Utilizing Pupil Diameter to Estimate Cognitive Load Changes During Human Dialogue: A Preliminary Study

Andrew L. Kun[1)], Zeljko Medenica[1)], Oskar Palinko[1)], Peter A. Heeman[2)]

[1)] University of New Hampshire
Durham, NH 03824
{first.last}@unh.edu

[2)] Oregon Health & Science University
Beaverton, OR 97006
heemanp@ohsu.edu

ABSTRACT

In-vehicle spoken dialogue systems are gaining in popularity. However, it is not always clear which system behaviors might result in increased driver cognitive load, which in turn could have negative safety consequences. In this paper we explore the use of pupil diameter to help in the evaluation of the effects of different dialogue behaviors on the cognitive load of the driver.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Eye tracking, pupillometry, cognitive load, driving simulator, dialogue.

1. INTRODUCTION

When developing a spoken dialogue system (SDS) for interactions with in-vehicle devices, designers have to decide on a number of characteristics of the SDS. These include the types of utterances to use, the timing, pace and volume of utterances, as well as how to handle switches between topics. We propose that the design process can be based on the characteristics of human-human dialogues, specifically those in which one conversant is operating a simulated vehicle.

One problem with using human-human dialogues is that we are likely to encounter a multitude of behaviors. Which ones should an in-vehicle human-machine interface (HMI) emulate? We expect that different dialogue behaviors will result in different levels of cognitive load experienced by the driver. We propose using human behaviors that do not unduly increase the cognitive load of the driver. One physiological estimate of cognitive load is pupil diameter. Unlike driving performance, we expect that pupil diameter will be sensitive enough to provide insight into how the different (rapidly changing) behaviors influence cognitive load.

In this preliminary study, we want to determine if we can use pupil diameter to detect major changes in cognitive load during a less-structured verbal task. We feel that less-structured verbal tasks are more representative of future HMI interaction than highly structured tasks (e.g. question-answer tasks). Furthermore, less-structured dialogues are likely to result in more complex dialogue behaviors, accompanied by more complex changes in cognitive load, and thus pupil diameter, than highly structured tasks. A key research question is how pupil diameter changes might be used to estimate the cognitive load related to behaviors that conversants employ in less-structured dialogues.

In this paper we will determine if we can detect differences in cognitive load between times when the driver is engaged in a verbal game with a remote conversant, and after the game finishes. Our hypothesis is that, once a game finishes, drivers will experience reduced cognitive load, and that this will be reflected in decreased pupil diameter.

2. BACKGROUND

In experiments conducted by Iqbal et al. [1; 2] participants performed manual-visual tasks in front of a computer screen. The authors found that the percent change of pupil size (PCPS) correlated well with the mental difficulty of the task. More complex tasks resulted in higher values of PCPS compared to easier tasks.

Schwalm et al. [9] conducted a driving simulator study in which study participants performed the standardized lane change task [5] and an additional visual search task. The authors explored the relationship between driving performance, and the index of cognitive activity (ICA). The ICA is calculated based on the characteristics of minute dilations of the pupil [4]. Driving performance was estimated using the mean difference between the path followed by a participants vehicle and a so-called optimal path. The study found that the ICA correlates well with driving performance: when the additional visual task was introduced, driving performance decreased and the ICA increased.

Both Iqbal and Schwalm used a high precision head-mounted eye tracking system (EyeLink 2). While head-mounted eye trackers are useful for precise eye measures, they can affect the results of the experiments [4]. Thus researchers have turned to remote eye tracking to estimate the cognitive load of the driver. Recarte and Nunes [8] used remote eye tracking in a naturalistic driving experiment in which participants performed secondary mental tasks. Pupil diameter measures showed differences between secondary task and no secondary task conditions.

Recently, Klingner et al. [3] reported on estimating cognitive load using remote eye tracking in front of a computer screen. Cognitive load was manipulated by instructing participants to perform tasks requiring mental multiplication, digit sequence repetition and aural vigilance. The authors concluded that using remote eye

AutomotiveUI'11, November 29-December 2, 2011, Salzburg, Austria
Adjunct Proceedings

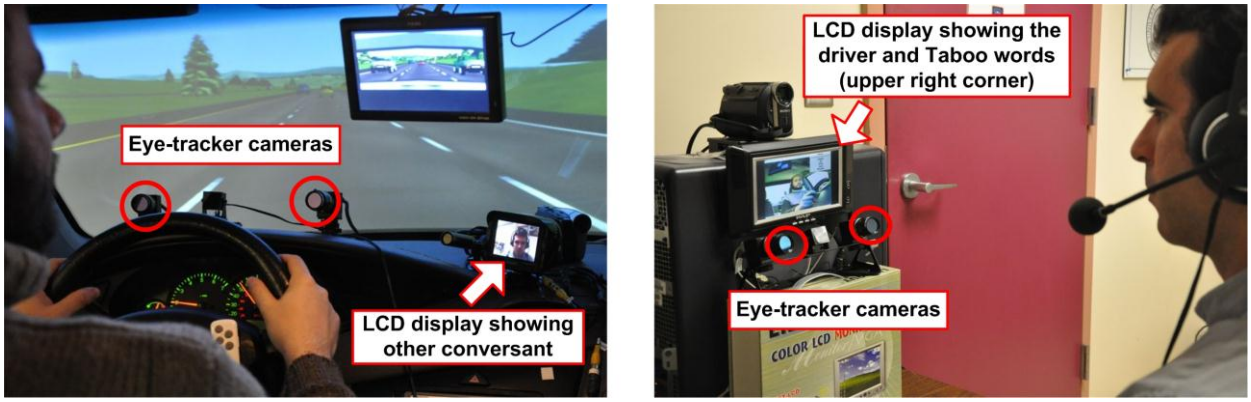

Figure 1 Driver and other conversant.

tracking is a viable way to measure pupil diameter for estimating cognitive load. Building partly on the results of that work, in past work we used a remote eye tracker in a driving simulator experiment to estimate the cognitive load of the driver while he is engaged in a spoken dialogue with a remote conversant [6]. Our pupillometric data indicated that cognitive load changes rapidly (within seconds) between different parts of the dialogue. Specifically, our participants played the last-letter game, in which participants have to utter a word that starts with the last letter of the word uttered by the other participant. During this game the driver's cognitive load estimate was higher when it was the driver's turn to speak than when it was the remote conversant's turn to speak. In contrast to this prior work, the current paper reports on results obtained when the conversants engage in a less-structured verbal game. This is an important distinction, as less-structured dialogues are likely to result in more complex dialogue behaviors than highly structured tasks. For example, in the last-letter game conversants took turns saying one word at a time, and they almost never interrupted each other. In contrast, in the current work, conversants used a variety of dialogue behaviors, for example those related to utterance delivery, such as pausing, fillers, and repetition disfluencies.

3. EXPERIMENTAL SETUP

In our experiment pairs of subjects (the *driver* and the *other conversant*) are engaged in a spoken dialogue. Additionally, the driver also operates a simulated vehicle.

3.1 Equipment

The driver and other conversant (see Figure 1) communicated using headphones and microphones. Their communication was supervised by the experimenter and synchronously recorded as a 48000 Hz mono signal.

The driver operated a high-fidelity driving simulator (DriveSafety DS-600c) with a 180º field of view, realistic sounds and vibrations, a full-width cab and a motion platform that simulates acceleration and braking. We recorded pupillometric data using a SeeingMachines faceLab 5.0 stereoscopic eye tracker mounted on the dashboard in front of the driver.

3.2 Method

3.2.1 Participants

To date the experiment was completed by 12 male participants (6 pairs) between the ages of 18 and 21 (the average age was 19.5). Subjects were recruited through email advertisement and received $20 in compensation. We plan to recruit two more subject pairs.

3.2.2 Driving task

Drivers drove in the middle lane of a three-lane highway in daylight. The highway had both straight and curvy segments. Each driver was instructed to follow a lead vehicle at a comfortable distance. The lead vehicle traveled at 89 km/h (55mph). There was also other traffic on the road travelling in adjacent lanes; however, the traffic did not interfere with the driver or the lead vehicle.

3.2.3 Spoken task: Taboo

The spoken task was the game of "Taboo," a game in which the other conversant is given a word, and needs to work with the driver to identify it, but cannot say that word or five related words. Participants played a series of Taboo games. We provided the words to the other conversant by displaying them on an LCD monitor, as shown in Figure 1. We imposed a time limit of 1 minute on each game.

The experimenter signaled the end of each Taboo game with an audible beep (0.5 second long, high pitched sine wave) heard by both conversants. The end of a game was reached in one of three ways:

- when the driver correctly guessed the word,
- when the other conversant used a taboo word, or
- when the conversants ran out of time.

The spoken task was played using two interaction conditions. In the *voice-only* condition the conversants could not see each other, and so could only use verbal communication. In contrast, in the *video call* condition conversants could also see each other on LCD displays.

3.2.4 Procedure

After filling out the consent forms and personal information questionnaires, participants were given an overview of the driving simulator, the Taboo game, and descriptions of the *voice-only* and *video call* conditions. Next, they completed two Taboo experiments, one for each interaction condition. Before each condition, we provided participants with about 5 minutes of training using that interaction condition. For training, participants completed Taboo games. In order to circumvent order effects, we plan to counterbalance the presentation order of the interaction conditions between eight participant pairs. However, in this paper we present preliminary results based on six participant pairs.

Each interaction condition was preceded by a short drive on a straight highway, which we used for training purposes for the given interaction condition. For both interaction conditions, participants drove a different route. Both routes started with a straight segment, which we used to establish baselines for our dependent variables. In the first route this initial straight segment (baseline) was followed by another straight segment and then a curvy segment. For the second route the initial segment was followed by the curvy and then by the final straight segment. The two routes were of the same length (about 15 km) and complexity. On average it took about 11 minutes to traverse a route. The presentation order of routes was the same for all subjects (but the interaction condition was varied).

After each interaction condition participants filled out a NASA-TLX questionnaire. Finally, at the end of the experiment, participants ranked their level of agreement with various statements pertaining to the interactions and provided written and verbal feedback about the experiment.

3.2.5 Design

In this paper we focus on the simplest scenario in our experiment: *voice-only* interactions on straight roads. From the perspective of pupil diameter measurements, *voice-only* interactions are simpler than *video-call* interactions, as we expect that in the *voice-only* condition the driver will keep his eyes on the road almost all the time. On the other hand, in the *video-call* condition, the driver might cast glances at the LCD display during verbal games, which may introduce noise in the pupil diameter signal as a result of changes in gaze angle and lighting conditions (simulator screen vs. LCD screen). At the same time, we expect that driving on straight segments will induce less (or at most as much) cognitive load as driving on curvy segments. Thus, as a result of driving alone, on straight segments the driver's pupil diameter should be smaller than (or at most as large as) on curvy segments. As pupil diameter is bounded, this smaller diameter is advantageous, because it allows for expansion due to increased cognitive load (due in turn to different dialogue behaviors) to be more pronounced.

We measured multiple dependent variables, of which in this paper we only report on the following:

- Dialogue characteristics: number of games completed, number of games completed successfully (driver correctly identified the word), game durations, and length of time from the end of a game (beginning of the beep) to the first contribution of the other conversant for the next game (excluding any fillers or other stalling devices).
- Cross-correlation to detect changes in pupil diameter after beeps that signaled the end of a Taboo game.

3.2.6 Measurement

Using an eye-tracker we obtained pupil diameter data. The sampling frequency of the eye-tracker was 60 Hz.

We recorded all dialogues in wav files and all beeps as log files created by custom software.

3.2.7 Calculation

Data segmentation: Using the time instants when the beeps started, we segmented each experiment into individual games. We performed calculations and analyze changes in cognitive load based on the pupil diameter data for each individual game.

Dialogue characteristics: From the audio recordings, we manually determined the start time of the first actual contribution by the other conversant. We also counted the number of games completed and the number completed successfully.

Cross-correlation: We estimated the cross-correlation between the beep vector (BV) and the pupil diameter vector (PDV). BV is a sequence of 0s and 1s, where a '1' represents the moment when the beep started (extracted from our custom log files), signaling the end of a Taboo game. Thus, there is a single '1' for each Taboo game. The PDV represents the processed measurements of the driver's left eye pupil diameter. We processed the raw measurements by interpolating short regions where the eye-tracker did not report pupil diameter measures, as well as by custom nonlinear smoothing (e.g. to reduce erroneous dips in pupil diameter caused by blinks).

The cross-correlation between BP and PDV was calculated as the average of cross-correlations for each of the segments and each of the 6 drivers. The lag variable indicates how much the change in pupil diameter lags behind the beginning of the beep signaling the end of the Taboo game. Thus, for positive values of lag, any drops in the cross-correlation function might represent drivers' reduced cognitive load after a beep.

4. RESULTS

The number of games completed by the conversants ranged from 11 to 16, with an average of 13.5. The percentage of successful games ranged from 63% to 100%, with an average of 85%. The game durations ranged from 16 to 24 seconds, with an average of 20.1 seconds. The first contribution by the other conversant happened between 3 and 5.6 seconds after the start of the beep, with an average of 4.6 seconds.

Figure 2 shows the average cross-correlation for all subjects between the BV and the PDV. As hypothesized, the cross-correlation drops in the seconds after the beep is initiated (which is at lag = 0 in the figure). The fact that the cross-correlation drops for about 5 seconds is consistent with the fact that the first contribution by the other conversant started on average about 4.6 seconds after the beginning of the beep (at lag = 0).

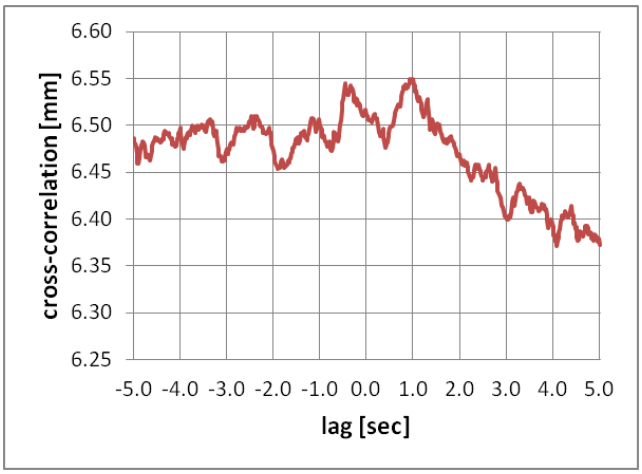

Figure 2 Cross-correlation for all six drivers.

The cross-correlations of two of the six drivers in this study did not clearly support our hypothesis. A number of factors could be

responsible, including differences in how the game was played by these participants (e.g. how engaged they were), and the noisiness of the pupil diameter measurements.

Figure 3 shows the cross-correlation (averaged over all segments) for the four drivers whose data did in fact support our hypothesis. In comparison to Figure 2, we can see that the drop is even more prominent. Additionally, the pupil diameter appears to be rising in the seconds before the end-of-game beep. We hypothesize that this rise is related to increased cognitive activity by the driver as he is attempting to find the word described by the other conversant. As correctly identifying this word is the most common cause of the end of the game, and thus the beep, it appears likely that cognitive load would indeed peak before the beep, thus at a negative value of lag. We should also expect to see a peak each time the driver makes a guess, but those peaks are not aligned with each other in time. Thus, they would not be visible after the cross-correlation operation.

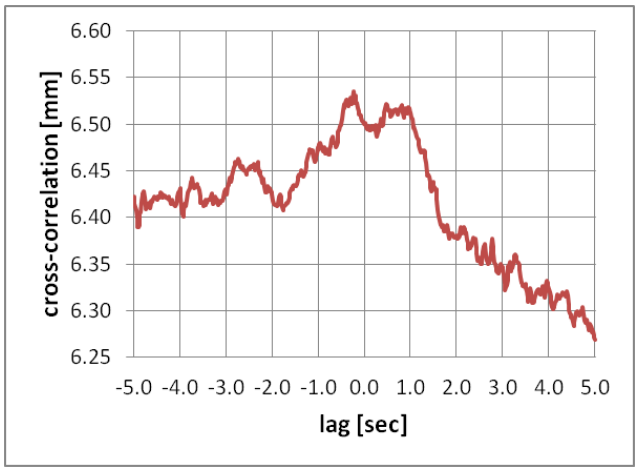

Figure 3 Cross-correlation for the four drivers whose results clearly supported our hypothesis.

5. CONCLUSIONS

The results for four of our six drivers support our hypothesis that pupil diameter can be used to identify major changes in cognitive load during a dialogue. Figure 3 indicates that for these drivers the pupil contracts by about 0.25 mm in the 4-5 seconds after the end of a Taboo game. Note that this effect size is similar to what we observed when we explored structured verbal tasks [6] as well as when we explored pupil diameter changes during an aural vigilance task [7]. These results are encouraging and indicate that using pupil diameter might be a viable approach to estimating the effects of dialogue behaviors on cognitive load changes.

Future efforts on this front should focus on collecting and processing large corpora of human-human dialogues and accompanying pupil diameter measurements. However, before such corpora are to be collected, researchers need to carefully identify potential confounding factors, such as effects resulting from the driving task, from the structure of the verbal task, and other effects on pupil diameter such as those due to changes in lighting (see e.g. [7]). Researchers would also benefit from improved signal processing algorithms for handling the effects of blinks and changes in gaze angle on pupil diameter measurements.

Once we establish the viability of our approach we expect that it will be useful in evaluating the effects of different dialogue behaviors on the driver's cognitive load. We expect that the approach will be useful both in human-human studies, which can result in behaviors that can inspire human-computer behaviors, as well as in human-computer studies, which will evaluate the behaviors implemented in different spoken dialogue systems.

6. ACKNOWLEDGMENTS

This work was supported by the US Department of Justice under grants 2006-DD-BX-K099, 2008-DN-BX-K221, 2009-D1-BX-K021 and 2010-DD-BX-K226.

7. REFERENCES

[1] Iqbal, S.T., Adamczyk, P.D., Zheng X.S., and Bailey, B.P. 2005. Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. *Proc. CHI'05.* Portland, OR: ACM, 311-320.

[2] Iqbal, S.T., Zheng, X.S., and Bailey, B.P. 2004. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. *Proc. CHI'04.* Vienna: ACM, 1477-1480.

[3] Klingner, J., Kumar, R., and Hanrahan, P. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. *Proc. ETRA '08*. Savannah, GA: ACM, 69-72.

[4] Marshall, S.P. 2002. The Index of Cognitive Activity: measuring cognitive workload. *Proc. IEEE Conf. on Human Factors and Power Plants*. 7-9.

[5] Mattes, S. 2003. The lane-change-task as a tool for driver distraction evaluation. *Proc. ISOES.*

[6] Palinko, O., Kun, A.L., Shyrokov, A., and Heeman, P. 2010. Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator. *Proc. ETRA '10*. Austin, TX: ACM, 141-144.

[7] Palinko, O., Kun, A.L. Exploring the Influence of Light and Cognitive Load on Pupil Diameter in Driving Simulator Studies. *Proc.* Driving Assessment 2011. Olympic Valley - Lake Tahoe, CA.

[8] Recarte, M.A., and Nunes, L.M. 2000. Effects of verbal and spatial-imagery tasks on eye fixations while driving. *J. Experimental Psychology*, 6(1):31-43.

[9] Schwalm, M., Keinath, A., and Zimmer, H.D. 2008. Pupillometry as a method for measuring mental workload within a simulated driving task. In *Human Factors for assistance and automation*, by Flemisch F., Lorenz B., Oberheid H., Brookhuis K. De Waard D. Maastricht: Shaker Publishing.