A PRELIMINARY ASSESSMENT OF PERCEIVED AND OBJECTIVELY SCALED WORKLOAD OF A VOICE-BASED DRIVER INTERFACE

Bryan Reimer, Bruce Mehler, Hale McAnulty, Daniel Munger, Alea Mehler,
Enrique Abdon Garcia Perez, Thomas Manhardt & Joseph F. Coughlin
The Massachusetts Institute of Technology AgeLab
New England University Transportation Center, Cambridge, Massachusetts, USA
Email: reimer@mit.edu

Summary: Interaction with a voice-command interface for radio control, destination entry, MP3 song selection, and phone dialing was assessed along with traditional manual radio control and a multi-level audio–verbal calibration task (n-back) on-road in 60 drivers. Subjective workload, compensatory behavior, and physiological indices of cognitive workload suggest that there may be both potential benefits and cautions in the implementation of a representative production level interface.

INTRODUCTION

Over the past several years there has been a shift in automotive driver-vehicle interfaces (DVIs) from purely visual-manual interactions to voice-based or voice-assisted interaction. However, few DVI functions are presently controlled entirely through voice commands. At minimum, most if not all current voice-based in-vehicle systems require manipulation of a “push-to-talk” button. Further, many systems that are assisted by voice commands continue to include traditional visual-manual interactions as well. For instance, in in-vehicle systems where an address is entered into a navigator verbally, confirmation of the address may be required on an in-vehicle display or a specific address may need to be selected from a list of candidates on the display with either a button press or additional voice interactions. The demands associated with most, if not all, current voice systems are therefore multimodal.

A number of studies have focused on assessing the demands of in-vehicle voice applications. In a test track study of 36 participants across three age groups, Ranney, Mazzae, Baldwin and Salaani (2007) investigated a set of voice driven navigation tasks through a “Wizard of Oz” 511 system. Deteriorations in all aspects of driving performance were found when drivers engaged with the simulated voice interface. In a field driving experiment of 12 subjects, Zheng, McDonald and Pickering (2008) assessed driver’s engagement with three different voice interfaces using vehicle and task performance measures and visual behaviors. Results show that the voice interface with cluster based text prompts had relative advantages over central based text prompts and traditional voice interfaces with no text prompts. Owens, McLaughlin and Sudweeks (2010) looked at the behavior of 21 drivers, approximately half older, who were current owners of vehicles with the SYNC system in a field study. The experiment was conducted in a 2010 Mercury Mariner equipped with the Ford SYNC™ voice interface and investigated the impact of different input modalities (voice control and handheld device) on phone dialing, conversations, and playing a music track. Results show that the voice interface interfered more with vehicle control than direct interaction with the hand held device. Various eye behaviors suggest a more optimal orientation towards the road with the voice interface. Finally, self-reported mental demand score of the NASA TLX was lower with the voice interface.

An underlying aspect of the studies discussed above is the focus on a driver's visual orientation, driving performance and self-reported demand while engaged in various types of voice dialog or in comparison to handheld operations. These are clearly key factors to consider in assessing the demands associated with voice systems. However, they may not fully provide an objective rating of the non-visual (cognitive) demands of the systems. In essence, although voice-based systems are intended to help keep drivers' eyes on the road, little is known about the "holistic", visual, manipulative and cognitive demand that the systems place on the driver.

Research suggests that physiological indices of workload reflect an individual's investment of cognitive resources corresponding to task demand (Brookhuis & de Waard, 1993; Lenneman & Backs, 2009; Mehler, Reimer, Coughlin, & Dusek, 2009; Reimer & Mehler, 2011; Mehler, Reimer & Coughlin, 2012; Wilson, 2002). Physiological measures have been shown to be sensitive to subtle increases in demand prior to overt breakdowns in driving performance are observed (Mehler, Reimer, Coughlin, & Dusek, 2009). In contrast to earlier work where demands exceeded a driver's capability or willingness to engage in a secondary activity (Engström et al. 2005), heart rate and skin conductance have been shown to scale relatively linearly with an increase in cognitive demand from an auditory presentation – verbal response working memory task (n-back) (Mehler, Reimer & Coughlin, 2012). In essence, the three levels of the n-back task create a three stage ruler, e.g. low, moderate and high, against which the relative demand of other cognitive activities can be objectively and non-invasively scaled.

The degree to which demand placed on the driver through artificial secondary tasks such as the n-back relate to the demands of voice-based interactions or other non-visual activities is an open question. However, we hypothesize that the demands of voice-based interactions will fall between the lowest (0-back) and highest (2-back) levels of the secondary task. This hypothesis is framed by the 0-back task, a simple mirroring activity consisting of verbally repeating back auditorily presented single digit numbers, and the 2-back task, a more demanding activity that taxes working memory and which approaches the limits of most drivers' spare capacity. This report focuses on detailing preliminary results from a field study conducted to measure drivers' visual behavior, physiological arousal and perceived workload while engaging in a number of tasks with a voice-based in-vehicle HMI, an implementation of the manual radio tuning reference task (AAM, 2006), and the three levels of the n-back task. The data was collected during a field experiment in which participants were given detailed training on the operation of the vehicle systems under study prior to the assessment of behaviors while driving.

METHOD

Subjects

Recruitment drew from the greater Boston area using online and newspaper advertisements and consisted of two age groups, 20-29 and 60-69 years. Participants were required to read and sign an institutional review board approved informed consent form, to present a valid driver's license and attest to having had their license for more than three years, to driving on average three or more times per week, and be in self-reported reasonably good health for their age. Additional screening for various health and medication considerations that might impact safety or physiological reactivity were carried out. Compensation was $90.

Apparatus

The study was conducted in a 2010 Lincoln MKS with a SYNC™ voice interface. The interface was engaged using a “push-to-talk” button on the steering wheel. The vehicle was instrumented for time synchronized recording of vehicle information from the controller area network (CAN) bus, a MEDAC System/3 physiology monitoring unit, FaceLAB® 5.0 eye tracking, cameras for capturing driver behavior and vehicle surroundings, and GPS tracking (see Mehler, Reimer and Coughlin (2012) for details on physiological recording). Subjective workload ratings were obtained using a single global rating per task on a scale consisting of 21 equally spaced dots oriented horizontally along a 10cm line with the numbers 0 through 10 equally spaced below the dots and end points labeled “Low” and “High” on the left and right respectively. A research associate was seated in the rear of the vehicle for ensuring safe operation.

Secondary tasks & procedure

There were six in-vehicle task areas: manual control of the radio, voice command control of the radio, navigation system destination entry, song selection (from an MP3 storage device), stored phone number dialing, and an auditory presentation / verbal response calibration task (n-back). Each task type was presented twice. N-back tasks consisted of single sets of the digits 0-9 presented in random order and were 30 seconds in duration (see Mehler, Reimer and Dusek, (2011) for details on task training). Basic radio interaction was modeled on guidelines established by the AAM (2006) and protocols developed as part of CAMP (Angell et al., 2006). An “easy” task consisted of changing a station by the single step of pressing a specified preset button in the radio-manual control version. The corresponding voice-command system interaction involved 3 steps (1 voice button press and 2 verbal inputs / confirmations, i.e. “preset-1”, “yes”). The voice system offered an advanced option for dropping confirmatory responses which would have reduced the number of steps in some of the interactions. Since this was not the default mode, it was not used. The “harder” radio-manual task required 4 steps (pressing the volume control to turn the radio on, pressing a ‘RADIO’ button to access the band selection, pressing a touch screen band button (i.e. ‘FM2’), and rotating the tuning knob to the specified frequency number). The corresponding voice-command interaction also involved 4 steps (1 voice button press and 3 verbal inputs / confirmations, i.e. “Radio”, “100.7”, “yes”).

Voice-command interaction with the navigation system consisted of two subtasks, entry of a street address and cancelation of the route request. Assuming there were no errors in interaction with the system, address entry required 11 steps (2 button presses and 9 verbal inputs / confirmations) and cancellation required 3 (1 button press and 2 verbal inputs / confirmations). For the song task, a USB drive containing MP3 files was pre-connected to the system. The primary task required 3 steps (1 button press, saying “USB” and then “Play Artist xxx’). Following this, participants were given a selection to request that did not exist on the device. This task was presented to observe how drivers interacted with the system when it was unable to comply with a request. The final task involved placing a phone call to a stored number using the voice interface. This required 3 steps (1 button, saying “phone” and then “call contact x”).
The driving portion of the study was conducted on roadways in the greater Boston area and divided into four segments. The first was an adaptation period of approximately 10 minutes of urban driving to reach interstate highway I-93 and continued north for an additional 20 minutes

or so to the I-495 intersection. The second consisted of driving south on I-495 to the exit 19 rest area and averaged approximately 40 minutes. The third was from the rest area back north on I-495 to I-93 and the fourth was the return on I-93 south. The radio-manual, radio-voice, navigation-voice, and song selection-voice tasks were presented in a counter-balanced order during segments two and three with the exception that the radio-manual and radio-voice tasks were never presented in the same segment. The 3 levels of the n-back were presented twice, once each in the middle of segments two and three; ordering of the levels was randomized. The phone task was always presented during segment four. Detailed training was provided in the MIT parking lot on the tasks to be completed during the first half of the drive. Training and practice on the remaining tasks were provided during the rest-stop between segments two and three.

RESULTS

Following exclusion of cases due to protocol, technical, weather, traffic, or other considerations, a total of 60 particpants, equally balanced by gender and age group, were used in the analysis. Mean age of the younger sample was 24.4 years (SD= 2.8) and 65.2 (SD= 3.1) for the older. Self-report workload ratings are summarized in Figure 1. An ANOVA with repeated measures indicates that the tasks different significantly in percieved workload ($F(11, 506) = 32.8$, $p < .001$).

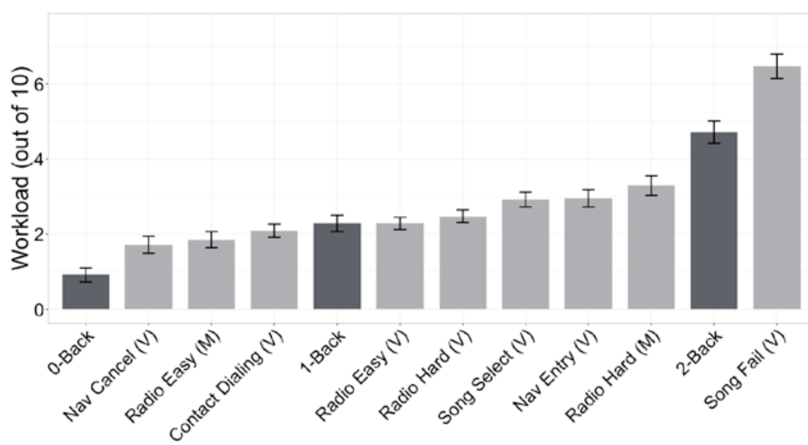

Figure 1. Mean self-reported workload on a low to high scale of 0 to 10 (with half point resolution). N-back reference tasks are denoted with darker bars. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions

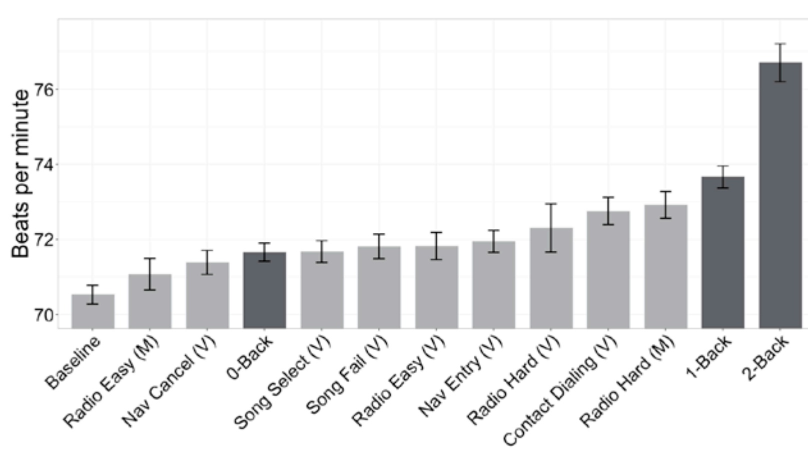

Figure 2. Mean heart rate during the reference period (Baseline) and each task

Two minute single task driving reference (Baseline) periods were pre-defined before each dual task period and combined for comparison with physiological and driving performance data during the dual task periods. Each task type was presented twice and the periods averaged for analysis. Heart rate data show that the tasks produced significantly different arousal effects ($F(12, 684) = 13.01$, $p < .001$, ANOVA with repeated measures)(Figure 2). Skin conductance levels also show significant differences between task types ($F(12, 576) = 4.72$, $p < .001$) (Figure 3). An analysis of mean speed during secondary task performance shows a significant effect of task ($F(12, 684) = 5.10$, $p < .001$) (Figure 4).

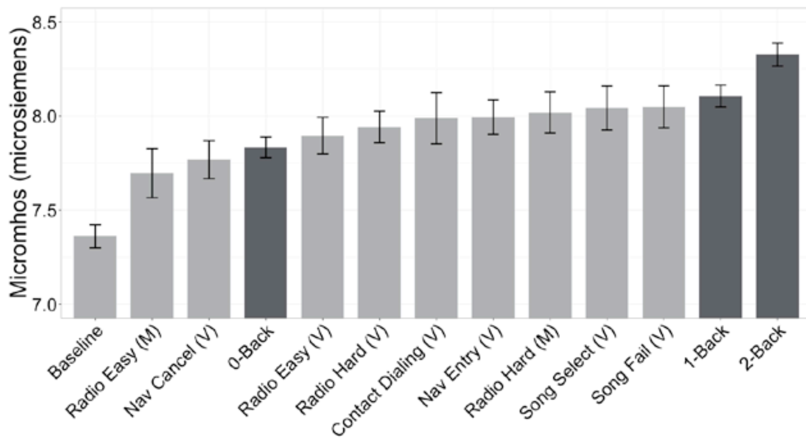

Figure 3. Mean skin conductance levels (SCL) during the reference period and each task

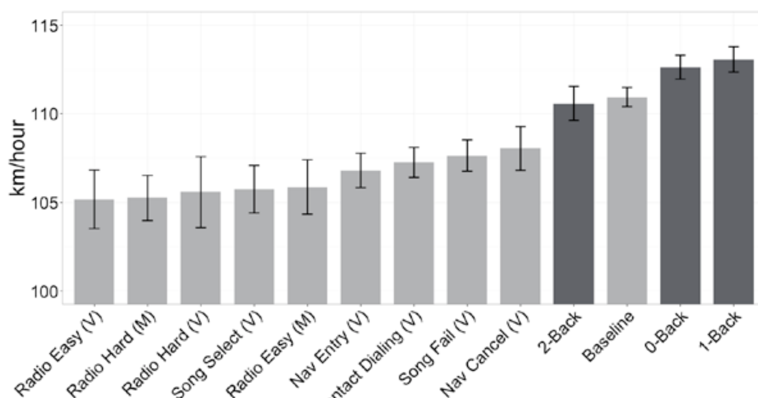

Figure 4. Mean vehicle speed during secondary task periods

DISCUSSION

This study employed a sample of 60 drivers equally balanced by gender and across two age groups (20-29 and 60-69 years) to assess interactions with a voice-command system during highway driving. An initial look at the data on perceived workload (subjective, self-report), compensatory behavior (relative speed), and objective estimates of cognitive workload (heart rate and SCL) suggests that there may be both potential benefits and cautions in the implementation of a production level voice-command interface. It is illustrative to consider various tasks and measures relative to the Radio-Manual Hard tuning task, which The Alliance guidelines (2006) recommend as a reference task for an upper bound of a generally acceptable level of secondary demand on a driver. This activity was given the highest subjective workload rating of all of the standard HMI interactions (except for the Song Fail, a deliberate test of a failure condition), resulted in the greatest compensatory behavior (lower speed), and was associated with the highest heart rate and a high level of SCL. The data for Radio-Voice method of carrying out this same task shows a lower subjective workload rating and comparable or nominally less speed compensation and physiological arousal measures.

A possibly more substantive advantage is suggested for destination entry. Manual entry of destination information is considered by some to exceed the level of demand that is acceptable while driving and this function is locked-out in many implementations while a vehicle is underway. Relative to the Radio-Manual Hard task, the voice-command method of entering a destination (Nav Entry) produced lower subjective workload, less compensatory speed behavior, and lower (heart rate) or equivalent (SCL) physiological indices of workload. In contrast, traditionally relatively easy secondary tasks such as selecting a pre-set radio station (Radio-Manual Easy), was given a higher subjective workload rating when operated using the voice-command interface (Radio-Voice Easy) and resulted in nominally higher physiological measures of workload (heart rate and SCL). Not surprisingly, this and other patterns suggest that a voice-command interface is not inherently superior or inferior to more traditional visual-manual interfaces. Careful analysis of where current implementations provide net enhancements and where they do not is likely to contribute to insights that can inform the overall optimization of multi-modal interface development.

The location of the 3-levels of the n-back task across the various measures aligns well for its proposed use as a cognitive workload calibration metric (Mehler, Reimer & Coughlin, 2012). In terms of subjective workload, the N-Back Easy (0-back) task received the lowest rating, N-Back Medium (1-back) was rated at an intermediate level across the tasks, and the N-Back Hard (2-back) was rated at the high end of the scale and above the Radio-Manual Hard reference task. Only the “failure” state (Song Fail), a task that was impossible to complete successfully, was given a higher subjective workload rating. In terms of physiological measures, all of the tasks scaled lower than both the medium and high levels of the n-back. Since, as a group, drivers decreased their speed during all of the HMI tasks, presumably to compensate for the added demand of the secondary tasks, effective cognitive workload (Mehler, Reimer & Zec, 2012) was maintained at or below the level of the 1-back task. Keeping effective cognitive load of an HMI task below the level of the 1-back might be a desirable design goal for highway driving.

While some of the voice-command results are encouraging, these data should be interpreted cautiously as other vehicle performance and eye behavior metrics need to be taken into account in developing a more complete understanding of overall impact, including visual distraction; analyses are ongoing. Similarly, breakdown of results by demographic considerations such as age, gender, and technology experience need to be examined in future reports.

ACKNOWLEDGMENTS

Acknowledgement is extended to The Santos Family Foundation and US DOT’s Region I New England University Transportation Center at MIT for providing the initial support for this project and to Toyota’s CSRC for providing funding to expand the original planned investigation. James Foley and Kazutoshi Ebe of CSRC provided valuable comment during the design of the study.

REFERENCES

Alliance of Automobile Manufacturers (2006). *Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems*. Driver Focus Telematics Working Group.

Angell, L., Auflick, J., Austria, P.A., Kochhar, D., Tijerina, L., Biever, W., et al. (2006). *Driver Workload Metrics Task 2 Final Report*. Washington, DC: U.S. DOT NHTSA.

Brookhuis, K. & de Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics,36(9),* 1099-1110.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F, 8*(2), 97-120.

Lenneman, J.K., & Backs, R.W. (2009). Cardiac autonomic control during simulated driving with a concurrent verbal working memory task. *Human Factors, 53*(3), 404-418.

Mehler, B., Reimer, B. & Coughlin, J.F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human Factors, 54(3),* 396-412.

Mehler, B., Reimer, B., Coughlin, J.F. & Dusek, J.A. (2009). The impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record, 2138, 6-12*.

Mehler, B., Reimer, B. & Dusek, J.A. (2011). *MIT AgeLab delayed digit recall task (n-back).* MIT AgeLab White Paper Number 2011–3B. Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B., & Zec, M. (2012). Defining workload in the context of driver state detection and HMI evaluation. *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications,* Portsmouth, NH, October 17-19, 2012, pp. 187-191.

Owens, J.M., McLaughlin, S.B., & Sudweeks, J. (2010). On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players. *SAE International Journal of Passenger Cars – Mechanical Systems, 3(1),* 734-743.

Ranney, T.A., Mazzae, E.N., Baldwin, G.H.S. & Salaani, M.K. (2007). *Characteristics of voice-based interfaces for in-vehicle systems and their effects on driving performance*. Washington, DC: U.S. DOT NHTSA.

Reimer, B. & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics, 54(10),* 932-942.

Wilson, G. F. (2002). Psychophysiological test methods and procedures. In S.G. Charlton & T.G. O'Brien (Eds.), *Handbook of Human Factors Testing and Evaluation* (pp. 127-156). Mahwah, NJ: Lawrence Erlbaum Associates.

Zheng, P., McDonald, M. & Pickering, C. (2008). Effects of intuitive voice interfaces on driving and in-vehicle task performance. *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China, October 12-15, 2008, pp. 610-615.