Stress and Cognitive Load in Multimodal Conversational Interactions

Andreea Niculescu, Yujia Cao, and Anton Nijholt

Human Media Interaction, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{niculescuai,y.cao,a.nijholt}@utwente.nl

Abstract. The quality assessment of multimodal conversational interactions is determined by many influence parameters. Stress and cognitive load are two of them. In order to assess the impact of stress and cognitive load on the perceived conversational quality it is essential to control their levels during the interaction. Therefore we present in this paper preliminary experiments carried out to determine the circumstances in which low/high levels of stress respectively cognitive load are achieved while interacting with the system. Different levels are manipulated by varying task difficulty (information complexity, task load, and simulated speech recognition errors), information presentation (modality usage, spatial organization and temporal order of information items) and time pressure. Heart rate variability (HRV) and galvanic skin response (GSR) as well as subjective judgments in the form of questionnaires are deployed to validate the induced stress and cognitive levels. Methods and preliminary results are presented.

Keywords: Stress, cognitive load, multimodal conversation.

1 Introduction

The new generation of multimodal conversational interfaces enable users to communicate with computer systems using a wide range of input/output modalities, such as speech, text, images, gestures etc. Thus the quality assessment of such interfaces is a complex process involving the analysis of a high number of influencing factors relating to interaction parameters and subjective user perceptions [1]. Among these factors are the cognitive load and stress experienced by users while interacting with the interface.

Cognitive load is often described as the degree of concentration required for a person to solve problems or to complete tasks at a given time [2]. The term, referred in the literature as 'cognitive effort' [3] or 'cognitive factor' [4] is often associated with the factor 'stress'. Both factors are summarized under the global concept of 'cognitive demand', which encloses the perceived level of effort needed to use the system and user feelings arising from this effort [5].

There are no doubts that cognitive load and stress are related. When the load reaches a certain level of demand people unconsciously appraise their abilities to meet the challenge: only if the situation is considered as exceeding the available resources stress would appear. This theory, formulated by Lazarus et al. [6] partially explains

why a tense situation might be perceived as stressful by one person, but not by another.

However, our assumption is that, even if related, these two factors can occur independently, i.e. there is no compulsory relationship between them. This means, they might have a different impact on the perceived conversational quality and consequently, they should be identified and measured separately. Hence, we propose in this paper an experiment meant to determine the circumstances in which low/high levels of stress, respectively cognitive load are achieved while interacting with a multimodal conversational system for crisis management.

2 Experiment Design

A first system simulation was build using the CSLU toolkit[1]. The simulation is meant to serve as a support tool for crisis managers, being currently under development. Crises information is presented in the form of speech, text and images by a virtual assistant, attached to the system. Users can interact with the assistant using speech and mouse pointing.

For the current experiment we designed the following crisis scenario: an explosion occurred in a chemical research lab and set on fire an entire floor; the crisis manager has to take essential decisions with the help of his virtual assistant.

The crisis manager role was played by test users who interacted with the system in four experimental trials. The trials were aiming to realize combinations of low/high stress and low/high cognitive load conditions in a 2x2 factor matrix. Each trial consists of three stages: first, the virtual assistant presents the current crisis situation, using narrative and assisting photos, maps and text; then the crisis manager has to find on an interactive map addresses to which rescue workers, fire fighters or wounded victims will be delivered; alternatively he has to memorize important event facts and insert them in a crisis report; finally, the crisis manager gets chemical description sheets to identify dangerous chemicals that have to be immediately removed by firemen in order to avoid further damages.

We achieved a stress level manipulation by combining six factors: background noise, speech speed, speech length, time limitation, simulated recognition mistakes and event description. In the low stress conditions, the virtual assistant presents calmly the crisis event using a clear voice with normal speed. The crisis situation is described as being under control; subjects are not urged to speed up their performance. In the high stress conditions however, noises (e.g. white noise, ambulance sound) are played in the background in order to induce stress [7]. The virtual assistant talks fast, using short sentences and an urgent tone. The crisis situation is described to be dramatic; subjects are put under time stress by being constantly reminded to make quick decisions; a simulated speech recognition mistake was also built in one of the trials.

The cognitive load level was manipulated by two factors: the task complexity, the presentation format. Task complexity variations were put into effect for addresses identification: in the trials with low cognitive load the subjects had to locate given

[1] http://cslu.cse.ogi.edu/toolkit/

addresses on a map by clicking on the street names; in the trials with higher load the subjects were required not only to identify but also to select the optimal address, according to several factors that needed careful analysis (e.g. hospital capacity, distance to the chemical lab, number of victims).

Variations in the presentation format were chosen for the chemical selection task. 'Well'-designed and 'badly'-designed information sheets were applied to achieve low and respectively high cognitive load conditions. Both sheets use a table to present the chemicals and their risk descriptions. The difference between the 'well'- and 'badly'-designed sheet lie in the way the information is spatially organized: the 'well'-designed sheet provides integrated chemicals and risk descriptions in a natural 'row-by-row' sequence facilitating the subjects 'scan'-reading; in contrast, the 'badly'-designed sheets provide numerical codes that links the chemicals to their corresponding risk descriptions summarized outside the table. As a consequence, the 'badly'-designed sheet requires additional mental effort causing a split-resource effect and an increase of cognitive load [8].

3 Measurements and Experiment Setup

In order to assess the stress and cognitive load level during the conversational interaction, we used physiological measurements and subjective reports.

Physiological measurements include heart rate variability (HRV) and galvanic skin response (GSR). Previous studies have shown that certain components of HVR exhibit systematic and reliable relationships with the mental demands of the task. In the frequency domain, higher levels of cognitive workload have been associated with decreased power in the 0.10 Hz band (LF) [9, 10]. Skin conductance response (SCR) is a measure that is traditionally associated with workload and especially with arousal states accompanied by mental effort and emotions. Higher workload normally yields higher number of responses (or longer SCR intervals) [11, 12].

Subjective reports were collected using NASA task load index (TLX) questionnaire. TLX contains six workload-related parameters: mental, physical and temporal demands, own performance, effort and frustration. We added to TLX questionnaire few more statements regarding subjects' concentration and tiredness degree, system's ease of use and understanding degree between subjects and system. A 20-level scale was used for rating the factors.

The experiments were performed using the Wizard-of-Oz technique: the speech recognition module was replaced by a human operator in order to ensure a controlled interaction. Four subjects participated in a first pilot study. After entering the lab and taking a seat the subjects were kindly asked to remain relaxed while the physiological sensors were applied by one of the experimenters. Before starting the trials a physiological baseline was recorded for 5 minutes. Afterwards the subjects received a brief introduction of the experiment and performed 4 experimental trials in the following predetermined order: T=00, T=01, T=11, T=10[2]. All trials were recorded

[2] T stands for the trial number; the sequences '00', '01','11','10' represent the factor combinations and their intensity: the first number stands for 'cognitive load', the second number for 'stress'; the values represent the intensity: 1=high, 0=low; for ex. T=01 means: trial with lower cognitive load and higher stress value.

using a webcam. A 5-minute break was placed between the trials. After each trial the subjects were asked to fill in the enhanced TXL questionnaire.

4 Results

The first two trials, 1and 2 (T=**0**0, T=**0**1), were designed to have a lower cognitive load level compared with the last two trials, 3 and 4 (T=**1**1, T**1**0). The results gathered from the questionnaires confirmed that test subjects indeed perceived the first two trials as being less mentally demanding, less hard to accomplish and requiring a lower degree of concentration compared with the last two trials. Subjects felt they were more successful performing the tasks and perceived the interaction with the system as easier, less tiring and less physically demanding. An interesting observation could be made comparing trial 1 (T=00) with trial 2 (T=01): even if both were designed to have similar cognitive levels, the first one was perceived to some extend as being more demanding and requiring more concentration. This finding might be explained by the fact that in the first trial the subjects were not familiar with the system; so they did not have a similar experience to compare with.

Regarding the stress factor we tried to induce a lower level of stress in the trials 1 and 4 (T=0**0**, T=1**0**) compared with a higher level in trials 2 and 3 (T=0**1**, T=1**1**). However, excepting the perception of task rushing which was considered indeed higher for trials 2 and 3, subjects felt more tense, annoyed, irritated, discouraged, insecure during the trials 1 and 3 compared with 2 and 4 (especially trial 3 had the highest negative values). This means that our expectations regarding the manipulation of stress level was successful only for trial 3. Trial 1 surprisingly achieved a higher stress level than expected. This might be again caused by the 'first impression' effect. The physiological measurements confirmed partly a similar learning effect achieved during the trials: the HP (heart rate) and LF (the HRV in the frequency domain) values showed a decreasing trend being higher for the trial 1 and falling off continuously for the following trials. GSRN (number of skin responses per minute) and GSL (the tonic level of the skin conductance) showed the same effect for two of the subjects. One subject's stress level seemed to be caused rather by the cognitive load level rather than by the planned manipulation: the values showed that the subject was more stressed when the task was more difficult. The values for the fourth subject did not deliver any meaningful results.

5 Conclusion

Our results showed that cognitive load and stress were better indicated by subjective reports than by physiological measurements. The cognitive load could be better manipulated compared to the stress, a fact that is not surprising, since stress if often caused by mentally demanding situations. The learning effect visible in the measurements disturbed the manipulation of the expected low/high levels. This effect might be weakened by a training session performed before the experiment's start.

In the future we plan to retrieve performance metrics from the video recording, to analyze them in details and to perform experiments with a larger number of subjects.

The interpretation of physiological data can be improved by measuring particular tasks inside each trial rather than using the whole trial. Further, enlarging the variance in cognitive load and stress between trials might also enhance the effectiveness of physiological measurements since their sensitivity is limited to minor variations.

In summary, the current experimental design can be considered as a good starting point for forthcoming investigations concerning the effects of stress and cognitive load on the conversational quality assessment.

Acknowledgments

This research is done in the framework of the Interactive Collaborative Information System (ICIS) Project, sponsored by the Dutch Ministry of Economic Affairs, under grant nr: BSIK03024. We thank all test participants for their effort and time.

References

1. Moeller, S.: Quality of telephone-based spoken dialogue systems, Springer, New York (2005)
2. Oviatt, S.: Human-centred design meets cognitive load theory: designing interfaces that help people think. In: Proc. of the 14th annual ACM international conference on Multimedia, Santa Barbara (2006)
3. Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., Stentiford, F.W.: Identifying salient usability attributes for automated telephone services. In: Proc. of the 3rd Conf. on Spoken Language (ICSLP), Yokohama, Japan, (1994)
4. Jack, M.A., Foster, J.C., Stentiford, F.W. M.: Intelligent dialogues in automated telephone services. In: Proc. 2nd Int. Conf. on Spoken Language Processing, (ICSLP), Banff, Canada (1992)
5. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech and System Interfaces (SASSI). Natural Language Engineering, 6(3–4), pp: 287–303, (2000)
6. Lazarus, R.S.: Theory based stress measurement. Psychological Inquiry, 1, 3–13, (1990)
7. Kryter, K.D.: The handbook of hearing and the effects of noise: Physiology, psychology, and public health. Academic Press New York (1994)
8. Chandler, P., Sweller, J.: Cognitive Load Theory and the Format of Instruction. Cognition and Instruction, Vol. 8, pp. 293–332 (1991)
9. Scerbo, M.W., Freeman, F.G., Mikulka, P.J., Parasuraman, R., Di Nocero, F.: The efficacy of psychophysiological measures for implementing adaptive technology. TP-2001-211018, (2001)
10. Wilson, G.F., Eggemeier, F.T.: Psychophysiological assessment of workload in multi-task environments. In: Damos, D.L. (ed.): Multiple-task performance. CRC Press (1991)
11. Verwey, W.B., Veltman, H.A.: Detecting short periods of elevated workload. A comparison of nine workload assessment techniques. Applied Experimental Psychology, Vol. 2, pp. 270-285 (1996)
12. Boucsein, W., Haarmann, A., Schaefer, F.: Combining skin conductance and heart rate variability for adaptive automation during simulated IFR flight. LNCS 4562, pp. 639–647 (2007)