Estimating Cognitive Load Using Pupil Diameter During a Spoken Dialogue Task

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

[1] Oregon Health & Science University
Center for Spoken Language Understanding
Portland OR, USA

[2] University of New Hampshire
Electrical and Computer Engineering
Durham NH, USA

heemanp@ohsu.edu, tmeshorer@hotmail.com, {andrew.kun,oskar.palinko,zeljko.medenica}@unh.edu

ABSTRACT

We explore the feasibility of using pupil diameter to estimate how the cognitive load of the driver changes during a spoken dialogue task with a remote conversant. The conversants play a series of Taboo games, which do not follow a structured turn-taking nor initiative protocol. We contrast the driver's pupil diameter when the remote conversant begins speaking with the diameter right before the driver responds. Although we find a significant difference in pupil diameter for the first pair in each game, subsequent pairs show little difference. We speculate that this is due to the less structured nature of the task, where there are no set time boundaries on when the conversants work on the task. This suggests that spoken dialogue systems for in-car use might better manage the driver's cognitive load by using a more structured interaction, such as system-initiative dialogues.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Natural Language

General Terms

Human Factors, Design, Experimentation, Measurement

Keywords

Speech user interfaces, cognitive load, driving simulator

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 interface should minimize the impact on the cognitive load of the driver, as high cognitive load might increase the probability of an accident.

Different behaviors exhibited by a SDS might have different effects on the driver's cognitive load. Furthermore, these behaviors might just span a portion of the dialogue. Hence, it is imperative that we are able to measure the short-term cognitive load of the driver. 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 [1].

In this paper, we will use a corpus of human-human dialogues between a driver operating a driving simulator and a remote conversant. The goal of this paper is to show that areas of a dialogue that should involve more cognitive resources should correlate to pupil dilations. We have already shown this in the last-letter game, where we contrasted regions where the driver had to wait for the other participant, with regions where the driver had to think of the next word [12]. However, the last-letter game is highly structured, where conversants alternate saying a single word. Thus, many of the complexities of dialogue are not present, such as co-ordinating turn-taking and showing initiative. Hence, we will extend our previous research to a less structured task. First, we specify regions in the dialogue where we expect the driver be under low cognitive load and regions where we expect the driver to be under high cognitive load. Second, we will determine whether the pupil diameter varies accordingly with these regions. The regions that we choose are as follows:

A: first 0.5 seconds of the remote conversant's speaking turn.

B: last 0.5 seconds before the driver's turn.

We expect the *A* regions to have low cognitive load for the driver, and the *B* regions to have high cognitive load. Thus, we should see larger pupil dilations for the B regions than the A regions. In the rest of the paper, we will discuss the related work, then describe the dialogue collection, the data analysis, and then discuss the results.

2. RELATED WORK

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 [4]. In our own work, we found evidence that certain characteristics of human-human spoken dialogues, such as switching from one task to another [11], 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 [8], can negatively influence driving performance. These results indicate that

AutomotiveUI'13, October 28-30, 2013, Eindhoven, The Netherlands

ACM 978-1-4503-2478-6/13/10...$15.00.
http://dx.doi.org/10.1145/2516540.2516570.

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 memorization [1], auditory and visual vigilance [7], the effects of generating spoken information [6], and simultaneous interpretation [5]. 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. For example, in our prior work [12], we used a remote eye tracker in a driving simulator experiment to estimate the cognitive load of the driver while the driver and a remote participant played the highly structured last-letter game, where participants take turns saying a word that begins with the last letter of the previous word. 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 a SDS, we must move from single-word, highly structured tasks, to real dialogue. Drews et al. [3] used engaging and naturalistic conversations in their work on the impact of conversation on cognitive load. Charlton [2] had conversants, who did not know each other, discuss any topic they wished, or choose from some predefined ones. 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 engaged in. Also, neither study made use of pupil diameter to estimate cognitive load.

We have conducted a study in which we used an eye-tracker to gauge cognitive load while conversants 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. Unlike our study with the last letter game [12], we cannot control when the driver should speak, or what the driver is thinking about. As a first step, we found that the driver's pupil contracted in 4-5 seconds after the end of each game [9]. We also contrasted the beginning of each dialogue, right before the remote conversant makes the first contribution, and right before the driver starts speaking [10]. We found an increase in the pupil diameter in 69% of the games. The current paper uses the same Taboo data, but we contrast the first dispatcher-driver pair with subsequent ones during the dialogues.

3. EXPERIMENT

In our experiment, pairs of participants (the driver and the remote conversant) are engaged in a spoken dialogue. Additionally, the driver operates 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 48 kHz 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 180° 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.

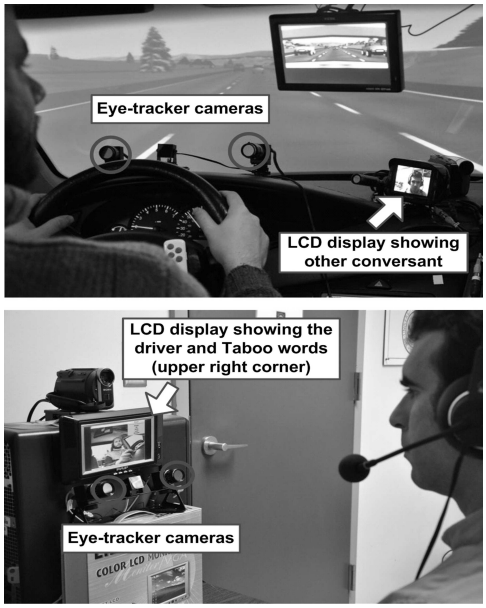

Figure 1: a) Driver and b) remote conversant

3.2 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 traveling 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.

3.3 Spoken Task

Participants played a series of "Taboo" games. We displayed the words to the remote conversant on an LCD monitor, as shown in Figure 1. The experimenter signaled the end of each game with an audible beep (0.5 second long, high pitched tone) heard by both conversants. The game ended when the driver correctly guessed the word, when the remote conversant used a taboo word, or when the conversants ran over the time limit of 1 minute.

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 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.

3.4 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.

3.5 Experimental conditions

In this within-subjects experiment we employed 3 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.

Road Type is whether the spoken task was performed on a straight or curvy road. In our previous work [10], we found that pupil diameter did not depend on Road type, so we collapse data between the two road types.

Dialogue Position indicated where in the game the conversants were, namely the first 0.5s of the remote conversant's speaking turn (A) or the first 0.5s of the driver's turn (B).

3.6 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 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. In each session drivers covered about 15 km of road in about 11 minutes, and played 11 to 16 Taboo games.

3.7 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 recorded all dialogues and beeps in audio files, and the timing of the beeps in log files. Two people transcribed the words that were said. 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.

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 a driver response. We also rejected one game in which the remote conversant did not know the meaning of the taboo word.

4. DATA ANALYSIS

We analyzed changes in cognitive load based on pupil diameter data for each individual game. We determined the start time of each turn by the driver and passenger. We omit any fillers at the beginning of turns. We paired each turn of the remote participant with the subsequent turn of the driver. We computed the driver's average pupil diameter for the first 0.5s of the remote conversant's turn A_p, and the last 0.5s before the driver spoke B_p, where p indicates the position of the pair in the dialogue (with $p = 1$ as the first pair). Similar with our earlier work [10], we computed the number of times in which $A_1 < B_1$ over all games for each speaker. For example, for subject pair 7, of the 25 games they played, the pupil diameter increased in 17 of the games. The results are reported in the first column of Table 1. Averaging over the results of each speaker, 71.0% of the time drivers had larger pupil diameter when they were about to speak for the first time in a game, than when the remote conversant started to speak. This is significant with the non-parametric Wilcoxon signed-rank test, N=8, W=34.0, p<.05, one-tailed.

Table 1: Pupil diameter changes within a game

Pair	First Pair $A_1<B_1$		Later Pairs $A_i<B_i$ $i>1$		Between $A_2<B_1$	
1	15/27	(55.6%)	16/29	(55.2%)	10/15	(66.7%)
2	15/29	(51.7%)	33/57	(57.9%)	9/16	(43.8%)
3	22/24	(91.7%)	28/45	(62.2%)	8/17	(47.1%)
4	22/26	(84.6%)	18/27	(66.7%)	3/15	(20.0%)
5	12/27	(44.4%)	14/37	(37.8%)	5/18	(27.8%)
6	26/28	(92.9%)	21/37	(56.8%)	7/15	(46.7%)
7	17/25	(68.0%)	9/29	(31.0%)	5/13	(38.5%)
8	19/24	(79.2%)	20/38	(52.6%)	6/16	(37.5%)
Total		(71.0%)		(52.5%)		(41.0%)

We next looked at the subsequent interaction cycles in the dialogues. This is reported in the second column in Table 1. Contrary to our hypothesis, these interaction cycles did not consistently show an increase in pupil diameter from when the remote conversant started speaking to the time when the driver was just about to speak (Wilcoxon signed-rank test, N=8, W<26, NS). Furthermore, we expected that pupil diameter would decrease from when the driver spoke in the first interaction cycle and when the remote conversant spoke in the second interaction cycle (if there was a second cycle). In other words, we expected to see that on average $A_2 < B_1$. However, as shown in the third column of Table 1, we found that this only held in 41% of the time in our corpus.

4.1 Between Games

We next compare cognitive load across games. For our variables A and B, introduced above, we use a superscript to indicate which game we are referring to. We compare the pupil diameter of the driver right before he spoke the last utterance of the game (and so guessed the right answer) with the first turn by the remote conversant of the next game: B^g_{end} with A^{g+1}_1. We expect pupil diameter to decrease when the participants start a new game. The results are in the first column of Table 2, which confirm our hypothesis, Wilcoxon signed-rank test, N=7, W=28.0, p<.05, one-tailed. These results also support the results in [9], where we found that the driver's pupil contracts after a game.

However, while the driver's pupil tends to contract after a game, it does not contract back to the size it was at the beginning of the experiment. Specifically, the second column of Table 2 shows that in 82.3% of the cases pupil size is smaller at the beginning of the first game (A^1_1, which is the beginning of the experiment) than it is at the beginning

Table 2: Pupil diameter changes between games

Pair	Between Games $B^g_{end}>A^{g+1}_1$	First Start $A^i_1>A^1_1 \quad i>1$	Later Starts $A^{i+1}_1>A^i_1 \quad i>1$
1	13/26 (50.0)%	23/26 (88.5%)	14/25 (56.0%)
2	17/28 (60.7)%	26/28 (92.9%)	15/27 (55.6%)
3	21/23 (91.3)%	19/23 (82.6%)	12/22 (54.5%)
4	24/25 (96.0)%	11/25 (44.0%)	15/24 (62.5%)
5	17/26 (65.4)%	26/26 (100.0%)	15/25 (56.0%)
6	23/27 (85.2)%	23/27 (85.2%)	14/26 (57.7%)
7	15/24 (62.5)%	22/24 (91.7%)	9/23 (39.1%)
8	22/23 (95.7)%	17/23 (73.9%)	10/22 (45.5%)
Total	(75.8)%	(82.3%)	(53.4%)

of subsequent games (A^i_1, $i>1$). We confirmed the significance of this result using the Wilcoxon signed-rank test, N=8, W=34.0, p<.05, two-tailed. Furthermore, as shown in the third column of Table 2, if we focus on games after the first one ($i>1$), we find no statistically significant difference between pupil diameter at the beginning of these games (Wilcoxon signed-rank test, N=8, W<26, NS).

5. DISCUSSION

Based on our work with the last-letter game, we hypothesized that we could divide each game of Taboo into a series of events that capture differences in the cognitive load of the driver, which we would see through changes in pupil diameter. Specifically, we expected drivers to show high cognitive load right before speaking, and low cognitive load when the remote conversant started speaking. Our reasoning was that after the driver spoke, he/she should just be waiting for the remove conversant to provide more information. Although we found this to be the case at the very beginning of each game, we did not find it throughout. We speculate that drivers' cognitive load is not decreasing after they finish their turn. After the driver speaks, the experimenter signals a successful guess with a beep. The absence of the beep informs the participants that the game should continue. In this case the driver might not idly wait for new information from the remote conversant; but instead continue to think of another guess, and perhaps even make another guess if the remote conversant is slow in giving another clue. Making another guess without waiting for the remote conversant is possible since, unlike the last letter game, Taboo does not have a rigid turn-taking structure. If no one is currently speaking, either participant can grab the speaking floor.

We also see that after the first game, the pupil diameter does not return to its initial (small) size. This might be because conversants move from one game immediately to the next, not allowing cognitive load to fully reset.

6. CONCLUSIONS

Our long term goal is to determine how to build a SDS that does not increase the cognitive load of the user. In this present work, we were not able to find events beyond the start and end of a dialogue where there are consistent changes in pupil diameter. However, it is interesting to contrast our experience with Taboo and the last letter game. The last letter game has a rigid turn-taking policy, in which conversants alternate saying a single word. After the driver says his/her word, there is nothing to think about until the remote conversant says his/her word. However, in Taboo, anyone can say anything at any time. The driver can continue to reason about what the taboo word is, even right after having said one. This difference might be why we were able to see changes in cognitive load with last letter game that we weren't able to see in Taboo. The implication of this on in-car SDSs is that it might be better to use a more structured interaction, where the system can better control how much the user will think about the secondary task. In fact, it might be much easier to control the driver's cognitive load with system-initiative dialogues rather than mixed initiative dialogues.

7. ACKNOWLEDGMENTS

Work at UNH was supported by the US DOJ under grants 2009-D1-BX-K021, and 2010-DD-BX-K226.

8. REFERENCES

[1] J. Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 1982.

[2] S. Charlton. Driving while conversing: Cell phones that distract and passengers who react. *Accident Analysis and Prevention*, 4:160–173, 2009.

[3] F. Drews, M. Pasupathi, and D. Strayer. Passenger and cell phone conversations in simulated driving. *J Exp Psychol Appl*, 14(4):392–400, 2008.

[4] W. Horrey, C. Wickens, and K. Consalus. Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *J Exp Psychol Appl*, 12(2):67–78, 2006.

[5] 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):598–612, 1995.

[6] S. Iqbal, Y. Ju, and E. Horvitz. Cars, calls, and cognition: Investigating driving and divided attention. In *Human Factors in Computing Systems*, 2010.

[7] 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):323–332, 2010.

[8] A. Kun, T. Paek, and Z. Medenica. The effect of speech interface accuracy on driving performance. In *Interspeech*, Antwerp Belgium, Aug. 2007.

[9] A. Kun, Z. Medenica, O. Palinko, and P. Heeman. Utilizing pupil diameter to estimate cognitive load changes during human dialogue: A preliminary study. In *Workshop on Cognitive Load and In-Vehicle Human-Machine Interaction*, Salzburg Austria, 2011.

[10] A. L. Kun, O. Palinko, Z. Medenica, and P. Heeman. On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues. In *Interspeech*, Lyon France, Aug. 2013.

[11] A. Kun, A. Shyrokov, and P. Heeman. Interactions between human-human multi-threaded dialogues and driving. *Personal and Ubiquitous Computing*, 2013.

[12] O. Palinko, A. Kun, A. Shyrokov, and P. Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. In *Eye Tracking Research and Applications*, pp. 141–144, Austin TX, Mar. 2010.