Jun 25th, 12:00 AM

A Secondary Assessment of the Impact of Voice Interface Turn Delays on Driver Attention and Arousal in Field Conditions

Thomas McWilliams
Massachusetts Institute of Technology (AgeLab), Cambridge, MA

Bryan Reimer
Massachusetts Institute of Technology (AgeLab), Cambridge, MA

Bruce Mehler
Massachusetts Institute of Technology (AgeLab), Cambridge, MA

Jonathan Dobres
Massachusetts Institute of Technology (AgeLab), Cambridge, MA

Hale McAnulty
Massachusetts Institute of Technology (AgeLab), Cambridge, MA

McWilliams, Thomas; Reimer, Bryan; Mehler, Bruce; Dobres, Jonathan; and McAnulty, Hale. A Secondary Assessment of the Impact of Voice Interface Turn Delays on Driver Attention and Arousal in Field Conditions. In: Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, June 22-25, 2015, Salt Lake City, Utah. Iowa City, IA: Public Policy Center, University of Iowa, 2015: 408-414. https://doi.org/10.17077/drivingassessment.1602

A SECONDARY ASSESSMENT OF THE IMPACT OF VOICE INTERFACE TURN DELAYS ON DRIVER ATTENTION AND AROUSAL IN FIELD CONDITIONS

Thomas McWilliams, Bryan Reimer*, Bruce Mehler, Jonathan Dobres, Hale McAnulty
MIT AgeLab & New England University Transportation Center
77 Massachusetts Avenue, E40-291 – Cambridge, MA 02139
*Corresponding author: reimer@mit.edu

Summary: Voice interface use has become increasingly popular in vehicles. It is important that these systems divert drivers' attention from the primary driving task as little as possible, and numerous efforts have been devoted to categorizing demands associated with these systems. Nonetheless, there is still much to be learned about how various implementation characteristics impact attention. This study presents a secondary analysis of the delay time between when users finish giving commands and when the system responds. It considers data collected on 4 different production vehicle voice interfaces and a mounted smartphone in field driving. Collapsing across systems, drivers showed an initial increase in heart rate, skin conductance level, and off-road glance time while waiting for a system to respond; a gradual decrease followed as delays continued. The observed attentional and arousal changes are likely due to an increase in anticipation following a speech command, followed by a general disengagement from the interface as delay times increase. Safety concerns associated with extended delay times and suggestion of an optimal range for system response times are highlighted.

INTRODUCTION

The frequent use of voice interface systems in automobiles has raised concerns about driver safety. These concerns are often reasonably founded and frequently met with reassurances that voice systems are far safer than manual input methods. However, a clearer understanding of the demands associated with voice interaction in the vehicle is needed to scientifically support the comprehensive assessment of all types of demands (visual, cognitive, etc.) associated with interactions. Research examining the relationship between voice interfaces and driver distraction use measures such as stimulus response time, cognitive load, lane position, and eye gaze to assess demands placed upon the driver (e.g. Reimer et al., 2014; Strayer et al., 2014). Other studies focus more on voice recognition accuracy, task time, and user preference (e.g., Walker, Kamm, & Litman, 2000; Hajdinjak & Mihelic 2006). These evaluation methods often face criticism for neglecting the demand that various systems place upon the user (Hajdinjak & Mihelic 2006; Lo & Green, 2013). Holistic assessments of the demands of voice systems need to consider all potential factors to support the development of highly optimal interfaces that minimize the impact on attention and driving behavior (Reimer et al., 2014).

One important, yet often ignored, aspect of voice interface system optimization is the delay time between when a command is spoken and when the system responds. Natural human interactions rely on turn taking, where two operators fluidly alternate speaking (Duncan, 1972; Thomaz & Chao, 2011). Voice interface systems attempt to recreate this turn taking style of communication with the goal of providing a natural feeling conversation. However, in practice, this is quite computationally difficult.

Syntax parsing and response location detection (Meena, Skantze, & Gustafson, 2014) methods are utilized to determine when a user has completed an utterance and when a voice system should begin processing and subsequently respond to the user. One simple way to do this is for the interface to wait for a set amount of time after the user stops speaking. However, such a method often results in long delays that lack a natural feel. More advanced response location detection methods consider changes in voice pitch, speaking duration, and various lexico-syntactic factors (Meena, Skantze, & Gustafson, 2014) to more optimally assess the end of an utterance.

Despite contemporary response location detection methods, many voice interface systems are still slow to respond. Long delay times may also be due to hardware constraints or voice recognition issues resulting from environmental noise (Hataoka et al., 2008). Regardless, self-report data show that people find long delays uncomfortable and unnatural (Skantze & Hjalmarsson, 2012; Meena, Skantze, & Gustafson 2014).

Extended delays between turns in conversations can have a variety of negative effects on users. During long wait times, users must maintain task relevant information in working memory. Under high working memory load, fewer resources are available for other processing priorities, interfering with the allocation of attention (Lavie, 1995). Ross et al. (2013) demonstrate that high working memory load has a negative impact on driving performance. Another issue raised by long delays between turns is that memory will decay over time, and this can eventually lead to difficulty completing tasks correctly (Brown 1958). Users may also actively inhibit responses to the upcoming feedback from the voice system during this wait time. For example, after cuing the navigation system the user must wait for the voice interface to ask for the address. The user knows this step is coming, but still has to wait and inhibit the urge to say the address before being prompted. Cognitive control processes may be involved during this wait time, and extended wait times may be an added source of stress.

In human conversations, long gaps between turns suggest confusion or a misunderstanding (Maat, Truong, & Heylen, 2011). Users may misattribute similar perceptions to voice interfaces, leading to further confusion and difficulty completing tasks. At the same time, long delays between turns increase overall task time, therefore drawing the user's attention away from other tasks. Distracted driving studies emphasize the importance of minimizing events that pull attention off-road, and to limit periods of distraction as much as possible (Reimer et al., 2013).

No research was identified investigating how drivers in the field behave during system delays. This secondary analysis aims to begin to fill this gap by examining the effect of delay times between turns on driver attention and arousal during the operation of production level voice interface systems in on-road driving situations. From the literature, it is clear that shorter delays are desirable. However, it is unclear if there is such a thing as a maximum acceptable delay time in vehicle settings where drivers are sharing attention between the road and a secondary activity.

METHODS

This secondary analysis draws upon data from a set of studies involving four different instrumented vehicles (2010 Lincoln MKS, 2013 Chevrolet Equinox, 2013 Volvo XC60, and 2014 Chevrolet Impala). In addition, to interacting with the embedded vehicle voice systems, drivers in Equinox and XC60 also utilized a mounted Samsung Galaxy S4 smartphone.

Participants

This study draws on data from 120 participants. From each dataset, 24 unique participants were randomly drawn. In the case of the smartphone, the 24 subject sample was comprised of 12 subjects drawn from both the Equinox and XC60 datasets that were distinct from the sample drawn for the vehicle interface assessment. The demographic characteristics of participants were such that each group of 24 was gender balanced and meet the recommended age groups for the NHTSA Phase I distraction guidelines. The studies were all approved by the local ethics board.

Apparatus and Procedure

All vehicles were instrumented for time-synchronized data collection from embedded sensors, including the vehicle's controller area network, a MEDAC System/3 physiological monitoring unit, and video and audio recorded from the vehicle's cab. Sensor data were logged at 10 Hz except for physiological signals, which were recorded at 250 Hz. Details on the processing of physiological data and double coded and mediated manual reduction of eye data can be found in Reimer et al., 2013.

The experiments were all conducted in a similar manner with participants completing a number of HMI activities while driving under highway conditions. Participants were extensively trained on HMI activities prior to driving and, where available (MKS and XC60), performed the interface's voice calibration procedure. Further details regarding the individual studies can be found in (Reimer et al., 2013; Mehler et al., 2014; Reimer et al., 2015).

This analysis draws from two full destination address voice navigation entry tasks performed in the MKS, three in the Equinox, and four in the Impala. In addition, participants in the Equinox and XC60 performed four voice based contact dialing tasks. In the Equinox and XC60, the number of tasks completed on the smartphone was equivalent to those completed using the embedded vehicle system. The MKS and XC60 embedded voice systems utilized a stepwise menu based entry of information, while the other systems used a "one-shot" approach. This analysis only considered the first turn with each voice interface. Delay times were measured from when a participant finished speaking the first step of a voice command, for example "Destination Street Address," to when the system responded to the command. During this delay participants waited for the system to ask for a confirmation of the previous input or to move to the next step. Plots present the delay per task per subject with a LOESS regression line. Due to high variability of delay times within each subject, we chose to avoid collapsing data per subject when possible. Statistical comparisons collapse data such that one mean delay is associated with each participant (i.e. removing the within subject statistical comparison). Where noted, two minute just driving baselines drawn prior to each task were used for normalization.

RESULTS

Figure 1 depicts the mean and standard error of the system delay times for each system as well as the individual delay times at the task level. There was a significant main effect of voice interface type (*F*(4,103)=137.90, *p*=<.001).

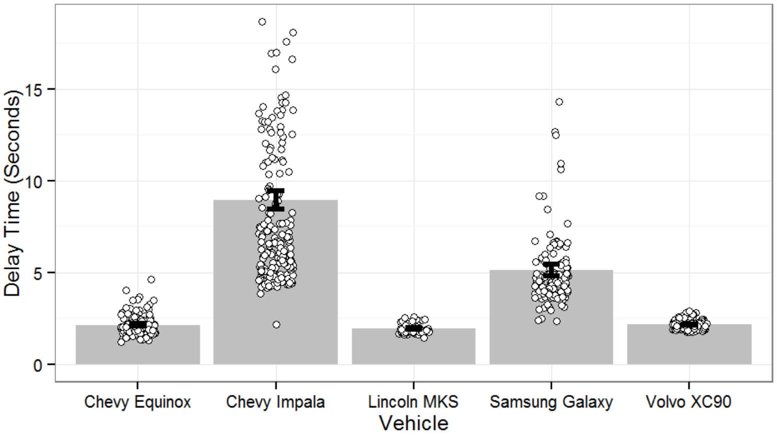

Figure 1. Mean delay by voice interface, dots represent individual data points and error bars represent SEM

Off-Road Glances

There were significant positive correlations between total off-road glance time and delay length (*r*(118)=.52, *p*<.01) and between off-road glance frequency and delay length (*r*(118)=.76, *p*<.01). In the percent of a delay time where drivers are looking off road, we see a similar trend as in HR and SCL. The shortest delay times are associated with very low or very high percentages of off-road glances since there is less time to change one's glance location during the delay. In the 66 percent of trials where delay time was below 2 seconds, participants did not change their glance location at all. There was a small peak around 3 seconds, followed by a gradual decrease and leveling out in percent of off-road glances as delay times increased (Figure 4).

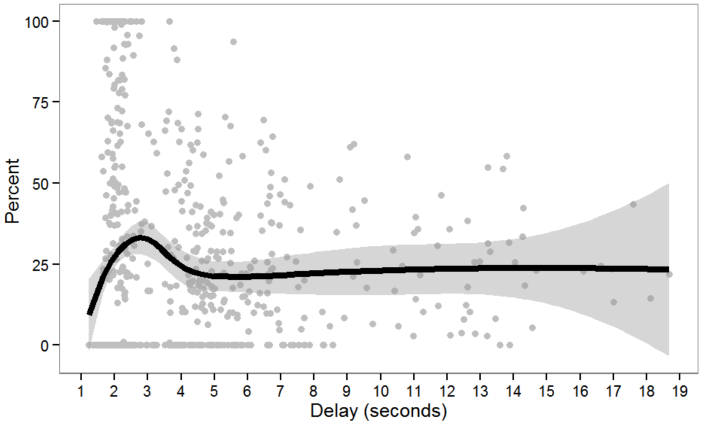

Figure 2. Delay time by percent of time looking off-road (gray area represents SEM for each delay time)

Skin Conductance Level (SCL)

During the delay period, SCL was significantly elevated compared to the baseline (*M*=13.88, *SE*=1.10) and (*M*=12.32, *SE*=1.05). SCL changes from baseline driving are relatively widely distributed for short delays and the initial increase within the 1.5 to 3 second range is characteristic of the time course of a normal anticipatory SCL response profiles. The elevation is sustained and then trends slightly higher as the delay approaches 3-5 seconds, then tapers off as delays become longer (Figure 2).

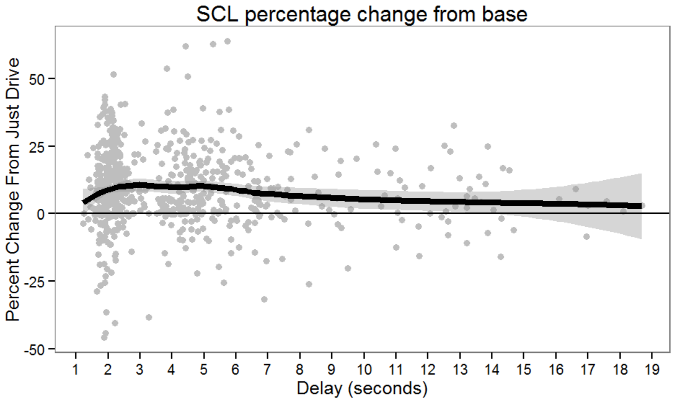

Figure 3. Delay time by mean SCL change from base (gray area represents SEM for each delay time)

Heart Rate (HR)

Mean HR was significantly elevated during delay times compared to the baseline, (*M*=76.13, *SE*=1.44) and (*M*=74.66, *SE*=1.25). As with SCL, HR changes peaked around 4-5 seconds; they then tapered at long delays, and gradually increased again during extreme delays (Figure 3).

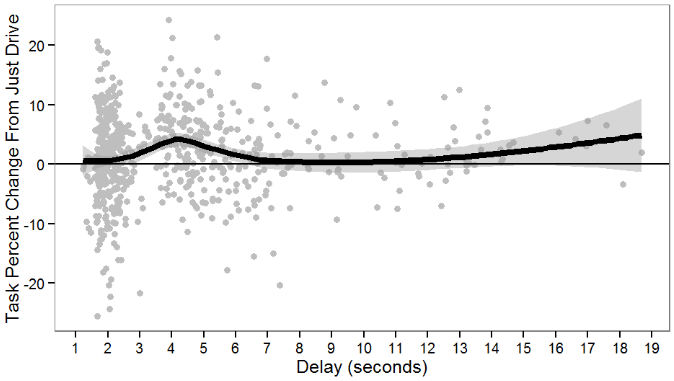

Figure 4. Delay time by mean HR change from base (gray area represents SEM for each delay time)

DISCUSSION

These data clearly show that some voice interface systems have much longer delay times than others. Combining data from multiple systems allows us to begin developing observations of effects over a range of system delays. Results suggest that there are complex relationships between delay time, glance behaviors, and changes in HR and SCL indicating that there may be variations in the level of visual and cognitive engagement across the waiting period. The results demonstrate that voice interfaces can engender visual and cognitive engagement during periods where the driver is not overtly required to interact with the system. Since longer delay times contribute to overall task time, it can also be deduced that longer delay times will lead to longer periods of driving where attentional resources are diverted from the primary task of driving. The impact of such temporal variation in task engagement on overall workload is not clear.

The common trend across HR, SCL, and off-road glances suggests that a driver's distribution of attention is impacted by the temporal nature of delay times. Total off-road glance time and glance frequency is positively correlated with delay time. Elevations in HR and SCL tend to peak at around 3 to 4 seconds, which may indicate that participants are experiencing a build-up of anticipatory arousal waiting for a response. Off-road glances suggest that the system is frequently checked for visual feedback during this time. The gradual decrease in all three measures suggests that attention for some drivers is pulled away from the voice system and reallocated to driving after around 4 seconds of waiting for a response. This may be due to users giving-up on waiting for feedback and shifting attention back to the primary task of driving. The latent increase in HR with extremely long delay times may stem from an increase in stress experienced by users who are unsure if the system is working correctly and growing frustration.

CONCLUSION

Long delay times between turns are not only unpleasant, but may even be taxing to drivers. At any length, delays between turns in user interfaces demand resources from users, as shown through three different measurements, creating suboptimal driving conditions. Delay times longer than 4 seconds are associated with indicators of decreased attention to the task, likely due to disinterest and a reallocation of control processes to bring attention back to the road. This suggests the possibility that system delays under 4 seconds may optimal.

LIMITATIONS

We chose to plot raw trial data without aggregating per subject and computed statistical comparisons at the aggregate level. We felt that the per subject view provided statistical integrity and an accessible characterization of effects without unduly compromising visualization through a complex statistical analysis. While this study is unable to make firm conclusions regarding delay times due to possible confounds of other features of the different voice systems, it does accomplish the goal of making an initial attempt to understand what occurs during these delay times. Future work may wish to develop a more sophisticated model of effects while controlling for per subject variability.

ACKNOWLEDGEMENTS

Support for this analysis was provided by the US DOT’s Region I New England University Transportation Center at MIT and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs’ class counsel. Data were drawn from studies supported by the Toyota Collaborative Safety Research Center (CSRC), the Santos Family Foundation, and the Insurance Institute for Highway Safety (IIHS).

REFERENCES

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292.

Hajdinjak, M., & Mihelič, F. (2006). The PARADISE evaluation framework: issues and findings. *Computational Linguistics*, *32*(2), 263–272.

Hataoka, N., Manabu Araki, Takashi Matsuda, Masayuki Takahashi, Ryoichi Ohtaki, & Obuchi,

Y. (2008). Evaluation of interface and in-car speech - many undesirable utterances and sever noisy speech on car navigation application. (pp. 956–959). IEEE.

He, J., Chaparro, A., Nguyen, B., Burge, R. J., Crandall, J., Chaparro, B., … Cao, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? *Accident Analysis & Prevention*, *72*, 287–295.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468.

Lo, V. E.-W., & Green, P. A. (2013). Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature. *International Journal of Vehicular Technology*, *2013*, 1–13.

Maat, M., Truong, K., & Heylen, D. (2011). How agents' turn-taking strategies influence impressions and response behaviors. *Presence: Teleoperators and Virtual Environments*, *20*(5), 412–430.

Meena, R., Skantze, G., & Gustafson, J. (2014). Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language*, *28*(4), 903–922.

Mehler, B., Reimer, B., Dobres, J., McAnulty, Mehler, A., Munger, D. & Coughlin, J.F. (2014). *Further evaluation of the effects of a production level "voice-command" interface on driver behavior*. Report 2014-2. MIT AgeLab, Cambridge, MA.

Reimer, B., Mehler, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Rumpold, A. (2014). Effects of an 'expert mode' voice command system on task performance, glance behavior & driver physiology. Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicle Applications, Seattle, WA.

Reimer, B., Mehler, B., Dobres, J. & Coughlin, J.F. (2013). *The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance*. Report 2013-17A. MIT AgeLab, Cambridge, MA.

Reimer, B., Mehler, B., Reagan, I, Kidd, D., & Dobres, J. (2015). *Multi-modal demands of a smartphone used to place calls and enter addresses during highway driving relative to two embedded systems.* Arlington, VA: Insurance Institute for Highway Safety.

Ross, V., Jongen, E. M. M., Wang, W., Brijs, T., Brijs, K., Ruiter, R. A. C., & Wets, G. (2014). Investigating the influence of working memory capacity when driving behavior is combined with cognitive load. *Accident Analysis & Prevention*, *62*, 377–387.

Skantze, G., & Hjalmarsson, A. (2013). Towards incremental speech generation in conversational systems. *Computer Speech & Language*, *27*(1), 243–262.

Strayer, D. L., Turrill, J., Coleman, J. R., Ortiz, E. V, & Cooper, J. M. (2014). *Measuring cognitive distraction in the automobile II: assessing in-vehicle voice-based interactive technologies*. AAA-FTS, Washington, D.C

Thomaz, A., & Chao, C. (2011). Turn taking based on information flow for fluent human-robot interaction. *Association for the Advancement of Artificial Intelligence*, *Winter 2011*, 53–63.

Walker, M., Kamm, C., & Litman, D. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering*, *6*(3&4), 363–377.