Multimed Tools Appl (2012) 56:553–567
DOI 10.1007/s11042-010-0619-7

Dual-task performance in multimodal human-computer interaction: a psychophysiological perspective

Domen Novak · Matjaž Mihelj · Marko Munih

Published online: 16 October 2010

Abstract This paper examines the psychophysiological effects of mental workload in single-task and dual-task human-computer interaction. A mental arithmetic task and a manual error correction task were performed both separately and concurrently on a computer using verbal and haptic input devices. Heart rate, skin conductance, respiration and peripheral skin temperature were recorded in addition to objective performance measures and self-report questionnaires. Analysis of psychophysiological responses found significant changes from baseline for both single-task and dual-task conditions. There were also significant psychophysiological differences between the mental arithmetic task and the manual error correction task, but no differences in questionnaire results. Additionally, there was no significant psychophysiological difference between performing only the mental arithmetic task and performing both tasks at once. These findings suggest that psychophysiological measures respond differently to different types of tasks and that they do not always agree with performance or with participants' subjective feelings.

Keywords Psychophysiology · Affective computing · Multitasking · Human-computer interaction · Mental workload

1 Introduction

Since the early studies of human-computer interaction (HCI), researchers have been interested in learning how humans respond to the physical and psychological demands of working with technology. However, while obtaining reliable information about a user's physical state is relatively simple, it is much harder to measure a person's psychological state. The concept of mental workload (the amount of mental work or effort necessary for a person to complete a task over a given period of time), for instance, cannot be detected directly, but through the measurement of some other variables that are thought to correlate highly with it [34]. In multimodal HCI, mental workload has been extensively studied in

D. Novak (✉) · M. Mihelj · M. Munih
Faculty of Electrical Engineering, University of Ljubljana, Tržaska 25, 1000 Ljubljana, Slovenia
e-mail: domen.novak@robo.fe.uni-lj.si

order to minimize the demands on the user and maximize the usability of the system. Applications where mental workload is important include information retrieval [10], supervision of automated equipment [22] and virtual reality [35].

One tool commonly used for mental workload estimation is psychophysiology, the study of physiological phenomena as they relate to behavior. Its basic assumption is that information about a person's mental state in a particular situation can be obtained from physiological processes. Its primary advantage is that it provides an objective, real-time assessment of a person's psychological state without any need for the person's active cooperation. Thus, it can be used to study workload and anxiety in diverse situations such as simulated driving [2], air traffic control [5] or even computer games [14].

Several studies in the field of multimodal HCI have begun implementing psychophysiology in a closed-loop fashion: to change a stimulus, task, game or virtual environment in response to the user's psychophysiological state. For instance, Octavia et al [20] propose a framework for virtual scenario adaptation where the user progresses to a different level of a task or even to a different task altogether based on psychophysiology. Ćosić et al [6] apply a similar adaptation procedure to virtual environments for treatment of post-traumatic stress disorder, advancing the patient through different scenarios based on psychophysiology. Liu et al [14] propose a method of dynamically adjusting a computer game in response to psychophysiological measurements.

Most of the aforementioned studies do mention interindividual differences in psychophysiological responses, which can occur due to differences in gender, age, experience, the presence of other people or a number of other factors [25, 27]. However, another factor needs to be considered. The type of task itself affects the relationship between task difficulty and user behavior [15], and different psychophysiological responses are sensitive to different types of tasks [1]. Thus, psychophysiological responses may not be reliable across different tasks, different virtual scenarios, or even different task difficulty levels.

Interpretation of psychophysiological responses becomes doubly difficult when a person is engaged in a multi-task environment. Tasks that are apparently dissimilar (e.g. talking and driving) can strongly interfere with each other when done together, both in laboratory conditions [26] and in real-world challenges [4]. One of the most useful methods for studying such situations is the dual-task paradigm. This paradigm involves performing two tasks concurrently, resulting in impaired behavioral performance on one or both tasks. Several different approaches to the study of the dual-task paradigm exist, but two are especially prominent. The first emphasizes structural and processing bottlenecks that cannot be devoted to carrying out two tasks at the same time [24]. The second emphasizes mental resources that can be shared between tasks and constrain how much information we can process at any given time. These resources can come from a single undifferentiated pool [12] or from multiple pools [32]. In the case of multiple resource pools, dual-task performance is worse when two tasks share common processing structures. For instance, vehicle drivers will have more success listening to instructions while driving than reading instructions while driving [23].

Psychophysiological measures have already been used in dual-task environments. Earlier studies examined simple dual-task situations such as manual tracking combined with auditory stimuli [1] while recent studies have focused on applications such as simulated driving [31]. However, few psychophysiological studies have examined dual-task performance in multimodal HCI. Thus, our paper evaluates two hypotheses:

a) Psychophysiological responses to two different tasks in multimodal HCI can be different despite no systematic difference in subjective responses.

b) When performing two tasks in multimodal HCI simultaneously, they may interfere with each other, and this would be reflected in psychophysiological responses.

The two tasks were chosen so that they would partially share resources and thus interfere with each other according to multiple resource theory. One task was a timed mental arithmetic task with a visual display and verbal operator response while the second was an error correction task with a visual display and manual operator response. Four different psychophysiological measures (heart rate, respiration, skin conductance and skin temperature) were compared with both objective (success rate) and subjective (self-report) measures of task performance.

2 Materials and methods

2.1 The tasks

Two different tasks were used for the dual-task paradigm: a mental arithmetic task and a manual error correction task. Each participant performed each task by itself (single-task period) and both tasks at the same time (dual-task period). The two tasks were selected to actively challenge the subject (since tasks which require active attention and thinking evoke stronger psychophysiological responses than, for example, passive viewing tasks), to use the same visual display and to use different input modalities (in our case, speech and haptics).

The mental arithmetic task presented the participants with two numbers that had to be multiplied. These numbers were randomly generated between zero and thirty for each subject. Four different possible answers were shown immediately underneath. One of the answers was correct while the other three were generated by adding or subtracting a random multiple of ten (between −40 and +40) to the correct answer. Using speech recognition, the participants verbally chose the answer they believed was correct (by saying "first", "second", "third" or "fourth"). If the participants answered correctly, their choice was coloured green. If the participant answered incorrectly, their choice was coloured red and the correct choice was coloured green. The participants had 15 s to answer each question; if they failed to answer within that time, the result was identical to making an incorrect choice (except that no number turned red). The time remaining was displayed using a large bar next to the numbers which grew progressively shorter and turned from green through yellow to red as time ran out. After a choice was made or time ran out, there was a five-second pause followed by the next two numbers to be multiplied.

The manual error correction task presented the participants with a virtual version of the classic inverted pendulum control problem. A thin pole with a weight at its top end is attached at its bottom to a moving cart. This vertical pendulum is inherently unstable and must be actively balanced by moving the cart horizontally. The participants were presented with a simulated cart and pole on the computer screen and moved the cart using a haptic interface. The cart moved in the same direction and with the same velocity as the end of the haptic interface. The model dynamics were adjusted in such a way to make balancing the pendulum moderately challenging. If the participants failed to balance the pendulum and it fell to a horizontal position, it was immediately reset to a nearly vertical position. Force feedback was also implemented with the haptic interface, allowing the participants to feel the reaction forces resulting from the movement of the cart. Though haptic systems are a relatively recent addition to multimodal systems, they are being used in a variety of applications in HCI [13], so we felt that such a task was suitable for our study. It has already been used in a study of psychophysiological responses to different difficulty levels in haptic workload [19], though that study did not involve more than one task.

During the dual-task period, both tasks were shown on the same monitor, one above the other (Fig. 1). During a single-task period, the task not being performed was replaced by a uniform gray background. Since the tasks share the same visual channel, they were expected to partially compete for the same mental resources. However, since the participant is required to respond both verbally and manually, the tasks were also expected to partially make use of different resources according to multiple resource theory [32].

2.2 Hardware configuration

The tasks were presented on a 22-inch display. A Phantom Premium from SensAble Technologies, Inc. was used as the haptic interface. This device provided a range of motion approximating hand movement pivoting at the wrist. It included a passive stylus and thimble gimbal and provided three degrees of freedom force feedback and three degrees of freedom positional sensing. Speech was recorded using a headset.

The electrocardiogram was recorded using pre-gelled, disposable surface electrodes affixed to the chest and abdomen. Skin conductance was measured using a g.GSR sensor (g.tec Medical Engineering GmbH). The electrodes were attached to the medial phalanxes of the second and third fingers of the non-dominant hand. Respiratory rate was obtained using a thermistor-based SleepSense Flow sensor placed beneath the nose. Peripheral skin temperature was measured using a g.Temp sensor (g.tec) attached to the distal phalanx of the fifth finger. The signals were amplified and sampled at 2400 Hz using a g.USBamp amplifier (g.tec).

A diagram of the entire system is shown in Fig. 2.

2.3 Software configuration

Both visualization and signal processing were implemented in Matlab/Simulink. The visual interface was programmed in Matlab's Virtual Reality toolbox and GUI design environment. Speech input was analyzed using Dragon NaturallySpeaking 9.0 Professional Edition, which determined the commands spoken by the participant and relayed them to

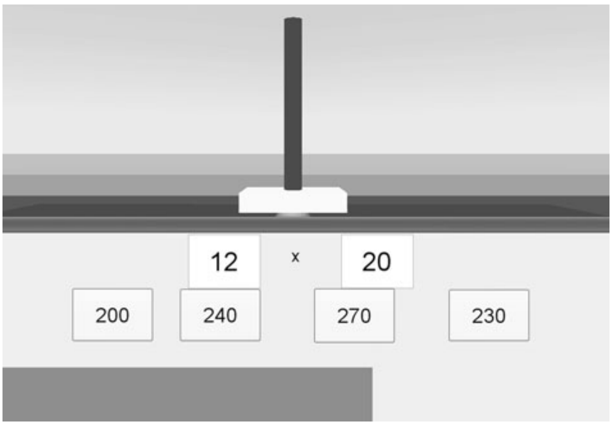

Fig 1 Computer display for the dual-task situation

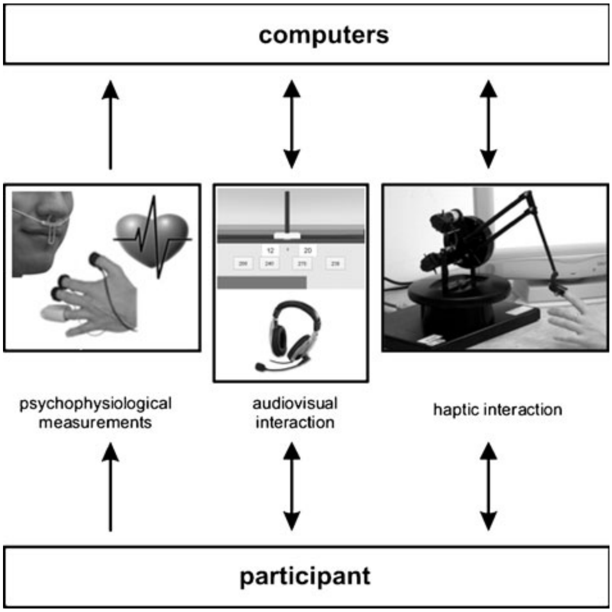

Fig. 2 Block diagram of the experiment hardware

Matlab. Physiological signals were imported directly into Simulink using drivers provided by the manufacturer of the signal amplifier. They were recorded raw at 2400 Hz, then filtered and analyzed offline. xPC Target 3.3 was used to control the Phantom haptic device. The task performance measurement was implemented in the same Simulink file used for physiological signal recording, and the software provided by the manufacturer of the signal amplifier performed the synchronization so that the maximum delay between two different signals was less than 0.02 s.

2.4 Participants

Twenty-four healthy male students and staff members of the Faculty of Electrical Engineering in Ljubljana (age range 20–46, mean age 28.0, standard deviation 6.6 years) participated in the experiment. Each subject signed an informed consent form after the purpose and procedure of the experiment was explained to him.

2.5 Experimental procedure

The experiment was conducted in a quiet area of the laboratory with constant temperature and humidity. There was never more than one participant inside the laboratory at any time. Upon arrival in the laboratory, the experiment was explained and the participant was allowed to practice each task as well as both tasks simultaneously. None chose to practice either a single-task or dual-task situation longer than 5 min. The measurement equipment was then attached and turned on. The signals were monitored until most effects of nervousness or novelty passed. Then, the three task periods (two single-task and one dual-task) were performed in random order. Each task period lasted 5 min and was preceded by a five-minute rest period during which the participant rested quietly and baseline values of psychophysiological parameters were obtained. After each task period, the participant was presented with a self-report questionnaire (detailed in Section 2.7). Between the rest period and the manual error correction task, participants were asked to move the Phantom left and right with the computer display turned off for 5 min. They were specifically asked to move as if they were performing the manual error correction task that they had previously practiced. This part of the experiment, hereafter referred to as the movement period, was necessary in order to gauge the effect that physical effort alone had on physiological responses. After completion of the tasks, the equipment was switched off and an informal interview was conducted about the entire experience. The entire experiment took about 60 min.

2.6 Task performance measures

Coordination task performance was evaluated by calculating the mean time between pendulum resets and the standard deviation of the time between pendulum resets. Cognitive task performance was evaluated by calculating the percentage of correct answers and the mean time needed to answer a question.

2.7 Self-report questionnaires

After each task, participants were presented with a short questionnaire that gauged their feelings during the task. They were asked to rate their level of satisfaction, frustration and concentration during the task on a six-point scale, with zero representing “did not feel at all”, one representing “very low” and five representing “very high”. They were also asked to rate the difficulty of the task on a five-point scale from “very low” to “very high”. An additional five-point scale was presented when performing both tasks simultaneously. On this scale, the participants had to specify how much their performance was affected by the divided attention, with 1 meaning “not affected at all” and 5 meaning “very strongly affected”.

2.8 Physiological measurements

Each subject’s psychophysiological state was evaluated using physiological signals recorded during the experiment. After the experiment, the signals were band-pass filtered offline and psychophysiological parameters were calculated for each rest and task period.

The skin conductance signal can be divided into two components: the skin conductance level (SCL) and nonspecific skin conductance responses (SCRs). The SCL is the baseline level of skin conductance in the absence of any particular discrete environmental event. Its mean value was calculated over the entire period. Since most skin conductance sensors (including ours) do not measure the absolute value of skin conductance, but only changes from an initial offset, the value of skin conductance at the beginning of the experiment was considered to be the zero value. SCRs are temporary increases in skin conductance followed by a return to the tonic level. They can occur in response to strong stimuli, but also occur in the absence of any specific event. Every increase in skin conductance was classified as a SCR if its amplitude exceeded 0.05 μS and the peak occurred less than 5 s after the beginning of the increase. In addition to SCR frequency, we calculated the mean SCR amplitude. Previous studies have found SCL and SCR frequency to be correlated with mental workload and general arousal [5, 8, 18].

Mean respiratory rate was calculated in breaths per minute. Additionally, respiratory rate variability was estimated by calculating the variance of the respiratory rate time series. Both respiratory rate and respiratory rate variability have been shown to be connected to mental workload (e.g. [3, 30]).

Analysis of the ECG began by extracting the times between two normal heartbeats (NN intervals) and converting them into mean heart rate in beats per minute. After calculating heart rate, several different measures were used to estimate heart rate variability (HRV) [29]. In the time domain, the standard deviation of NN intervals (SDNN) and the square root of the mean squared differences of successive NN intervals (RMSSD) were calculated. For frequency-domain analysis of HRV, NN intervals were converted into an instantaneous time series using cubic spline interpolation (as suggested in [29]) and the power spectral density (PSD) of this time series was calculated using Welch's method. The PSD has two frequency bands of interest to us: the low-frequency band (LF) between 0.04 Hz and 0.15 Hz and the high-frequency band (HF) between 0.15 Hz and 0.4 Hz. Three frequency-domain estimates of HRV were calculated: total power in the LF band, total power in the HF band (commonly referred to as respiratory sinus arrhythmia—RSA) and the ratio of the two (commonly referred to as the LF/HF ratio). Heart rate and the various estimates of heart rate variability have been shown to be connected to mental workload (e.g. [2, 5]).

Final skin temperature was recorded for each period by averaging temperature during the last 5 s of the period. Skin temperature has been shown to be connected to mental workload and work stress [21].

2.9 Statistical methods

Statistical significance was calculated using a one-way repeated-measures ANOVA or, if the assumptions for ANOVA were not met, a repeated-measures ANOVA on ranks. Differences were considered significant for $p<0.05$. The Kolmogorov-Smirnov test with Lilliefors' modification was used to test for normality.

3 Results

3.1 Task performance

During the cognitive task, participants correctly answered 80.4% of all questions. The time needed to answer a question was 7.7±3.9 s (mean±standard deviation). When performing both tasks simultaneously, participants correctly answered 73.3% of all questions (fewer than during the cognitive task, $p=0.03$). The time needed to answer a question was 8.5±4.8 s (higher than during the cognitive task, $p=0.006$).

Without participant input, the inverted pendulum fell to the ground and was reset every 6.0 s. When balanced by participants during the manual error correction task, it fell and was reset every 19.2±5.4 s. When the participants performed both tasks simultaneously, the pendulum fell and was reset every 13.5±7.6 s (higher than during the manual error correction task, $p=0.004$).

3.2 Subjective evaluation

Data from the questionnaires is summarized in Table 1. All four parameters passed the normality test. There was no significant difference between the two single-task periods for any of the four reported subjective feelings. The dual-task period resulted in the highest subjective feelings of frustration, concentration and difficulty together with the lowest satisfaction. The differences in frustration and task difficulty between the dual-task period and either single-task period were significant ($p<0.01$ for either variable when compared to

Table 1 Results of self-report questionnaires

	Manual error correction		Mental arithmetic		Dual-task	
	mean	st. dev.	mean	st. dev	mean	st.dev.
satisfaction	2.5	1.3	2.8	1.1	1.9	1.3
frustration	2	1.3	1.8	1.2	2.9	1.3
concentration	3.6	0.9	3.8	1.1	4.2	0.7
difficulty	3.2	1	3.3	0.5	4.5	0.7

either single-task period). Participants rated the degree to which divided attention affected their performance as 3.7±0.8 (mean±standard deviation) out of five.

3.3 Physiological measurements

Table 2 shows the differences in physiological responses between baseline (rest) and task as well as between different task periods. The only exception to this is the SCL, which is already measured as deviations from an initial value and is therefore presented as the absolute difference between periods. Three psychophysiological parameters that exhibit particularly significant differences between task types are shown as box plots in Figs. 3 (mean SCL), 4 (LF power) and 5 (final skin temperature). On a box plot, the central line represents the median value, the bottom and top of the box represent the 25th and 75th percentiles, and the whiskers represent the 10th and 90th percentiles.

The following physiological parameters passed the normality test: respiratory rate variance, mean HR, SDNN, RMSSD and HF power.

Table 2 Physiological differences between different task periods. All differences other than mean SCL are presented as percentage of baseline (rest) value. Bold underlined values and asterisks indicate statistical significance: * for $p<0.05$, ** for $p<0.01$ and *** for $p<0.001$. Top row abbreviations: *B* baseline, *M* movement period, *MEC* single-task manual error correction, *MA* single-task mental arithmetic, *D* dual-task period

	M–B	MEC–B	MA–B	D–B	MEC–M	MA–MEC	D–MEC	D–MA
mean SCR frequency	13.1	**122.7***	**165.2***	**227.3***	**109.6**	42.5	104.6	62.1
mean SCR amplitude	3.8	**22.9***	**34.9***	**46.1***	**19.1***	12	23.2	11.2
mean SCL (μS)	0	**7.9****	**12.6*****	**11.2*****	**7.9***	**4.7***	**3.3***	−1.4
mean respiratory rate	**16.6*****	**21.8*****	**17.9*****	**16.8*****	5.2	−3.9	−5	−1.1
respiratory rate variance	**−24.0***	**−39.1*****	22.7	2.9	−15.1	**61.8****	**42.0*****	−19.8
mean HR	**2.6***	1.5	**2.4***	**3.8*****	−1.1	0.9	2.3	1.4
SDNN	−4	**−12.0***	13.6	5.5	−8	**25.6****	**17.5****	−8.1
RMSSD	−9.6	0	10.5	10.2	9.6	10.5	10.2	−0.3
LF/HF ratio	39.3	−9.2	33.5	35.2	**−48.5***	42.7	44.4	1.7
HF power (RSA)	−12.8	19.5	**42.4***	**33.3***	**32.3***	22.9	13.8	−9.1
LF power	7.1	**−4.6***	79.9	62.8	**−11.7***	**84.5****	**67.4***	−17.1
final temperature	0.8	−0.7	**−2.2*****	**−2.7*****	**−1.5***	**−1.5***	**−2.0***	−0.5

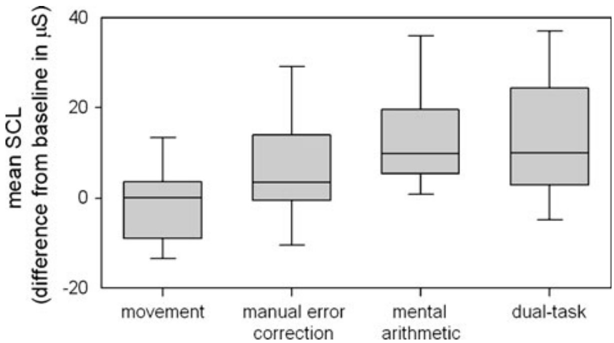

Fig. 3 Mean SCL during different tasks as a percentage of baseline value

4 Discussion

4.1 Psychophysiological responses

All three skin conductance parameters: skin conductance level (SCL), SCR frequency and SCR amplitude significantly increased from baseline during both single- and dual-task periods, but not the movement period, thus indicating a strong influence of mental workload.

Mean respiratory rate significantly increased from baseline during both single-task and dual-task periods as well as during the movement period. Respiratory rate variance significantly decreased from baseline during the movement period and the manual error correction task, but not during the mental arithmetic task or the dual-task period. Since respiratory variability in general has been previously linked to panic and anxiety [16] or boredom [21], differences in respiratory rate variance could be linked to differences in satisfaction or frustration. However, this cannot adequately explain our results since self-report measures found no significant difference between the two single-task periods. A second possible explanation is that artefacts resulting from verbal responses to the mental arithmetic task increased respiratory rate variance despite our best efforts to remove them.

Heart rate significantly increased from baseline during the movement period, the mental arithmetic task and the dual-task period, but not during the manual error correction task. The increase in heart rate during the movement period was not surprising, as the connection between physical activity and heart rate has been well-established in countless studies. The increased heart rate during the mental arithmetic task and during the dual-task period, on the other hand, must be caused by mental activity. Still, it is puzzling why there was no significant increase in heart rate during the manual error correction task.

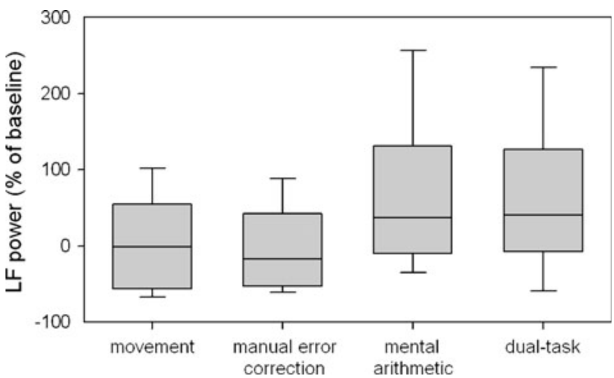

Fig. 4 LF power during different tasks as a percentage of baseline value

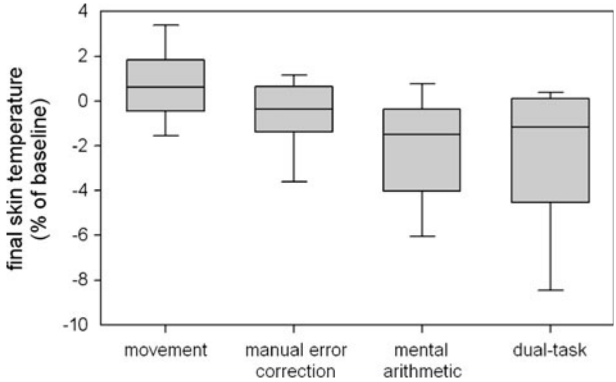

Fig. 5 Final skin temperature during different tasks as a percentage of baseline value

HRV decreased significantly from baseline only during the manual error correction task (and only for SDNN and total LF power). It has been previously shown that, while HRV does decrease as mental effort increases, it increases again if the effort needed for a task increases beyond the capacity of working memory [17]. This explains why heart rate variability did not decrease from baseline during the dual-task period. However, it does not explain why SDNN and total LF power decreased during the manual error correction task but not the manual arithmetic task since self-report measures did not show any significant difference between the two single-task periods.

Skin temperature significantly decreased during both the mental arithmetic task and the dual-task period, but not during the manual error correction task. Since there was no significant self-reported difference between the two single-task periods, the psychophysiological difference is again surprising.

4.2 Influence of divided attention and multiple resources

As shown by results of the self-report questionnaires, participants were aware of the added challenge posed by performing both tasks simultaneously. Their performance was also adversely affected in the dual-task condition: the pendulum fell more often, more incorrect answers were given, and more time was needed for each answer.

While self-report measures found no difference between the two single-task periods, there were several significant psychophysiological differences (see Table 2 and Figs. 3, 4 and 5). It is possible that these differences were caused by different resource requirements. Previous studies have demonstrated that not all psychophysiological responses are sensitive to the same task types and changes in difficulty [1]. Since the two different tasks were performed using different modalities (verbal vs. manual), they may have drawn on separate mental resources and thus also evoked different physiological responses.

Interestingly, psychophysiological measures did not reliably differentiate between single-task and dual-task periods (Table 2 and Figs. 3, 4 and 5) despite significant differences in self-report measures and task success. As a follow-up, we attempted to differentiate between single-task and dual-task periods using MANOVA, thus combining multiple physiological parameters. However, the MANOVA also did not yield a significant difference between the mental arithmetic and dual-task periods. This is troubling, as psychophysiological measures cannot be reliable in multitasking if they cannot show a significant difference between performing a single task and performing two tasks at once.

The lack of a significant difference between the mental arithmetic condition and dual-task periods may have been caused by mental overload: if the mental arithmetic task already uses a majority of mental resources, adding the manual error correction task would only divert resources from the mental arithmetic task instead of using additional resources. Wilson and Russell [33] have shown that psychophysiological measurements can discriminate between normal performance and mental overload. Thus, mental overload is a phenomenon with observable physiological effects. In our study, the two tasks can be expected to at least partially share mental resources since they are both displayed on the same screen. The visual modality can thus be described as a shared resource according to multiple resource theory [32] or as a processing bottleneck that cannot be devoted to multiple tasks at once [24]. Additionally, both performance measures and self-report measures clearly showed that mental overload did occur during the dual-task period. If psychophysiological measurements reflect the total amount of resources used, they would not show whether resources are focused on a single task or divided among several tasks. This requires the assumption that the mental arithmetic task is more resource-intensive than the manual error correction task, which is not supported by self-report measures but is partially supported by psychophysiological measures.

4.3 Discrepancy between performance, self-report and physiological measures

The discrepancies between self-report and psychophysiological measures observed in Sections 4.1 and 4.2 are not specific to our study. Other studies have also failed to find evidence for associations between psychophysiological responses and self-reported emotions [9]. Other psychological assessment methods such as observation of the subject by an independent observer have been shown to correlate much more reliably with psychophysiological responses [28].

Theoretical and practical studies of attention and mental workload have highlighted a multidimensional nature of mental workload [7]. Dissociations between performance, self-report and physiological responses may occur for a number of reasons: responses having different mappings to task demands, differential sensitivity (i.e. gain functions) of responses to task demands, differential effects of background physiological state (which itself is affected by the psychological and physical task demands in addition to individual differences and environmental conditions) etc. Backs [1], for instance, found that different psychophysiological measures responded to different ways of manipulating task difficulty, thus demonstrating dissociation between psychophysiology and task performance. Dissociations between objective and self-report measures of task performance have also been observed [11]. Thus, it is evident that the dissociations between task performance, subjective opinion, and psychophysiology are an unavoidable facet of user state estimation.

4.4 Relevance for human-computer interaction

Our study has demonstrated that, in multimodal human-computer interaction, psychophysiological responses are sensitive not only to interindividual differences, but also exhibit significantly different responses to different tasks even though participants report no significant subjective differences. When participants do feel significant subjective differences, however, these may not be evident in psychophysiological responses. These lessons should be heeded by planners and developers of adaptive, context-aware systems such as those described by [20] or [6].

Psychophysiological measurements can be taken continuously as users of adaptive systems are exposed to different stimuli, different task difficulty levels and different tasks altogether, Various parameters such as SCR frequency and respiratory rate can also be calculated in real time with the same methods regardless of the task being performed. However, the rules and methods for the interpretation of psychophysiological parameters need to be adjusted depending on the task being performed. Different psychophysiological parameters may need to be taken into account during different tasks, for instance [1]. Even small changes such as the addition of background music may require different rules for interpretation of psychophysiology. In multitasking situations such as the one examined by our study, psychophysiological measures could fail altogether as some mental resources become overloaded.

This complexity also suggests that simple interpretation rules such as ‘if this parameter is above a threshold, user is stressed’ may be insufficient to accurately identify psychophysiological states. At the very least, a combination of multiple parameters using methods such as discriminant analysis is likely to be necessary. More sophisticated approaches such as neural networks or Gaussian mixture models may be preferred for greater accuracy.

4.5 Study limitations

Finally, we would like to point out a few limitations of our study. First of all, our findings apply primarily to the four psychophysiological measures included in the study: heart rate, skin conductance, respiration and skin temperature. Though these four are some of the most popular psychophysiological measures currently used in multimodal HCI, they are not the only possibilities. Additionally, all four primarily measure the activity of the autonomic nervous system and thus cannot provide a full image of the user’s psychophysiological state. Expanding the system with measurements such as eye movements, facial electromyography or electroencephalography would make it more robust and potentially allow for better discrimination between different task types. Of course, it must also be noted that measurements such as electroencephalography or facial electromyography could be deemed intrusive by the user and would thus potentially increase the accuracy of the system while decreasing its user-friendliness.

Second, the study does not take differences in individual personalities and subjective factors into detailed account. This is partially due to the study design; interindividual differences have already been previously studied, so our focus was instead on systematic differences between task types. Since the study is a repeated-measures design, systematic differences (or lack of them) between different tasks should not be strongly dependent on individual personality features. Nonetheless, we acknowledge that including more detailed questionnaires about current feelings, the participant’s opinion of his/her task performance and the participant’s general personality type would have allowed for a more accurate interpretation of results. In the same vein, the set of participants is, in a way, limited since all are male and technically experienced. A more varied set of participants would have made the results more applicable to the general population.

Finally, the tasks chosen are not necessarily the most representative for a multimodal HCI environment. The mental arithmetic task represents a cognitively challenging task, but may be more of a ‘laboratory’ task than a real-world task. In a similar vein, while haptic interfaces are becoming increasingly common, they are usually used with different tasks than the one in our study. Nonetheless, despite the somewhat abstract nature of the tasks, we feel that the findings derived from them illustrate the principle that, in HCI, different tasks can evoke different psychophysiological responses despite no systematic difference in subjective responses, and that psychophysiological measurements are less reliable in dual-task conditions due to possible mental overload.

5 Conclusions

While our study demonstrated that psychophysiological measures are sensitive to mental workload caused by different tasks, it demonstrated that they do not always agree with subjective feelings. This may be because different types of tasks utilize different mental resources and thus evoke different physiological responses without the participants' awareness. Additionally, the use of psychophysiological measures in multitasking appears limited since they cannot always differentiate between single-task and dual-task conditions. This may be due to mental overload: after a certain level of workload is reached, the participant is using all available mental resources and is forced to divide them between tasks. Nonetheless, this does not make psychophysiology useless for multimodal human-computer interaction. Rather, system designers should be aware of both its strengths and its limitations. First, as the environment around the user changes, that rules and methods for interpretation of psychophysiological responses need to change as well, even if the user remains the same. Second, in certain conditions psychophysiological responses may not accurately reflect the user's actual mood.

Acknowledgments The authors would like to thank the volunteers who participated in this study. The work was funded by the EU Information and Communication Technologies Collaborative Project MIMICS grant 215756 and additionally supported by the Slovenian Research Agency.

References

1. Backs RW (1997) Psychophysiological aspects of selective and divided attention during continuous manual tracking. Acta Psychol 96:167–191
2. Backs RW, Lenneman JK, Wetzel JM, Green P (2003) Cardiac measures of driver workload during simulated driving with and without visual occlusion. Hum Factors 45:525–538
3. Boiten F (1993) Component analysis of task-related respiratory patterns. Int J Psychophysiol 15:91–104
4. Chen JYC (2009) Concurrent performance of military and robotics tasks and effects of cueing in a simulated multi-tasking environment. Presence–Teleop Virt 18: 1–15
5. Collet C, Averty P, Dittmar A (2009) Autonomic nervous system and subjective ratings of strain in air traffic control. Appl Ergon 40:23–32
6. Ćosić K, Popović S, Jovanovic T, Kukolja D, Slamić M (2007) Physiology-driven adaptive VR system: Technology and rationale for PTSD treatment. Annual Review of CyberTherapy & Telemedicine 5:179–191
7. Derrick WL (1988) Dimensions of operator workload. Hum Factors 30:95–110
8. Detenber BH, Simons RF, Bennett GG (1998) Roll 'em!: the effects of picture motion on emotional responses. J Broadcast Electron 42:113–127
9. Feldman PJ, Cohen S, Lepore SJ, Matthews KA, Kamarck TW, Marsland AL (1999) Negative emotions and acute physiological responses to stress. Ann Behav Med 21:216–222
10. Gwizdka J (2010) Distribution of cognitive load in web search. J Am Soc Inf Sci Technol, in press
11. Horrey WJ, Lesch MF, Garabet A (2009) Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. J Safety Res 40:7–12
12. Kahneman D (1973) Attention and Effort. Prentice Hall, Englewood Cliffs
13. Kahol K, Panchanathan S (2008) Neuro-cognitively inspired haptic user interfaces. Multimed Tools Appl 37:15–38
14. Liu C, Agrawal P, Sarkar N, Chen S (2009) Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. Int J Hum-Comput Int 25:506–529
15. Liu J, Gwizdka J, Liu C, Belkin NJ (2010) Predicting task difficulty for different task types. In: Proceedings of the 73rd Annual Meeting of the American Society for Information Science & Technology. Pittsburgh, PA, in press
16. Martinez JM, Kent JM, Coplan JD, Browne ST, Papp LA, Sullivan GM et al (2001) Respiratory variability in panic disorder. Depress Anxiety 14:232–237
17. Mulder G, Mulder LJ, Meijman TF, Veldman JB, van Roon AM (2000) A psychophysiological approach to working conditions. In: Backs R, Boucsein W (eds) Engineering psychophysiology: Issues and applications. Lawrence Erlbaum Associates, Mahwah, pp 139–159

18. Nikula R (1991) Psychological correlates of nonspecific skin conductance responses. Psychophysiology 28:86–90
19. Novak D, Mihelj M, Munih M (2010) Psychophysiological responses to different levels of cognitive and physical load in haptic interaction. Robotica, in press
20. Octavia JR, Raymaekers C, Coninx K (2010) Adaptation in virtual environments: conceptual framework and user models. Multimed Tools Appl, in press
21. Ohsuga M, Shimono F, Genno H (2001) Assessment of phasic work stress using autonomic indices. Int J Psychophysiol 40:211–220
22. Parasuraman R, Cosenzo KA, De Visser E (2009) Adaptive automation for human supervision of multiple uninhabited vehicles: Effects on change detection, situation awareness and mental workload. Mil Psychol 21:270–297
23. Parkes AM, Coleman N (1990) Route guidance systems: a comparison of methods of presenting directional information to the driver. In: Lovesey J (ed) Contemporary Ergonomics 1990. Taylor & Francis, London, pp 480–485
24. Pashler H, Johnston JC (1998) Attentional limitations in dual-task performance. In: Pashler H (ed) Attention. Psychology Press, East Sussex, pp 155–189
25. Phillips AC, Carroll D, Hunt K, Der G (2006) The effects of the spontaneous presence of a spouse/partner and others on cardiovascular reactions to an acute psychological challenge. Psychophysiology 43:633–640
26. Posner MI, Sandson J, Dhawan M, Shulman GL (1989) Is word recognition automatic? A cognitive anatomical approach. J Cogn Neurosci 1:50–60
27. Ritvanen T, Louhevaara V, Helin P, Väisänen S, Hänninen O (2006) Responses of the autonomic nervous system during periods of perceived high and low work stress in younger and older female teachers. Appl Ergon 37:311–318
28. Schwerdtfeger A (2004) Predicting autonomic reactivity to public speaking: don't get fixed on self-report data! Int J Psychophysiol 52:217–224
29. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996) Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 17:354–381
30. Veltman JA, Gaillard AW (1998) Physiological workload reactions to increasing levels of task difficulty. Ergonomics 41:656–669
31. Wester AE, Böcker KB, Volkerts ER, Verster JC, Kenemans JL (2008) Event-related potentials and secondary task performance during simulated driving. Accident Anal Prev 40:1–7
32. Wickens CD (2002) Multiple resources and performance prediction. Theor Issues Ergon Sci 3:159–177
33. Wilson GF, Russell CA (2003) Operator functional state classification using multiple psychophysiological features in an air traffic control task. Hum Factors 45:381–389
34. Xie B, Salvendy G (2000) Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. Work Stress 14:74–99
35. Zhang T, Kaber D, Hsiang S (2010) Characterisation of mental models in a virtual reality-based multitasking scenario using measures of situation awareness. Theor Issues Ergon Sci 11:99–118

Domen Novak is a PhD candidate at the Laboratory of Robotics and Biomedical Engineering, University of Ljubljana, Slovenia. He received his B.Sc. degree in electrical engineering from the University of Ljubljana in 2003. His research interests include human-robot interaction, virtual reality, and biomedical signal processing. He is currently working toward a PhD on the use of psychophysiological measurements in rehabilitation robotics.

Matjaž Mihelj received the M.Sc. and PhD degrees (1999 and 2002) in electrical engineering from the University of Ljubljana. Since 2009 he has been affiliated with the university as an associate professor. In 2001 he was a visiting researcher at Sendai University, Japan. From May 2005 to April 2006 m he was a visiting researcher at the Automatic Control Lab, ETH Zurich with the Marie Curie Intra-European fellowship. His rehabilitation robotics research performed at ETH Zurich was awarded with the Swiss Technology Award. His current research interests are in man/machine interaction, haptic interfaces as well as modeling and control of biological systems with the main focus on restoration of lost functions in people with disabilities.

Marko Munih holds a PhD degree in electrical engineering from the University of Ljubljana. His research interests include functional electrical stimulation, measurement and control, biomechanics and robotics. In 1995/1996, he was a visiting scientist at University College London. He is now full professor and head of the Laboratory for Robotics and Biomedical Engineering and was head of the Department for Measurements at the university of Ljubljana in 2003–20077. In the last 5 years, he was and still is principal investigator of five EU-RTD projects and cooperated in six other projects. Professor Munih received the National Zois research award in 2002.