Cognitive Workload Using Interactive Voice Messaging Systems

James R. Coleman, Jonna Turrill, Joel M. Cooper, & David L. Strayer
University of Utah

The current research sought to understand the sources of cognitive distraction stemming from voice-based in-vehicle infotainment systems (IVIS) to send and receive textual information. Three experiments each evaluated 1) a baseline single-task condition, 2) listening to e-mail/text messages read by a "natural" pre-recorded human voice, 3) listening to e-mail/text messages read by a "synthetic" computerized text-to-speech system, 4) listening and composing replies to e-mail/text messages read by a "natural" voice, and 5) listening and composing replies to e-mail/text messages read by a "synthetic" voice. Each task allowed the driver to keep their eyes on the road and their hands on the steering wheel, thus any impairment to driving was caused by the diversion of non-visual attention away from the task of operating the motor vehicle.

INTRODUCTION

The present research sought to determine the mental workload associated with using voice-based in-vehicle infotainment systems (IVIS) to send and receive textual information (e.g., text messages and e-mails). Our previous research (Strayer et al., 2015) developed a method for assessing cognitive workload in the automobile. We found that using a voice-based e-mail/text messaging system that converted text-to-speech and translated speech into text was associated with a surprisingly high level of cognitive workload (3.1 on our 5-point rating system). By comparison, talking on either a hand-held or hands-free cell phone had a cognitive workload rating of 2.3, significantly lower than that observed with the IVIS interactions. These findings highlight the need to understand why IVIS interactions are so cognitively demanding.

One factor that differed between the cell-phone conditions and the voice-based IVIS system tested by Strayer et al., (2015) is that the former involved both listening to and talking to another person whereas the latter involved interactions using computerized synthetic speech. Prior research (Harbluk, 2005; Jamson, Westerman, Hockey, & Carsten, 2004; Lee, Caven, Haake, & Brown., 2001; Raney, Harbluk, & Noy. 2005) has found that synthetic speech leads to higher levels of cognitive workload. However, in the last decade there have been considerable improvements to the computerized speech algorithms. We sought to determine if speech quality was responsible for higher levels of workload associated with voice-based IVIS interactions.

Another factor that has been shown to affect mental workload is the degree to which participants are listening to speech (i.e., speech comprehension) as compared to the degree to which participants are generating speech (i.e., speech production). Prior research has found that the latter is more mentally demanding that the former (Bergen et al., 2014). Are the workload differences observed by Strayer et al. (2015) due to differences in speech comprehension vs. speech production?

To examine these issues, we conducted three experiments, each of which employed a 2X2 factorial design, where natural human speech versus synthetic speech was crossed with conditions where the driver listened to messages without generating a reply versus conditions where the driver listened to messages and composed a reply when it was required (i.e. not a "spam" message). Experiment 1 assessed the mental workload associated with performing the "secondary tasks" in isolation. Experiment 2 assessed the mental workload when these tasks were paired with driving in a driving simulator. The simulator provided a standardized driving environment upon which to evaluate dual-task performance. Experiment 3 assessed the mental workload when participants were driving an instrumented vehicle on residential streets. Utilizing this design allowed us to localize the sources of mental workload when the driver interacted with a perfectly reliable speech-to-text and text-to-speech system.

EXPERIMENT 1

Method

Participants. Forty-five participants (27 men and 18 women) completed the experiment. They ranged in age from 18 to 40 years ($\bar{x}$ = 24.8 years). All reported having a valid driver's license and were fluent in English. Participants' years of driving experience ranged from 2.5 to 24 ($\bar{x}$ = 8.5 years). All of the participants owned a cellular phone and 87% reported that they used their phone regularly while driving. They were recruited via flyers posted on campus bulletin boards and through word of mouth within the community. Eligible participants had a clean driving history (e.g., no at-fault accidents in the past five years). They received $15/hour in compensation for their participation in the experiment.

Equipment. Microsoft PowerPoint 2013 was used to coordinate an interactive messaging service with text-to-speech features. Participants were given a short list of commands (i.e., *Repeat, Reply, Delete, Next Message,* and *Send*) that were used to control the messaging program. The experimenter, who reacted to the participants' verbal commands, mimicked a speech detection system with perfect reliability, implementing the "Wizard-of-Oz" technique (Kelley, 1983; Lee et al., 2001; Strayer et al., 2015). TEAC CD-X70i Micro Hi-Fi system speakers were used for the presentation of the audio for each of the conditions.

A peripheral detection response task (DRT) was used to quantify the workload associated with task performance (ISO, 2012). We adopted the protocol used by Strayer et al. (2015), in which a red/green LED light was attached on the

participant's head via a headband. The light was adjusted to an average 15° to the left and 7.5° above the participant's left eye. Response reaction time (RT) was recorded with millisecond accuracy via a microswitch attached to participants' left thumb that was depressed in response to the green light.

Procedure. Participants were asked to complete five different 9-minute conditions, each of which are described below. The order of conditions was counterbalanced across participants using a balanced Latin Square design. The participants sat in front of a computer screen displaying a fixation cross and were asked to look forward and avoid making excessive head and eye movements during the performance of each task. Before each condition began, participants were familiarized with the procedure for interacting with the voice-command system and they were required to demonstrate proficiency before data collection commenced.

The first of the conditions was a single task condition that was selected to provide a baseline level of performance in the DRT task (i.e., no concurrent secondary task). There were also four conditions, each described in detail below, in which participants interacted with the IVIS system. Each condition contained messages matched in type and duration.

In the second condition (Natural Listen), participants interacted with a simulated email/text messaging service. The system was fully automated with perfect speech recognition capability. Before beginning the condition, the participant was familiarized with the basic commands, which were: *Repeat*, *Delete*, and *Next Message*. The email and text messages and the system instructions and feedback were pre-recorded using a high-fidelity female voice. Participants were asked to listen to the messages, but they were not allowed to compose or send messages in reply. The messages were designed to be representative of text/email messages that individuals receive on a regular basis from friends, family, coworkers, and service providers (i.e., spam).

In the third condition (Synthetic Listen), participants interacted with the same system design as in the second. However, the messages and system interactions were pre-recorded using a synthetic, computerized female voice, "Kate," from NeoSpeech (NeoSpeech, 2012). NeoSpeech was selected because of its superior synthetic speech generation capabilities. Prior to beginning the condition, the participant was familiarized with the program's basic commands, which were the same as the previous condition.

In the fourth condition (Natural Listen + Compose), participants interacted with the same system design that was used in the second condition, with the exception that they were allowed to compose replies to the textual messages. Before beginning the condition, the participant was familiarized with the program's basic commands, which were: *Repeat*, *Reply*, *Delete*, *Next Message,* and *Send*. The messages and system confirmations were pre-recorded using the same high-fidelity female voice used in the second condition. Thus the only difference between the second and fourth conditions is that participants were asked to *compose and send* a response to messages that required one.

The fifth condition (Synthetic Listen + Compose), was identical to the fourth except that the messages and system confirmations were pre-recorded using the same synthetic NeoSpeech female voice used in the third condition. Before beginning, the participant was familiarized with the program's basic commands, which were identical to the commands used for the fourth condition.

Results

The DRT data reflect the participant response times to the green lights in the peripheral detection task. The RT and accuracy data for the DRT task are plotted in Figures 1 and 2, respectively. RT for correct responses (i.e., green light responses) was measured to the nearest msec. The accuracy data were converted to the non-parametric measure of sensitivity, A', where a response to a green light was coded as a "hit," non-responses to a red light were coded as a "correct rejection," non-responses to a green light were coded as a "miss," and responses to a red light were coded as a "false alarm" (Pollack & Norman, 1964). A repeated measures Analysis of Variance (ANOVA) found that RT increased across condition, $F(4, 176) = 27.26$, $p < .01$, partial $\eta^2 = .38$. An analysis restricted to the 2 (speech quality: natural vs. synthetic) X 2 (task type: listen vs. listen + compose) factorial found a significant effect of task type on RT, $F(1, 44) = 8.58$, $p < .01$, partial $\eta^2 = .16$; but neither the quality of speech nor the 2-way interaction were significant. There were no significant effects on A'.

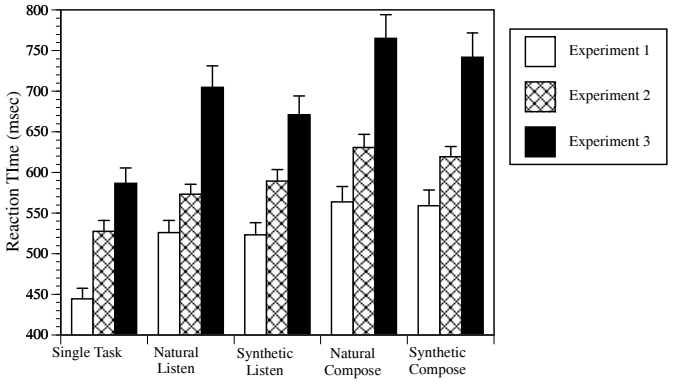

Figure 1. DRT RT across the three experiments.

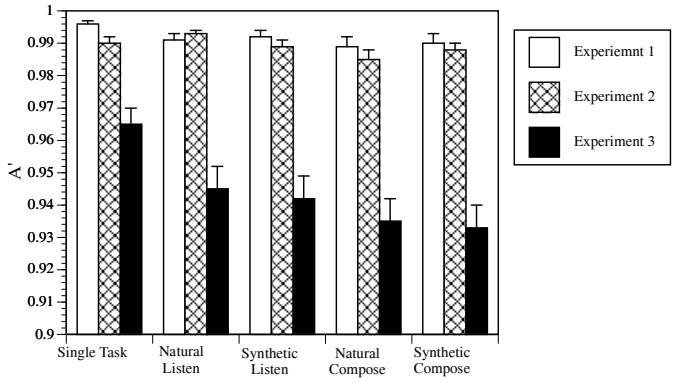

Figure 2. DRT A' across the three experiments.

EXPERIMENT 2

Experiment 2 assessed the mental workload of the five conditions used in Experiment 1 when they were paired with a driving task in a high-fidelity driving simulator.

Method

Participants. Forty-one participants (21 men and 20 women) completed the experiment. Participants ranged in age from 18 to 40 ($\bar{x}$ = 25.2 years). All reported having a valid driver's license and were fluent in English. Participant's years of driving experience ranged from 2.5 to 24 ($\bar{x}$ = 9 years). All participants owned a cellular phone and 84% reported that they used their phone regularly while driving. They were recruited via flyers posted on campus bulletin boards and via word of mouth within the community. Eligible participants reported a clean driving history (e.g., no at-fault accidents in the past five years). They received $15/hour in compensation for their participation in the experiment.

Equipment. In addition to the equipment used in Experiment 1, the present study used a fixed-base high fidelity driving simulator (made by L-3 Communications) with high-resolution displays providing a 180-degree field of view. The dashboard instrumentation, steering wheel, gas, and brake pedals were from a Ford Crown Victoria sedan with an automatic transmission. The simulator incorporated vehicle dynamics, traffic-scenario, and road-surface software to provide realistic scenes and traffic conditions. All other equipment was identical to Experiment 1.

Procedure. The procedures used in Experiment 1 were also used in Experiment 2, with the following modifications. In Experiment 2, a car-following scenario was used whereby participants drove on a multilane freeway with moderate traffic (approximately 1500 vehicles/lane/hour) traveling at 65 miles per hour. Participants were instructed not to change lanes or pass the pace car, and were asked to maintain a 2-second following distance behind the pace car. They were given a 5-minute practice session to familiarize themselves with the driving simulator and pace vehicle following distance.

Results

The RT and accuracy data for the DRT task are plotted in Figure 1 and 2, respectively. A repeated measures ANOVA found that RT increased across condition, $F(4, 160) = 28.49$, $p < .01$, partial $\eta^2 = .41$, and that A' decreased across condition, $F(4, 160) = 2.43$, $p = .05$, partial $\eta^2 = .06$. An analysis restricted to the 2 (speech quality: natural vs. synthetic) X 2 (task type: listen vs. listen + compose) factorial found a significant effect of task type on RT, $F(1, 40) = 6.32$, $p < .01$, partial $\eta^2 = .13$ and the 2-way interaction was significant $F(1, 40) = 6.83$, $p < .01$, partial $\eta^2 = .15$, but the main effect of quality of speech was not significant. There were no significant effects on A' in the 2X2 factorial analysis.

EXPERIMENT 3

Experiment 3 assessed the mental workload of the five conditions used in Experiment 1 when they were paired with driving an instrumented vehicle on residential streets. The instrumented vehicle provided a real-world driving environment within which to evaluate dual-task performance.

Method

Participants. Forty participants (23 men and 17 women) engaged in the experiment. Participants ranged in age from 20 to 39 ($\bar{x}$ = 26.1 years). All reported having a valid driver's license and were fluent in English. Participants' years of driving experience ranged from 2 to 24 years ($\bar{x}$ = 9.9 years). All participants owned a cellular phone and 89% reported that they used their phone regularly while driving. Recruitment techniques and demographic criteria were identical to Experiment 1 except that the Division of Risk Management Department at the University of Utah ran a Motor Vehicles Record report on each participant to ensure a clean driving history (e.g. no at fault accidents in the past five years).

Equipment. In addition to the equipment used in Experiment 1, the present study used an instrumented 2010 Subaru Outback. All other equipment was identical to Experiment 1 except instead of the speakers from experiments 1 and 2, we presented audio through the vehicle's stereo system. It is important to note here that the light from the DRT was bright enough to be seen outside during the daytime.

Procedure. The procedures used in Experiment 1 were also used in Experiment 3. The experiment was conducted in the daytime and any participants who had appointments during periods of inclement weather were rescheduled. Before beginning the study, the driver was familiarized with the controls of the instrumented vehicle, adjusted the mirrors and seat, and was informed of the tasks to be completed while driving. Next, participants drove one circuit on a 4.3 km loop in the Avenues section of Salt Lake City, UT in order to become familiar with the route itself. The route was in a residential driving environment, which contained seven all-way controlled stop signs, one two-way stop sign, and two stoplights. A research assistant and an experimenter accompanied the participant in the vehicle at all times. The research assistant sat in the rear of the vehicle. The experimenter sat in the front passenger seat and had access to

a redundant braking system and notified the driver of any potential roadway hazards. Participants were trained for each condition while stopped on the side of the road.

The driver's task was to follow the route defined above while complying with all local traffic rules, including a 25 mph speed restriction. Throughout each condition, the driver responded to the DRT. Each condition lasted approximately 9 minutes, the average time required to make one loop around the 4.3 km track.

Results

The RT and accuracy data for the DRT task are plotted in Figures 1 and 2, respectively. A repeated measures ANOVA found that RT increased across condition, $F(4, 156) = 27.16$, $p < .01$, partial $\eta^2 = .41$, and that A' decreased across condition, $F(4, 156) = 8.30$, $p < .01$, partial $\eta^2 = .18$. An analysis restricted to the 2 (speech quality: natural vs. synthetic) X 2 (task type: listen vs. listen + compose) factorial found a significant effect of task type on RT, $F(1, 39) = 8.11$, $p < .01$, partial $\eta^2 = .17$ and the 2-way interaction was significant $F(1, 39) = 15.90$, $p < .01$, partial $\eta^2 = .29$, but the main effect of quality of speech was not significant. There were no significant effects on A' in the 2X2 factorial analysis.

Finally, an ANOVA performed on the DRT data across the three experiments found that RT increased, $F(2, 123) = 23.7$, $p < .01$, partial $\eta^2 = .28$ and A' decreased, $F(2, 123) = 78.7$, $p < .01$, partial $\eta^2 = .56$, from Experiment 1 to 3. Overall, there was consistent agreement across experiments. If anything, the controlled laboratory- and simulator-based studies would appear to provide a more conservative estimate of the impairments to driving associated with in-vehicle technology use.

GENERAL DISCUSSION

The present research examined the factors driving the mental workload associated with voice-based IVIS interactions to send and receive textual information. We found that modern computerized synthetic speech did not impose extra cognitive workload upon the driver as compared to a natural/human voice. Prior research (Harbluk, 2005; Jamson et al., 2004; Lee et al., 2001; Raney, Harbluk & Noy. 2005) had found that synthetic speech led to higher levels of cognitive workload. The computerized text-to-speech algorithms have evidently improved so that they are no longer a major source of cognitive workload. With respect to cognitive workload, there is little more to be gained by improving the quality of synthetic speech.

In the listening conditions of our experiments (i.e., conditions 2 and 4), the participant was required to issue voice commands (e.g., *Repeat*, *Delete*, and *Next Message*) in addition to listening to the textual messages. The requirement to issue voice-commands introduced a modicum of speech production as compared to a "pure" listen only condition, such as listening to a book on tape. Our prior research has found that listening to a book on tape with the foreknowledge that participants would be quizzed on the content resulted in a workload significantly lower than that of talking on a cell phone (e.g., 1.7 on our 5-point scale for the book on tape, and a 2.4 for the handheld cellphone, see Strayer et al., 2015). Thus, the listen only conditions of the current research resulted in workload levels midway between listening to a book on tape and talking on a cell phone. The elevated workload relative to that associated with just listening to a book on tape is most likely due to participants issuing voice commands.

It is noteworthy that the speech-to-text system used in the present research had perfect reliability, with no errors in translation; however, current real-world systems often introduce errors that drive workload levels considerably higher (e.g., Strayer et al., 2015). However, even in this best case (i.e., perfect reliability), the workload was significantly elevated compared to just listening to the textual messages. This finding indicates that the speech production aspect of voice-based IVIS interactions is a significant source of cognitive workload and should be taken into consideration with regard to the overall workload of the driver (i.e., voice-based IVIS interactions are cognitively taxing).

Finally, moving from the laboratory to the driving simulator to the instrumented vehicle increased the intercept of the cognitive workload functions; however, similar effects of condition were obtained for DRT dependent measures. This experimental cross-validation establishes that the effects obtained in the simulator generalize to on-road driving. We did not collect data on the behavior of the vehicle (e.g lateral deviation of lane position, steering angle variance, etc.), but we did notice an increase of overall mental workload while operating a vehicle versus driving in a simulator.

ACKNOWLEDGMENTS

This research was supported by a grant from the AAA Foundation for Traffic Safety.

REFERENCES

Bergen, B., Medeiros-Ward, N., Wheeler, K., Drews, F., & Strayer, D. L. (2014). The crosstalk hypothesis: Language interferes with driving because of modality-specific mental simulation. *Journal of Experimental Psychology: General, 142*, 119-130.

Harbluk, L., & Noy, I. (2002). *The impact of cognitive distraction on driver visual behaviour and vehicle control.* Canada: Ergonomics Division, Road Safety Directorate and Motor Vehicle Regulation Directorate.

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention*, 372-379.

ISO. (2012). Road vehicles -- Transport information and control systems -- Detection-Response Task (DRT) for assessing selective attention in driving. ISO TC 22 SC 13 N17488 (Working Draft). *Under development by Working Group 8 of ISO TC22, SC 13*.

Jamson, A. H., Westerman, S. J., Hockey, G. R. J., & Carsten, O. M. (2004). Speech-based e-mail and driver behavior: Effects of an in-vehicle message system interface. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 625-639.

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. *Proceedings of ACM SIG-CHI '83 Human Factors in Computing Systems* (pp. 193-196). Boston: New York, ACM.

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interactions with in-vehicle computers; The effect of speech-based e-mail on drivers' attention and roadway. *Human Factors, 43*, 631-640.

NeoSpeech. (2012). NeoSpeech Text-to-Speech voices [computer software]. Santa Clara, CA. Available from http://www.neospeech.com/.

Parasuram, R., & Davies, D. R. (1984). Varieties of Attention. Academic Press.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1-12), 125-126.

Ranney, T. A., Harbluk, J. L., & Noy, Y. I. (2005). Effects of voice technology on test track driving performance: Implications for driver distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(2), 439-454.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors, 53*, 1300-1324.