *Presence-Teleoperators and Virtual Environments*, 9:413-434.

SLATER, M., USOH, M. & CHRYSANTHOU, Y. (1995). The influence of dynamic shadows on presence in immersive virtual environments. In *Selected papers of the Eurographics workshops on Virtual environments '95*. Barcelona, Spain: Springer-Verlag, 1995, pp. 8-21.

SLATER, M. & WILBUR, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence-Teleoperators and Virtual Environments*, 6:603-616.

STEED, A., MORTENSEN, J. & FRECON, E. (2001). Spelunking: Experiences using the DIVE System on CAVE-like Platforms. In *Immersive Projection Technologies and Virtual Environments*. Translated by B. Frohlicj, J. Deisinger & H.-J. Bullinger: Springer-Verlag/Wien, 2001, pp. 153-164.

STOICA, P. & MOSES, R. L. (1997). *Introduction to spectral analysis*. Upper Saddle River, N.J.: Prentice Hall.

TASK-FORCE. (1996). 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. *European Heart Journal*, 17:354-381.

USOH, M., ARTHUR, K., WHITTON, M. C., BASTOS, R., STEED, A., SLATER, M. & FREDERICK P. BROOKS, J. (1999). Walking > walking-in-place > flying, in virtual environments. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 359-364.

USOH, M., CATENA, E., ARMAN, S. & SLATER, M. (2000). Using presence questionnaires in reality. *Presence-Teleoperators and Virtual Environments*, 9:497-503.

WATSON, D. & FRIEND, R. (1969). Measurement of Social-Evaluative Anxiety. *Journal of Consulting and Clinical Psychology*, 33:448-&.

WITMER, B. G. & SINGER, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence-Teleoperators and Virtual Environments*, 7:225-240.