Measuring Stress and Cognitive Load Effects on the Perceived Quality of a Multimodal Dialogue System

Andreea Niculescu
niculescuai@ewi.utwente.nl
+31534894654

Yujia Cao
y.cao@ewi.utwente.nl
+31534892826

Betsy van Dijk
e.m.a.g.vandijk@ewi.utwente.nl
+31534893781

Anton Nijholt
a.nijholt@ewi.utwente.nl
+31534893686

HMI Group, University of Twente
Postbus 217, 7500 AE, Enschede, The Netherlands

ABSTRACT

In this paper we present the results of a pilot study investigating the impact of stress and cognitive load on the perceived interaction quality of a multimodal dialogue system for crisis management. Four test subjects interacted with the system in four differently configured trials aiming to induce low/high levels of stress and cognitive load. Physiological sensors and subjective ratings were collected to measure the level of stress and cognitive load. After each trial the subjects filled in an evaluation questionnaire regarding the system interaction quality. In the end we conducted an in-depth interview with each subject. The trials were recorded with a webcam to facilitate the behaviour analysis. Results showed that both factors had an influence on the way subjects perceived the interaction quality, whereas the cognitive load seems to have a higher impact. Further quantitative experiments are needed in order to validate the results and quantify the weight of each factor.

Author Keywords

Multimodal conversational interactions, qualitative evaluations, behavior analysis.

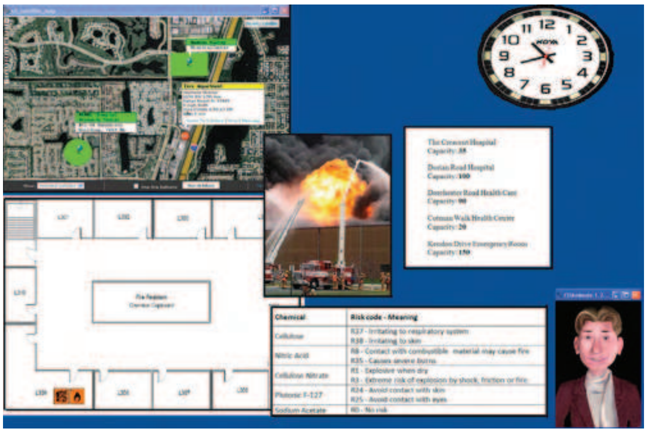
ACM Classification Keywords

H.5.1 Multimedia information systems: Animations, Audio input/output, Evaluation/methodology, H.5.2 User Interfaces: Evaluation/methodology, Graphical user interfaces, Natural language, Voice I/O.

INTRODUCTION

The quality assessment of interactive systems is a complex construct of interdependent factors relying on system design and performance, and user perception. Among these factors are the cognitive load and stress experienced by users during the interaction [1].

In the literature these two factors are often summarized and measured together under the global concept of 'cognitive demand'. There are surely no doubts that these two factors are related but their relationship is not exclusive: stress can be caused not only by a highly loaded cognitive task but also by frequent input recognition mistakes or poor sound quality [2], whereas a highly loaded cognitive task would be perceived as stressful only in situations considered as exceeding available resources. [3]. Therefore, our long term research goals are to investigate whether stress and cognitive load can be successfully manipulated and measured separately. Also, we are interested in the impact these two factors might have on the perceived interaction quality of a multimodal dialogue system.



The Crescent Hospital
Capacity: 35

Dorian Road Hospital
Capacity: 100

Dorchester Road Health Care
Capacity: 90

Cotman Walk Health Centre
Capacity: 20

Kendon Drive Emergency Room
Capacity: 150

Chemical	Risk code - Meaning
Cellulose	R37 - Irritating to respiratory system R38 - Irritating to skin
Nitric Acid	R8 - Contact with combustible material may cause fire R35 - Causes severe burns
Cellulose Nitrate	R1 - Explosive when dry R3 - Extreme risk of explosion by shock, friction or fire
Pluronic F-127	R24 - Avoid contact with skin R25 - Avoid contact with eyes
Sodium Acetate	R0 - No risk

Figure 1. System screen shot.

METHODS

Since crises represent situations in which people experience high levels of stress and cognitive load they offer perfect test environments for our experiment. Accordingly, we developed a small prototype of a multimodal dialogue system for crisis management. The system has attached an embodied conversational agent representing the crisis manager (see figure 1). The crisis manager provides information about a crisis event, such as event description, geographical maps, available rescue resources and estimated number of victims. Users can interact with the system using speech or mouse clicks and receive information in the form of text, speech, images or videos.

Experiment Design

The experiments were configured in four trials. Each trial combined several parameters in order to achieve low/high stress (S) and cognitive load (CL) (see table 1).

Trial no.	Factor combination
1	low CL / low S
2	low CL / high S
3	high CL / high S
4	high CL / low S

Table 1. Combinations of cognitive load and stress levels per trial.

To manipulate the stress factor we used six parameters, such as background noise, speech speed, speech length, time limitation, simulated recognition mistakes and dramatic event description. For the cognitive load manipulation we varied task complexity and presentation format. A detailed description of the trials configuration can be found in [8]. A compulsory break of a minimum of five minutes was placed between the trials to lower the stress level.

Participants

Four male test subjects, aged between 24 and 30, participated in the experiment. All of them had a technical background, but vague or no idea about crisis management. Also, they were not particularly familiar with the use of multimodal dialogue systems; only one subject had spoken to a computer before. Due to the fact that our pilot study was performed with a small number of participants we decided to perform the trials in the same order for all four subjects (see table 1). However, in future we plan to counterbalance the trial order among participants to avoid learning effects or other biases that might arise by being exposed to a certain factor combination before the others.

Measurements

To collect measurements we used physiological sensors, questionnaires, a two-session in-depth qualitative interview and behavior analysis.

- **Physiological sensors**. We used the heart rate variability (HRV) [4, 5] as an indicator for the cognitive load and the galvanic skin response (GSR) [6, 7], as a measurement of the stress.
- **Questionnaires**. The NASA (TLX) questionnaire was used to control the results collected from the physiological sensors. TLX contains six workload-related factors: mental, physical and temporal demands, own performance, effort and frustration. Statements concerning subjects' concentration and degree of tiredness, ease of the system use, overall system quality and degree of understanding between subjects and the system were added to the TXL questionnaire. The factors were rated on a 20-point scale. The questionnaires were filled in after each trial.
- **In-depth interview.** Each test subject was interviewed based on a qualitative open questionnaire. The interviews were recorded on tape and manually transcribed. The interview aimed to explore relationships between the induced level of stress and cognitive load, and the overall interaction quality. The interview was performed in two sessions: before and after the experiment. In the first session subjects were interviewed about their expectations and background knowledge concerning the experiment topic (crisis management) and the use of multimodal conversational interfaces. In the second session subjects were asked about expectation fulfilling, problems encountered during the interaction, system comprehensibility and transparency, content informativeness, information presentation and interaction easiness. In the end the subjects could make additional comments if they had any.
- **Behavior analysis.** From the log files and videos analysis we extracted various parameters concerning response competition time, reaction time, number and type of errors, total number of words, verbal hesitations, breaks and mispronunciations. The videos also facilitated the qualitative analysis of several other behavioral cues, such as speaking style (polite, rude, key-words vs. sentences) verbal and non-verbal reactions to system errors or increased task difficulties, gestures and gaze.

RESULTS AND DISCUSSION

Our results showed that our manipulation was successful for the cognitive demand, and only partly for the stress.

Also, both stress and cognitive load were better indicated by subjective rating than by physiological measurements (for more details see [8]).

In the first trial the planned manipulation was altered by a "first impression" effect, the trial achieving a much higher level of stress than expected[1]. Thus, the effect intended for trial 1 (low CL/low S) was instead achieved in trial 2.

[1] In the future we plan to add a "base-line" trial at the experiment's begin in order to avoid getting unplanned manipulations.

Accordingly, three out of four subjects ranked the interaction quality with the system for trial 2 as being the best. Trial 3 had the highest level of stress and cognitive load as perceived by subjects. However, only one subject rated the interaction quality in this trial as being the lowest; the other subjects chose instead trial 4 (high CL/low S) as having the lowest system interaction quality. This finding might lead to the conclusion that the cognitive load could have a bigger impact on the perceived interaction quality compared with the stress. Nevertheless, repeated measurements with a higher number of subjects are needed to confirm this assumption.

Further, the interview showed that subjects considered the system as presenting relevant and informative content in a clear and systematical way, a fact that decreased the cognitive load, according to one subject. They did not encounter communication problems, except for one subject and were pleasantly surprised to know how to handle the system right from the beginning. The interaction with the system was in general perceived as being easy and 3 subjects enjoyed it. However, the system was criticized as lacking basic functionalities (such as zooming in the crisis maps or help options) and being not transparent and not flexible enough: during the third trial most subjects were unsure how to answer the system's questions and were not allowed to return to previous conversation stages to ask for clarifications. Another negative point was the synthesized voice of the agent reminding the subjects unpleasantly that they were talking with a machine.

Analyzing the subjects' behavior we observed that they acted congruently to the golden rule "treat others as you want to be treated"; for example, one subject who expressed the wish to be treated politely used polite markers, such as "thank you" and "please" during the entire interaction (even in stress loaded trials); another subject wishing the system to present only facts and no other redundant information used only keywords or extremely short sentences. In general, subjects displayed a very different behavior in terms of performance and reactions to stress: some expressed their frustration using loud verbal expressions or showing a constant "joke" attitude; others became impatient and started clicking the mouse button to "increase" the interaction speed; some remained apparently calm, showing their response to stress only through frowning. Interestingly, most of the speech disfluencies, breaks and errors were made in the low stress condition and mostly by two of the subjects. These two subjects had in general a poor performance completing the trials and gave lower quality rankings, as compared with the other two subjects. The verbal response time values were very different among the subjects, but showed a common trend: they were lower in high stress conditions and higher in high cognitive load circumstances.

CONCLUSIONS

We performed a pilot study on the way users perceive the quality of the interaction with a multimodal dialog system for crisis management while being exposed to stress and cognitive load variations. Our manipulation was successful for the cognitive demand, and only partly for the stress. We encountered difficulties in achieving accurate stress manipulations, as stress appears to be a highly complex phenomenon to which humans respond very differently. For our experiment we used physiological sensors, questionnaires, qualitative interviews and behavioral analyses. Since objective measurements methods, (i.e. physiological sensors) could not provide meaningful results additional quantitative investigations are required to determine whether stress can be measured apart from cognitive load. Also, further analyses are needed to validate and weight the factors' impact on the interaction quality assessment.

ACKNOWLEDGMENTS

This research was funded by the EC Artemis project on Human-Centric Design of Embedded Systems (SMARCOS, Nr. 100249).

REFERENCES

1. Moeller, S.: Quality of telephone-based spoken dialogue systems, Springer, New York, 2005.
2. Kryter, K.D.: The handbook of hearing and the effects of noise: Physiology, psychology, and public health. Academic Press New York, 1994.
3. Lazarus, R.S.: Theory based stress measurement. *Psychological Inquiry*, 1 (1990), 3-13.
4. Scerbo, M.W., Freeman, F.G., P.J Mikulka., R. Parasuraman, and F. Di Nocero,: The efficacy of psycho-physiological measures for implementing adaptive technology. TP-2001-211018, (2001).
5. Wilson, G.F. and F.T. Eggemeier: Psycho-physiological assessment of workload in multi-task environments. In: Damos, D.L. (ed.): Multiple-task performance. CRC Press, 1991.
6. Verwey, W.B. and H.A Veltman.: Detecting short periods of elevated workload. A comparison of nine workload assessment techniques. *Applied Experimental Psychology, 2* (1996), 270-285.
7. Boucsein, W., Haarmann, A., Schaefer, F.: Combining Skin Conductance and Heart Rate Variability for Adaptive Automation During Simulated IFR Flight. In: Harris, D. (ed.) HCII 2007 and EPCE 2007. LNCS, 4562 (2007), Springer, Heidelberg, 639–647..
8. Niculescu, A.I., Y. Cao and A. Nijholt Manipulating Stress and Cognitive Load in Conversational Interactions with a Multimodal System for Crisis Management Support, in Development of Multimodal Interfaces: Active Listening and Synchrony, A. Esposito, N. Campbell, C. Vogel, A. Husain and A.Nijholt (eds), LNCS, vol. 5967 (2010), Springer, Heidelberg, 134-148.