MENTAL WORKLOAD OF VOICE INTERACTIONS WITH 6 REAL-WORLD DRIVER INTERFACES

Joel M. Cooper[1] & David L. Strayer[2]
Precision Driving Research[1]
University of Utah[2]
Salt Lake City, Utah, USA
joel.cpr@gmail.com

Summary: Hands-free voice interaction is an increasingly common option in new vehicles. Recent research suggests that hands-free interactions with speech-to-text systems may require significantly more cognitive effort than previously anticipated. This high level of mental workload may both keep drivers from using the technology and potentially create additional safety concerns for the driver. However, little prior research has measured the cognitive demands of simple voice based tasks using real-world systems. The current study evaluated the mental demands of a small set of auditory-vocal vehicle commands using five 2013 and one 2012 model year OEM infotainment systems. Results indicate that well executed voice systems impose little additional cognitive demand while poorly executed systems may significantly elevate workload.

INTRODUCTION

Driving is a highly complex activity that requires a significant amount of visual and cognitive attention to be performed successfully. In order to allow drivers to maintain their eyes on the driving task, nearly every vehicle sold in the US and Europe can now be optionally equipped with a hands-free voice system. Using structured voice commands, drivers can access functions as varied as contact and number dialing, music selection, destination entry, and even climate adjustment. Hands-free, voice activated convenience features are a natural development in vehicle safety. Yet, a sizable body of literature cautions that even auditory-vocal tasks may lead to unexpected task demands (Delogu, Conte, & Sementina, 1998; Harbluk & Lalande, 2005; Recarte & Nunes, 2007). Research on cognitive distraction suggests that even if a driver's eyes remain on the forward roadway, their ability to detect and respond to targets within their visual field may be impaired if their cognitive focus is not also on the forward roadway (Hyman, et al. 2010; Simons, 2000). Furthermore, prior research suggests that the cognitive demands associated with speech-to-text system interactions may be higher than other common auditory-vocal tasks such as talking on a cell phone or listening to the radio (Strayer et al. 2013 & 2014). However, prior research has almost exclusively looked at voice based interactions with systems that differ in some way from those used in actual vehicles. Thus, it is not clear whether or how estimates of cognitive task load obtained using synthetic systems might apply to real world systems.

One ongoing effort to understand cognitive distraction in vehicles is being led by Strayer and colleagues (See Strayer et al., 2013, and Strayer et al., 2014). In their research Strayer et al. (2013) investigated a comprehensive set of common cognitive tasks across various data collection environments, using a complementary set of primary, secondary, physiological, and

subjective measures. Through the use of a consistent protocol, Strayer at al. (2013) completed a laboratory, simulator, and on road assessment of common auditory-vocal tasks performed by drivers. This allowed a variety of every-day secondary cognitive driving tasks to be directly compared. Results indicated that the tasks could be roughly clustered into 3 distinct groups. This clustering is based both statistical and practical task differentiation. Listening to the radio or a book on tape led to low levels of cognitive demand, similar to baseline driving. Conversation, whether with a passenger or through a hand-held or hands-free cellular phone led to slightly elevated levels of mental workload, comprising the second group. Finally, a synthetic speech-to-text email interaction system led to still elevated levels of mental workload, forming a separate workload category that was greater than the other two.

In order to create the most broadly applicable results, the speech-to-text tasks evaluated by Strayer et al. (2013) and Strayer et al. (2014) were carefully scripted and controlled using functional mock-ups rather than actual systems. Thus, it is unknown how the cognitive demands reported in that research might compare to similar tasks using actual vehicle systems. The purpose of the current study is to assess cognitive workload associated with selected speech-to-text functionality of actual OEM systems in order to extend results from Strayer et al. (2013) and Strayer et al. (2014) to interactions with commonly available vehicle systems. This research addresses the following questions: How cognitively demanding are auditory-vocal vehicle interactions with actual OEM systems? How similar/dissimilar is the cognitive demand associated with different OEM systems? How do the cognitive demands of OEM speech interactions compare to the cognitive demands of the various tasks evaluated in Strayer et al. (2013, 2014)?

METHOD

Following Institutional Review Board approval, participants were recruited through ads placed on online local classifieds websites, flyers posted around a local university campus, and by word of mouth. All data were collected from December 5th through December 10th of 2013. A total of 36 participants completed this research (18 male, 18 female). Participants ranged in age from 22 to 36 years (M = 28.1, SD = 3.89). All participants were required to have a valid driver's license and have fewer than 2 accidents in the past 2 years. Additionally, participants were selectively recruited to balance gender.

Systems from six different vehicle manufacturers were investigated. All were chosen because of their popularity and hands-free voice controlled functionality. These cars included a 2013 Ford Explorer Limited featuring SYNC with MyFord Touch, a 2013 Chevy Cruz Eco featuring Chevrolet MyLink, a 2013 Chrysler 300 with the Uconnect System, a 2012 Toyota Prius V Three with Entune, a 2013 Mercedes E350 featuring the COMMAND® system, and a 2013 Hyundai Sonata SE with a Blue Link Telematics System. These six systems had many features in common, including steering wheel-mounted controls, Bluetooth phone pairing, voice-activated music functions, voice-activated CD playing, voice-controlled satellite radio, hands-free calling, and access to calling features (i.e. phonebook, call log, etc.). An Alcatel One-Touch Fierce phone

was paired via Bluetooth to each of the voice-controlled systems. Phones were placed in an out of the way location and were never directly viewed or manipulated by participants during the study.

During all phases of testing, participants wore a head-mounted reaction time assessment device. These Detection Response Task (DRT) devices were assembled for the purpose of this study and follow the specifications outlined in ISO WD 17488 rev 10.1 (ISO, 2012). The devices consisted of an LED light mounted to a flexible arm that was connected to a headband. The light was positioned in the periphery of the participants' left eye so that it could be seen while looking forward at the road but did not obstruct their view. The devices featured a simple user interface which was optimized to assess mental workload. The precise configuration used in this research differed from the draft standard in two ways. First, the stimulus lights were configured to flash red or green every 3-5 seconds. Each time a light turned on, there was a 60 percent chance it would be red and 40 percent chance it would be green. Second, participants were given a response button and instructed to respond only to green lights as quickly as possible by clicking the button against the steering wheel. Timing was controlled on Asus Transformer Book T100s with quad-core Intel® Atom™ processors running at 1.33GHz.

Participants were outfitted with a Zephyr BioHarness 3 heart rate monitor. These professional quality heart rate monitors, and their accompanying software algorithms, have been tested to be within +/- 1 beat per minute of accuracy. The BioHarness 3 collects and stores comprehensive physiological data about the person wearing the monitor, including heart rate, heart-rate variability, breathing rate, posture, and activity level. The monitors attach around the chest with a flexible strap. For the purposes of this study, only Heart Rate, operationalized as the beats per minute, was used. Prior research suggests that of the many potential cardiovascular measures, Heart Rate is the most sensitive to mental workload (Mehler, Reimer, & Wang, 2011). Following each drive, participants completed a written form of the NASA TLX, a subjective workload rating scale.

Prior to data collection in the vehicles, participants were able to familiarize themselves with the course by driving one circuit. Each loop took approximately 7-9 minutes depending on stop lights, driving speed, and traffic at stop signs. After familiarization with the course, participants received instructions about the DRT task and were given the opportunity to practice while sitting in a parked vehicle. During experimentation, a high and low workload baseline task was given. The low workload baseline task consisted of routine driving in a single-task condition. The high workload baseline consisted of driving while performing the Operation Span (Ospan) task used by Strayer et al. (2013, 2014).

Condition A was a single-task drive which provided a baseline for the assessment of cognitive workload in the other conditions. In condition B participants drove while concurrently performing the Ospan task. Conditions C through H were in-vehicle system interactions. Each voice interaction condition corresponded with one of the six vehicles (i.e. condition C corresponded to vehicle one, condition D corresponded to vehicle two, etc.).Each participant completed just one drive in the single-task condition and one drive in the Ospan condition. These

two baseline assessments were collected using the same vehicle. Thus, a total of 36 (6 in each vehicle) low workload baseline drives and 36 (6 in each vehicle) high workload baseline drives were completed. The order of the 8 conditions was counterbalanced across subjects.
Prior to driving in each vehicle, participants were given instructions on how to complete the calling and music tasks in the vehicle and practiced with the system until they could complete the tasks without error. The Voice interactions with each of the six vehicles were functionally equivalent; the only thing that differed between vehicles was the precise sequence of commands that were required to complete the tasks. Each of the 6 tasks that were completed at a specific geographical location on the course. When the participant reached pre-specified locations on the course, the facilitator gave an instruction to begin the indicated task. Participants were not told where on the course the new tasks would be given but the task onset location remained constant for all interactions. All tasks began with the press of a steering wheel mounted button to initialize the voice command systems. Once initiated, each of the tasks was completed through auditory + vocal system interactions. System interactions alternated between completing a phone related task and a music functions task. They were as follows: Task 1: "Call from your contacts ____ on his cell", Task 2: "Tune your radio to 99.5 FM," once completed: "tune your radio to 1320 AM", Task 3: "Dial your own phone number", Task 4: "Play your CD*", once completed: "tune your radio to 98.7 FM", Task 5: "Call from your contacts _____ on his cell", Task 6: "Tune your radio to 103.5 FM," once completed: "play your CD.

RESULTS

Core Measures

Heart Rate. A one-way repeated measures ANOVA was used to test for differences in Heart Rate among the 8 experimental conditions. The overall test was significant, $F(7, 245) = 5.97$, $p < .001$, partial $\eta^2 = .146$, indicating that the measurement of heart-rate in the vehicle was sensitive to the experimental conditions. The range of the mean Heart Rate values between the low and high workload baseline conditions was 3.17. Pairwise comparisons indicated that mean Heart Rate was significantly lower during music selection and call placement using Toyota's Entune system than with Hyundai's Blue Link, Chevrolet's MyLink, Chrysler's Uconnect, and Mercedes' COMMAND systems (all p's < .05). On the flip side, music selection and call placement using Cheverolet's MyLink system led to a mean Heart Rate that was greater than all other vehicle systems except Mercedes COMMAND (all p's < .05). In short, Toyota's Entune system elicited the lowest mean Heart Rate and Chevrolet's MyLink system elicited the highest mean Heart Rate, while all other systems were statistically undifferentiated.
NASA TLX. The six subscales of the NASA TLX were combined through an equally weighted average. The resulting aggregate scores were then subjected to a one-way repeated-measures ANOVA. Results indicated a highly significant main effect of experimental condition, $F(7, 245) = 56.3$, $p < .001$, partial $\eta^2 = .62$. Pairwise comparisons of the 8 experimental conditions revealed

a pattern that was very similar to that obtained from the Heart Rate measure reported above. In general, music selection and call placement on all of the systems elicited responses that were differentiated from the low and high workload baselines. Toyota's Entune and Hyundai's Blue Link systems were rated slightly more demanding than the low workload baseline; the Chrysler, Ford, and Mercedes systems were rated somewhat more demanding; and finally, music selection and call placement on Chevrolet's MyLink system were rated the most demanding, but still substantially less than the high workload baseline.

DRT Reaction Time. A one-way repeated measures ANOVA indicated that the music selection and call placement tasks significantly affected reaction time, $F(7, 245) = 16.1$, $p < .000$, partial $\eta^2 = .315$. Pairwise comparisons indicated a very similar pattern to that seen in the Heart Rate and NASA TLX measures. One exception was that mean reaction time while selecting music or placing a call using the MyFord Touch system was significantly slower than with all other systems. Otherwise, the same consistent pattern was observed, reaction times while interacting with any of the voice based systems was significantly slower than the low workload baseline but faster than the high workload baseline. Again, the one exception was that reaction times while using the MyFord Touch system were nearly as delayed as those observed in the high workload baseline.

Workload Rating Scale

Based on the combinatorial procedure that was presented by Strayer et al. (2013), Heart Rate, NASA TLX, and DRT Reaction Time data were Z-transformed and linearly combined with equal weighting to generate a summary variable (Please refer to Strayer et al. (2103) for a complete description of this analytic approach). This summary variable was then analyzed using a one-way repeated measures ANOVA revealing a significant overall effect of condition, $F(7, 245) = 49.4$, $p < .000$, partial $\eta^2 = .585$. A table with the mean Z-scores for each of the measures and experimental conditions is presented below. Pairwise comparisons for the measures cleanly distinguished 6 groups (see Figure 1). As expected, the single-task and Ospan conditions were statistically different from all of the voice based system interactions. On the low end, Toyota's Entune system produced moderately more mental workload than the single-task condition. Based on our prior findings, this resulted in a workload estimate that is similar to listening to the radio or an audio book. Music selection and call placement using Hyundai's Blue Link system led to a significant increase in workload from the Toyota system, a level which was similar to holding conversation over a cell phone or with a passenger. Music selection and call placement using the Chrysler, Ford, and Mercedes systems led to a level of workload that was similar to the error free speech-to-text system evaluated in Phase-1. Finally, music selection and call placement using Chevrolet's MyLink system led to a level of workload that was greater than any of the other system interactions.

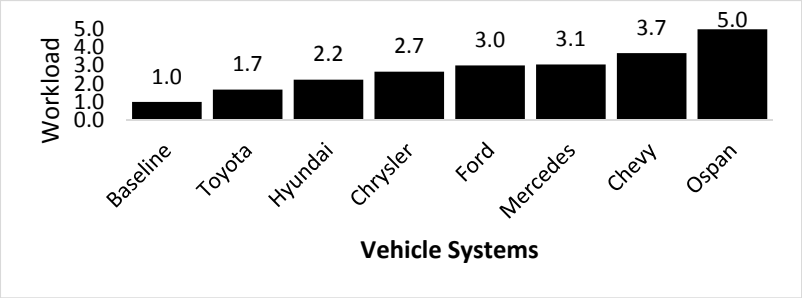

Figure 1. Mental Workload Scale for each of the 8 research conditions

DISCUSSION

The purpose of this study was to evaluate the mental demands of simple auditory-vocal vehicle interactions across five 2013 and one 2012 model year OEM infotainment systems. This research was designed to address three novel questions which were: How cognitively demanding are auditory/vocal vehicle interactions with actual OEM systems? How similar/dissimilar is the cognitive demand associated with different OEM systems? How do the cognitive demands of OEM speech interactions compare to the cognitive demands of the various tasks evaluated in Strayer et al. (2013, 2014)?

Results obtained in this investigation indicate that simple auditory-vocal interactions with vehicles may significantly elevate mental workload in drivers. On the standardized rating scale developed by Strayer et al. (2013), and refined in 2014, a mean demand score of approximately 3 was observed across the 6 OEM systems. A score of 3 is midway between the workload associated with the single-task baseline and the Ospan mental math condition and indicates a moderate level of cognitive load. Workload ratings for all voice interactions were greater than that observed in the Single-Task driving condition and less than that observed in the Ospan task. Thus, all systems imposed some demand but no system imposed more demand than the Ospan math task.

In the best case, evaluated voice commands using Toyota's Entune system imposed modest additional demands as compared to the single-task baseline condition. In the worst case, those same activities using Chevy's MyLink system imposed mental demands that were approaching the high workload baseline (Ospan mental math). Not surprisingly, the most critical element of workload appeared to be the duration of the interaction. For the tasks selected in this analysis, Toyota's Entune system required the least amount of interaction time while Chevrolet's MyLink required the most. Most of the auditory-vocal interactions evaluated in this research were more demanding than the 3 conversation tasks evaluated in by Strayer et al. (2013). The exception, of course, was that call voice interactions using Toyota's Entune system were less demanding than the book-on-tape condition used by Strayer et al. (2013). In general, most systems elicited about the same, or less, mental workload than the speech-to-text task evaluated in by Strayer et al. (2013). The exception here was that the Chevy MyLink system was significantly more demanding (see Strayer et al., 2014 for additional research contrasts). Overall, observed mental

workload during voice interactions ranged from being as low as listening to audio media to higher than any of the non-math tasks earlier measured.

Currently, the association between cognitive driver distraction and safety risk is not well understood. Because of the complex manifestations of cognitive driver distraction, it is not clear whether and how the findings from this research might result in changes to real world safety risk. As of yet, there is no unambiguous correspondence between variations in mental workload and the actual risk of a crash. Clearly, additional research is needed to gain a better understanding of the crash risks of various cognitive tasks. This research evaluated a call placement and a music selection task that were similar implemented across each of the vehicles included in this study. This restricted set of tasks allowed us to make direct comparisons between vehicles. However, it is unknown how the results might generalize to other in-vehicle voice commands afforded by the various vehicle systems. Critically, this research does not allow us to generalize to the full set of functions afforded by any of the systems that we evaluated. If we had evaluated a different set of tasks it is highly likely that the relative performance of the systems would have changed.

REFERENCES

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, *24*(2), 153-168.

Harbluk, J. L., & Lalande, S. (2005). Performing e-mail tasks while driving: The impact of speech-based tasks on visual detection. In *Proceedings of the 3rd International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design* (pp. 304-310).

Hyman, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, *24*(5), 597-607.

Mehler, B., Reimer, B., & Wang, Y. (2011). A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive load. Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. Olympic Valley – Lake Tahoe, California. Jun 27-30.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119-137.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in cognitive sciences*, *4*(4), 147-155.

Strayer, D. L., Cooper, J .M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013).Measuring Cognitive Distractionsin the Automobile. AAA Foundation for Traffic Safety, Washington, DC.

Strayer, D. L., Turrill, J., Coleman, J., & Cooper, J.M. (2014). Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies. AAA Foundation for Traffic Safety, Washington, DC.