On the Feasibility of Using Pupil Diameter to Estimate Cognitive Load Changes for In-Vehicle Spoken Dialogues

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

[1] Electrical and Computer Engineering, University of New Hampshire, Durham, NH, USA
[2] Center for Spoken Language Understanding, OHSU, Beaverton, OR, USA

andrew.kun@unh.edu, oskar.palinko@unh.edu, zeljko.medenica@unh.edu, heemanp@ohsu.edu

Abstract

In a driving simulator study, we explore the feasibility of using pupil diameter to estimate how the cognitive load of the driver changes during a spoken dialogue with a remote conversant. We confirm that it is feasible to use pupil diameter to differentiate between parts of the dialogue that increase the cognitive load of the driver, and those that decrease it. Our long term goal is to build a spoken dialogue system that can adapt its behavior when the driver is under high cognitive load, whether from the driving task or the dialogue task.

Index Terms: dialog, cognitive load, pupil diameter, driving

1. Introduction

In-vehicle spoken dialogue systems (SDS) hold the promise of allowing drivers to accomplish secondary tasks without compromising their ability to safely operate the vehicle. However, the design of such interfaces must be done with care, as research shows that engaging in spoken interaction with humans [1,2] and machines [3] can have a detrimental effect on driving performance, and more generally on cognitive load. And while we do not have a formula linking different levels of driver cognitive load to the probability of crashes, it is generally accepted that the higher the cognitive load, the higher this probability is. Thus, in designing an in-vehicle SDS, we should endeavor to minimize (and ideally eliminate) the negative effects of the SDS on the driver's cognitive load.

Given that different behaviors exhibited by the SDS might have different effects on cognitive load, how can we evaluate the effect of the SDS on the driver's cognitive load? And how should we do this in the complex context of driving? We propose using changes in pupil diameter. Pupil diameter is a physiological measure of cognitive load: when people are faced with a challenging cognitive task, their pupils dilate. This phenomenon is called the Task Evoked Pupillary Response (TEPR) [4].

In this study, we explore the feasibility of using changes in pupil diameter to estimate the size and timing of cognitive load changes in spoken dialogue while the person is driving. Specifically, we explore the case of a driver and a remote conversant playing a verbal game, while keeping the driving difficulty constant. Our hypothesis is that the driver's pupil diameter will reflect the current difficulty level of the game. As this is a feasibility study, we will test our hypothesis by comparing the driver's pupil diameter only at two distinct locations in the game. First, we will observe a baseline pupil diameter when the remote conversant is preparing to speak. At this point in time, the spoken dialogue does not impose significant cognitive load on the driver. Second, we will observe pupil diameter when the driver is preparing to speak. This is the time when the driver is formulating a response to the remote conversant, which requires cognitive resources. Thus, if we find pupil diameter to be larger when the driver is preparing to speak than when the remote conversant is preparing to speak, our results will support our hypothesis.

2. Related research

A number of researchers have explored the effects of engaging in spoken dialogue on driving. Much of the attention was devoted to research on talking with a remote conversant on a mobile phone, and the results clearly indicate that such interactions can be detrimental to driving performance [1]. In our own work, we found evidence that certain characteristics of human-human spoken dialogues, such as switching from one task to another [2], can have a detrimental effect on driving performance, and more generally on cognitive load. We also found that certain characteristics of a speech user interface, such as low recognition rate [3], can negatively influence driving performance. These results indicate that designers must carefully evaluate the effects of the SDS on cognitive load and confirm that drivers can safely operate their vehicles even while using the SDS.

Pupil diameter has been used to assess cognitive load for a variety of tasks, such as mental arithmetic [5], auditory and visual vigilance [6], the effects of listening to, and identifying [7, 8], as well as generating [9] spoken information, and simultaneous interpretation [10]. In many of these studies, the tasks are highly structured and simple (e.g., participants are presented with one word at a time), and certainly cannot be viewed as extensive dialogues. In our prior 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 [11]. As part of this work participants played the highly structured last-letter game, in which they utter a word that starts with the last letter of the word uttered by the other participant. We found that the driver's pupil diameter was higher when it was the driver's turn to think of a word and utter that word, than when it was the remote conversant's turn to do the same. This result provides evidence that pupil diameter can be used to estimate cognitive load changes for in-vehicle speech interaction.

For SDS, we need to move away from single-word, highly structured tasks, to real dialogue. Drews et al. [12] used engaging and naturalistic conversations in their work on the impact of conversation on cognitive load. Charlton [13] had conversants, who did not know each other, discuss any topic they wished, or choose from some predefined ones (e.g., "a list of 10 songs to put on a mix tape to listen to on a long car trip"). Although both approaches resulted in naturalistic conversations, the dialogues were not task-based, and so are not representative of the dialogues that an SDS will be

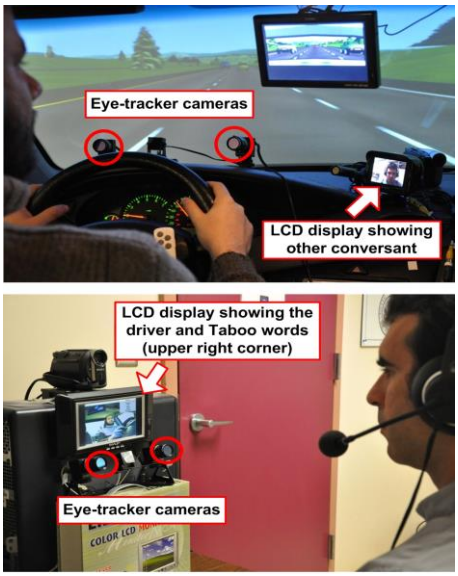

Figure 1: *a) Driver and b) remote conversant.*

engaged in. Also, note that neither study made use of pupil diameter to estimate cognitive load.

In a preliminary exploration of using eye-tracking with a less structured game [14], we found that pupil diameter can be used to identify major changes in cognitive load during dialogue. Specifically, we found that, on average, the driver's pupil contracts in the 4-5 seconds after the end of a word game. The current paper reports on new findings that utilize the same word game data set that we used in [14]. The major difference is that now we delve deeper into the problem by exploring changes in pupil diameter within a dialogue, and not only at the boundaries of dialogues.

3. Experiment

In our experiment, pairs of participants (the driver and the remote conversant) were engaged in a spoken dialogue. Additionally, the driver operated a simulated vehicle.

3.1. Equipment

The driver and remote conversant (see Figure 1) communicated using headphones and microphones. Their communication was supervised by the experimenter and synchronously recorded as a 48000 Hz audio file. Due to a technical problem, the audio was recorded as a mono signal rather than each conversant on their own channel. 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.

3.2. Tasks

3.2.1. Driving task

Drivers drove in the middle lane of a three-lane highway in daylight. They were instructed to follow a lead vehicle at a comfortable distance. The lead vehicle traveled at 89 km/h (55 mph). 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. Half of the highway was straight and the other half curvy.

3.2.2. Spoken task

The participants played the game of "Taboo," where the remote conversant is given a word, and needs to help the driver identify it, but cannot say that word or five related words. Participants played a series of Taboo games. We displayed the words to the remote conversant 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 game with an audible beep (0.5 second long, high pitched sine wave) heard by both conversants. The end of a game was reached when the driver correctly guessed the word, when the remote conversant used a taboo word, or when the conversants ran out of time.

The game was played using two interaction conditions. In the *speech-only* (SO) condition the conversants could not see each other, and thus could only use speech communication. In contrast, in the *video call* (VC) condition conversants could also see each other on LCD displays. Figure 1 a) and b) demonstrates the VC condition from the driver's and the other conversant's perspective. The SO condition was played in the same way, but the two LCD displays did not show the conversants to each other.

3.3. Participants

The experiment was completed by 16 male participants (8 pairs) between the ages of 18 and 21 (the average age was 19.4). Participants were recruited through email advertisement and received $20 in compensation.

3.4. Experimental conditions

We employed three independent variables: *Interface*, *Road Type* and *Dialogue Position*.

Interface had two levels: speech-only (SO) and video call (VC), as discussed above. In this paper we report *only* on the SO condition. We do this as we expect that in the VC condition pupil diameter will be affected by glances to the LCD due to changes in the amount of light reaching the driver's retina, which will require more complex estimation of pupil diameter.

Road Type determined the type of road where the spoken task was performed, namely, *straight* (Figure 1 a) or *curvy*.

Dialogue Position indicated where in the game the conversants were, namely just before the remote conversant's *first contribution* (FC) or just before the driver's *response* (R).

Since this was a within-subjects experiment, all drivers experienced all experimental conditions.

3.5. Procedure

After completing the consent forms and personal information questionnaires, participants were given an overview of the driving simulator, the Taboo game, and descriptions of the SO and VC conditions. Next, they completed two sessions, one for each interaction condition. We counterbalanced the presentation order of the interaction conditions between the 8 participant pairs. Before each session, we provided the participants with about 5 minutes of training using the

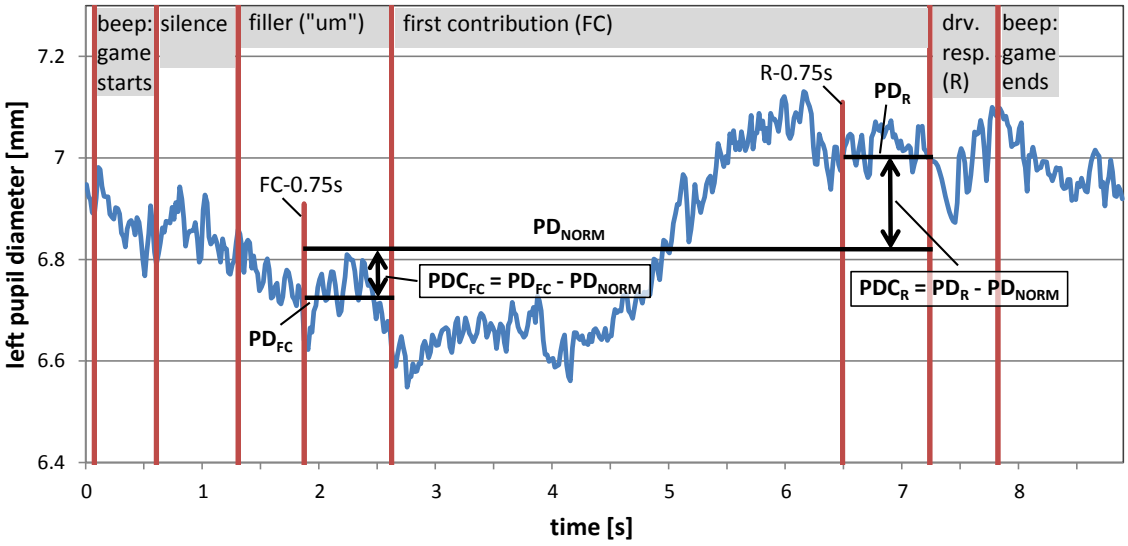

Figure 2: *Example pupil diameter change over the course of a game.*

interaction condition for that session. For training, participants played Taboo games, with the driver operating the simulated vehicle.

Sessions started with a short drive on a straight road during which the driver could adjust to the driving task. Next, participants completed Taboo games while the driver was presented with two longer road segments: one straight and one curvy. For the first interaction condition drivers drove on the straight segment first, followed by the curvy segment. For the second interaction condition drivers encountered the curvy segment first and the straight second. This order of presentation was the same for all drivers. In each session drivers covered about 15 km of road in about 11 minutes, and played 11 to 16 Taboo games.

3.6. Measurement and dialogue transcription

We measured multiple dependent variables; in this paper we only report on pupil diameter, which we obtained using the eye-tracker. We measured the left pupil diameter at a sampling frequency of 60 Hz. 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 to reduce erroneous dips in pupil diameter caused by blinks.

We also recorded all dialogues and beeps in audio files and separately all beeps as log files created by custom software. Two people transcribed the words that were said in the audio files. They compared their transcriptions and came to a consensus on any differences. As the audio files are single channel, when the speakers overlapped each other or overlapped with the beep, it was not always possible to determine the exact words that were said, or their timing.

3.7. Calculation and statistical analysis

Using the start times of the beeps, we segmented each session into individual games. We rejected games in which the remote conversant used a taboo word during the first contribution, which ended the game without driver response. We also rejected one game in which the remote conversant did not know the meaning of the taboo word.

We analyzed changes in cognitive load based on pupil diameter data for each individual game, as illustrated in Figure 2, which provides an example for how pupil diameter changes over a game in our corpus. Specifically, we determined the start time of the first contribution by the remote conversant (FC) and the start time of the driver's response (R). We then found the driver's average pupil diameter during 0.75 seconds before the remote conversant's first contribution (PD_{FC}) and before the driver's response (PD_R). Next, we calculated the pupil diameter change for the times before the first contribution (PDC_{FC}) and before the response (PDC_R). We did this by subtracting from PD_{FC} and PD_R the average pupil diameter from the start of PD_{FC} to the end of PD_R (the value labeled PD_{NORM} in Figure 2). This allows us to compare data for participants with different pupil sizes, and focus on the change in the pupil size. Finally, we found the overall mean pupil diameter change for each participant (averaged over all games), both before the first contribution ($MPDC_{FC}$) and before the response ($MPDC_R$).

In the example game in Figure 2 the driver's pupil diameter is around 6.6-6.8 mm before the remote conversant's first contribution (which starts around 2.6 seconds). The pupil diameter rises during the first contribution, and is around 7 mm before the driver's first response (which starts around 7.25 seconds). The game ends at around 7.8 seconds, as the driver guesses the taboo word.

Our selection of window length (0.75 seconds) for calculating the PDC is based on data from a number of studies that indicate that pupil diameter can change rapidly with changes in cognitive load (see e.g. [2], [6], and [8]). In fact, such rapid change is illustrated in Figure 2. However, we did not systematically vary the window length in order to explore the effect of window length on our results.

4. Results

Figure 3 shows the mean pupil diameter for the experimental conditions explored in this paper (straight vs. curvy road, before remote conversant's first contribution vs. before driver response). A repeated-measures multivariate analysis of variance (MANOVA) showed a significant main effect for *Dialogue Position* ($F_{1,7}$=13.37, p<.01), but not *Road Type*

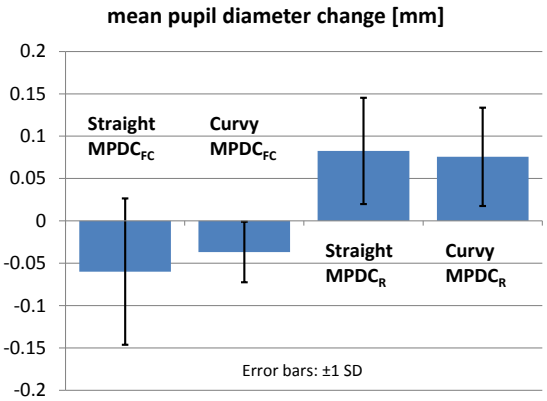

Figure 3: *MPDC for the time periods before the first contribution ($MPDC_{FC}$) and before the driver's response ($MPDC_R$) for straight and curvy roads.*

($F_{1,7}$=0.42, ns). No interaction was observed between *Dialogue Position* and *Road Type* ($F_{1,7}$=0.62, ns). Exploring interface usage within the individual road types using a univariate ANOVA, a significant difference was observed between $MPDC_R$ and $MPDC_{FC}$ for both straight ($F_{1,7}$=8.72, p<.05) and curvy ($F_{1,7}$=15.16, p<.01) roads. Collapsing data between the two road types, the difference between $MPDC_R$ and $MPDC_{FC}$ is about 0.13 mm, which is an effect size similar to what we observed in [11] and [14].

These results support our hypothesis that it is feasible to use pupil diameter to identify the current difficulty level of the game from the driver's perspective: when the driver does not have to pay attention to the game (before the remote conversant's first contribution) pupil diameter is smaller than at a time that the driver is formulating his response. However, while the results are encouraging, they represent only an initial step, as they do not evaluate the feasibility of comparing the effects of *different dialogue behaviors* on the driver's cognitive load.

To obtain the results above, we relied on averaging data from about 200 games and 8 participants. In addition, we wanted to assess how robust our approach is on a case-by-case basis. Thus, we analyzed the difference between PDC_{FC} and PDC_R for each of the games played. The result of this analysis is shown in Table 1. The second column shows that taking into account both curvy and straight segments, participants played anywhere from 23 to 28 games (not including 16 rejected games, constituting 7.2% of the total games played). The third column of Table 1 indicates that PDC_{FC} was less than PDC_R in 46% to 89% of the games. In other words, the comparison of PDC_{FC} and PDC_R supports our hypothesis (pupil diameter allows us to observe that the driver's cognitive load is lower before the remote conversant's first contribution than before the driver's response) in 46% to 89% of the games played, and overall, in 69% of the games we examined. Just as the MPDC calculations above, this indicates that the simple approach to evaluating cognitive load associated with the remote conversant's first contribution and the driver's response (use mean pupil diameter change calculated during 0.75 seconds before an utterance) can be useful in many, but certainly not all, cases.

Table 1. *Game statistics. The last column shows the number of games supporting our hypothesis.*

Participant pair	Games (curvy + straight)	Games with $PDC_{FC} < PDC_R$
1	27	14 (52%)
2	28	13 (46%)
3	23	19 (82%)
4	24	21 (88%)
5	25	13 (52%)
6	27	24 (89%)
7	25	16 (64%)
8	25	21 (84%)
Total	203	141 (69%)

5. Conclusions

Overall, the above results are encouraging, as they provide a proof-of-concept that it is indeed possible to use pupil diameter to differentiate between parts of the dialogue that increase the cognitive load of the driver, and those that decrease it. Specifically, averaging over roughly 200 games played by eight participant pairs, as well as in 69% of the individual Taboo games we explored, we were able to identify that the driver's pupil diameter (and thus presumably cognitive load) increases when the driver is thinking of a response to the first contribution of the remote conversant. In fact, for four of our participant pairs (pairs 3, 4, 6 and 8), our approach appears to be very successful in finding differences in driver cognitive load before the remote conversant's first contribution and the driver's response. However, the less-favorable results for the other four participant pairs suggests that the time interval before the driver starts to speak might not always have higher cognitive load. We expect that there are many factors that influence cognitive load. Some speakers might start speaking before they have formulated a complete utterance, and use disfluencies to amend what they are saying if need be [15]. Or some of the remote participants might give better hints, thus reducing the amount of cognitive load that the driver experiences. Alternatively, while we attempted to keep the driving task uniform during sessions, in some cases fluctuations in cognitive load (and thus pupil diameter) due to the driving might have masked the effect of the dialogue task. Similarly, it is possible that in some cases the driver's pupil diameter was affected by the pupillary light reflex in such a way as to mask the effect of cognitive load change due to the dialogue task [16]. In future work, we will take into account these and other additional factors (such as utterance delivery and higher level dialogue processing) to model the driver's expected cognitive load.

Our long term objective is to determine how, for task-oriented dialogues, an SDS can interact with drivers without negatively impacting their cognitive load. As Campana et al. point out, this requires tools to compare the effect of different SDS features on cognitive load [17]. Our results indicate that pupil diameter can be one such tool.

6. Acknowledgements

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] W. J. Horrey and C. D. Wickens. Examining the impact of cell phone conversations on driving using meta-analytic techniques. Human Factors: The Journal of the Human Factors and Ergonomics Society 48(1), pp. 196-205. 2006.

[2] A. L. Kun, A. Shyrokov and P. A. Heeman. Interactions between human–human multi-threaded dialogues and driving. Accepted for publication in Personal and Ubiquitous Computing.

[3] A. Kun, T. Paek and Z. Medenica. The effect of speech interface accuracy on driving performance. Interspeech 2007.

[4] J. Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91(2), pp. 276. 1982.

[5] S. P. Marshall. The index of cognitive activity: Measuring cognitive workload. 2002 IEEE Conference on Human Factors and Power Plants.

[6] J. Klingner, B. Tversky and P. Hanrahan. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. Psychophysiology 48(3), pp. 323-332. 2010.

[7] M. Zellin, A. Pannekamp, U. Toepel and E. van der Meer. In the eye of the listener: Pupil dilation elucidates discourse processing. International Journal of Psychophysiology 81(3), pp. 133-141. 2011.

[8] S. E. Kramer, A. Lorens, F. Coninx, A. A. Zekveld, A. Piotrowska and H. Skarzynski. Processing load during listening: The influence of task characteristics on the pupil response. Language and Cognitive Processes, 2012.

[9] S. T. Iqbal, Y. Ju and E. Horvitz. Cars, calls, and cognition: Investigating driving and divided attention. CHI 2010.

[10] J. Hyönä, J. Tommola and A. Alaja. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. Q. J. Exp. Psychol. 48(3), pp. 598-612. 1995.

[11] O. Palinko, A. L. Kun, A. Shyrokov and P. Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. 2010 Symposium on Eye-Tracking Research & Applications.

[12] F. A. Drews, M. Pasupathi and D. L. Strayer. Passenger and cell phone conversations in simulated driving. Journal of Experimental Psychology: Applied 14(4), pp. 392. 2008.

[13] S. G. Charlton. Driving while conversing: Cell phones that distract and passengers who react. Accident Analysis and Prevention 41(1), pp. 160. 2009.

[14] A. L. Kun, Z. Medenica, O. Palinko and P. A. Heeman. Utilizing pupil diameter to estimate cognitive load changes during human dialogue: A preliminary study. AutomotiveUI Adjunct Proceedings 2011.

[15] W. J. Levelt. Speaking: From Intention to Articulation 1993.

[16] O. Palinko and A. L. Kun. Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. Driving Assessment 2011.

[17] E. Campana, M. Tanenhaus, J. Allen and R. Remington. Evaluating cognitive load in spoken language interfaces using a dual-task paradigm. 5th International Conference on Spoken Language Processing (ICSLP'04), 2004.