Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems

Bruce Mehler, David Kidd, Bryan Reimer, Ian Reagan, Jonathan Dobres & Anne McCartt

To cite this article: Bruce Mehler, David Kidd, Bryan Reimer, Ian Reagan, Jonathan Dobres & Anne McCartt (2016) Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems, Ergonomics, 59:3, 344-367, DOI: 10.1080/00140139.2015.1081412

To link to this article: https://doi.org/10.1080/00140139.2015.1081412

Published online: 12 Oct 2015.

ERGONOMICS, 2016
VOL. 59, NO. 3, 344–367
http://dx.doi.org/10.1080/00140139.2015.1081412

Taylor & Francis
Taylor & Francis Group

Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems

Bruce Mehler[a], David Kidd[b], Bryan Reimer[a], Ian Reagan[b], Jonathan Dobres[a] and Anne McCartt[b]

[a]MIT AgeLab, New England University Transportation Center, Cambridge, MA, USA; [b]Insurance Institute for Highway Safety, Arlington, VA, USA

ABSTRACT

One purpose of integrating voice interfaces into embedded vehicle systems is to reduce drivers' visual and manual distractions with 'infotainment' technologies. However, there is scant research on actual benefits in production vehicles or how different interface designs affect attentional demands. Driving performance, visual engagement, and indices of workload (heart rate, skin conductance, subjective ratings) were assessed in 80 drivers randomly assigned to drive a 2013 Chevrolet Equinox or Volvo XC60. The Chevrolet MyLink system allowed completing tasks with one voice command, while the Volvo Sensus required multiple commands to navigate the menu structure. When calling a phone contact, both voice systems reduced visual demand relative to the visual–manual interfaces, with reductions for drivers in the Equinox being greater. The Equinox 'one-shot' voice command showed advantages during contact calling but had significantly higher error rates than Sensus during destination address entry. For both secondary tasks, neither voice interface entirely eliminated visual demand.

Practitioner Summary: The findings reinforce the observation that most, if not all, automotive auditory–vocal interfaces are multi-modal interfaces in which the full range of potential demands (auditory, vocal, visual, manipulative, cognitive, tactile, etc.) need to be considered in developing optimal implementations and evaluating drivers' interaction with the systems.

Social Media: In-vehicle voice-interfaces can reduce visual demand but do not eliminate it and all types of demand need to be taken into account in a comprehensive evaluation.

ARTICLE HISTORY
Received 20 December 2014
Accepted 17 June 2015

KEYWORDS
Voice interface; visual demand; distraction; workload; human–machine interface

1. Introduction

Manufacturers are equipping vehicles with embedded systems that allow occupants to interact with entertainment, communication and driver support systems built into the vehicle (e.g. radio, navigation) and connected portable devices (e.g. cell phones, MP3 players). Increasingly, embedded systems allow occupants to interact with different functions or devices with voice commands in addition to traditional visual–manual interactions using buttons, knobs or a touch screen. A perceived advantage of voice inputs compared with manual inputs is that they eliminate or reduce the competition for visual and manual resources between a secondary activity and the primary task of driving. Therefore, voice interfaces have been widely considered as an appealing approach for giving drivers access to a range of entertainment and connectivity options while minimising the potential impact on driving performance and safety. At the same time, there remains a concern that performing any secondary task can increase crash risk, and some caution that adding even easy-to-use interfaces may raise the total amount of attention drivers give to non-driving tasks. Regardless, a deeper understanding of the various demands originating from drivers' interactions with voice interfaces is needed to more objectively optimise tasks in which drivers engage and to identify tasks that can lead to an inappropriate level of disruption in driving.

The apparent benefits of using voice inputs to interact with a device while driving compared with manual inputs are well documented in experimental research using various simulated driving performance measures. For instance, the standard deviations of lane position and reaction time are not as degraded relative to baseline driving, and may be less than baseline driving, when drivers perform a secondary auditory–vocal task vs. a secondary visual–manual task (e.g. Haigney, Taylor, and Westerman 2000; Maciej and Vollrath 2009; Ranney, Harbluk, and Noy 2005; Tsimhoni, Smith, and Green 2004). Schreiner, Blanco, and Hankey (2004) have shown parallel findings with a simulated voice system on a closed roadway. Drivers also look away from the roadway less often and for less time during voice interactions (Chiang, Brooks, and Weir 2005; Mehler et al. 2014; Owens, McLaughlin, and Sudweeks, 2010; Reimer et al.

CONTACT Bruce Mehler bmehler@mit.edu

The Appendix was originally published with errors. This version has been amended. See Corrigendum (http://dx.doi.org/10.1080/00140139.2016.1179501).

2013; Reimer et al. 2014; Schreiner, Blanco, and Hankey, 2004; Shutko et al. 2009). Attempts to use voice interactions as a means of reducing the amount of time that a driver's eyes are directed away from the road are easily understood. In studies of small samples of drivers who were continuously monitored over an extended period of time, the risk of a crash, near-crash, or other type of safety conflict increased the longer the driver's eyes were off the road (e.g. Klauer et al. 2006). Thus, systems that allow drivers to look away from the roadway less often may degrade safety less. However, experimental research also indicates that some voice interactions affect driving performance in ways that could increase crash risk. Speech generation, speech comprehension and even simple cognitive tasks can affect simulated driving performance (e.g. speed variability, lane maintenance) and mental workload, especially when the information is complex or poorly implemented within the vehicle (e.g. Blanco et al. 2006; Kubose et al. 2006; Lee et al. 2001; Strayer et al. 2013).

1.1. On-road research with production voice systems compared with visual–manual interaction

Although the potential benefits of voice-based interactions compared with visual–manual interactions are well documented in simulated or prototype implementations (see previous section and reviews by Barón and Green (2006), Lo and Green (2013), and Reimer et al. (2013)), only a few studies have examined whether these benefits exist with production-level embedded systems (e.g. Graham and Carter 2000; Harbluk et al. 2007; Owens, McLaughlin, and Sudweeks 2011; Shutko et al. 2009), and even fewer have examined the use of production systems on actual roadways. Chiang, Brooks, and Weir (2005) conducted two studies where drivers entered a destination into a navigation system using a point-of-interest selection, the destination's phone number or the street address with similar built-in navigation systems in a 2004 Accord and 2005 Acura RL. Participants completed the destination entry tasks using voice or manual inputs while driving on city streets and a freeway. Drivers spent a smaller percentage of time looking at the navigation system interface, had less variability in lane keeping performance, and reported lower subjective ratings of mental workload using voice inputs compared with manual inputs. In another on-road study, Owens, McLaughlin, and Sudweeks (2010) reported that using the embedded voice interface of the Ford SYNC® system to complete several infotainment tasks lowered drivers' visual demand, steering variability and subjective mental workload relative to using a portable hand-held cell phone. Owens, McLaughlin, and Sudweeks (2010) did not examine performance during manual interactions with the vehicle's embedded system or voice interactions with the cell phone.

In a series of studies, Reimer, Mehler, and colleagues (Mehler et al. 2014; Reimer and Mehler 2013; Reimer et al. 2014) also examined Ford SYNC® and an embedded navigation system in a 2010 Lincoln MKS. While driving on an interstate highway, drivers used voice commands in a set of navigation, phone calling and entertainment tasks. As comparison points, drivers also manually changed radio stations using preset buttons, engaged in a more intensive visual–manual reference task of locating specific radio stations that required multiple button presses and rotation of the tuning knob, and completed several levels of a working memory cognitive reference task. In line with the studies cited above that showed some advantages for voice command systems, drivers looked away from the roadway less when using the voice interface to select a radio station than when using the multi-step manual radio tuning interface. On the other hand, compared with multi-step manual radio tuning, drivers looked away from the roadway for substantially longer periods of time when using voice inputs to enter a destination address into the navigation system or attempted to select a song that did not exist in the entertainment system. Reimer and Mehler (2013) noted that the user interface for voice input presented information on the centre console display (e.g. voice command options, street or city name selection options), which inherently provides a reason to look at the screen.

Together, the findings from these on-road studies suggest that, compared with manual interaction, voice interaction with embedded and portable systems can reduce visual demand as intended, but do not necessarily eliminate it. Some voice interactions appear to result in moderate to large visual engagement when considered in terms of metrics such as glance time to device or total eyes-off-road time that have been used or proposed for evaluating visual–manual interfaces (Driver Focus-Telematics Working Group 2006; NHTSA 2013).

1.2. Variation across system implementations

However, the designs of embedded systems vary, and some interface designs may be more effective at minimising visual demand than others. Reagan and Kidd (2013) used hierarchical task analysis to count the steps required to dial a 10-digit phone number, dial a contact in a cell phone contact book and tune to a radio station to identify differences in the manual and voice interfaces of four embedded systems in 2013 model year vehicles. Two distinct design approaches for voice interactions emerged from this static evaluation. The first was a menu-based approach where tasks were completed using contextual voice commands to progress through a series of menus and submenus, often mimicking the sequence of manual inputs required to complete the task. The second was a 'one-shot' approach where a single compound command was used to execute most or all of the tasks in a single step.

As initially observed in Reagan and Kidd (2013), the differences in these two approaches were most apparent between the Chevrolet MyLink and Volvo Sensus systems. For example, calling a contact in the phone book could be performed using a single voice command with the Chevrolet MyLink (e.g.

the driver saying 'Call home', which resulted in the system response, 'calling home on cell' and initiation of the phone call). The same task required four separate voice commands with the Volvo Sensus as the user moved through different menus and verified previous commands (e.g. to call 'home' a driver said 'Phone, call contact'; waited for the system prompt 'name please'; said the contact name 'home'; waited for the system prompt, 'please say a line number'; said 'one'; waited for the system prompt 'dial home mobile – confirm: yes or no'; and then said 'yes', after which Sensus made the call). Calling a contact with the Volvo Sensus took more steps when using voice inputs than when using manual inputs. Furthermore, many of the system prompts asked the driver to look at the centre stack display to choose among options for the contact (e.g. home, cell, work) or to confirm input, and the prompt asking for a line number always occurred whether the contact ('home' in the example here) had one or multiple line numbers. This integration of visual information to support the voice interface was similar to that noted by Reimer and Mehler (2013) in their initial evaluation of the Ford SYNC® system. Providing visual information to support voice input may help alleviate the cognitive demand associated with memorising and recalling voice commands, or an auditory list of options. However, it is unclear whether and to what extent safety or other trade-offs are associated with reducing cognitive demand at the expense of increased visual demand.

While calling a contact in the phone book with Chevrolet MyLink required fewer voice commands than Volvo Sensus, MyLink required a deeper understanding of system operation in that it did not provide as much prompting, visual support or confirmation, which could potentially result in more calling errors. The cognitive demand associated with recalling complex voice commands might negate the benefits associated with reducing overall task duration and the potential for visual engagement. Previous research has shown that cognitive demand from a secondary task can interfere with visual information processing (Just, Keller, and Cynkar 2008; Strayer, Drews, and Johnston 2003) and constrict visual scanning patterns (Recarte and Nunes 2000; Reimer et al. 2012).

A recent study by Garay-Vega et al. (2010) found that the differences between a menu-based voice interface and one-shot voice interface are not negligible. Drivers completed a music retrieval task using an iPod™, a multiple turn voice interface (i.e. menu-based voice interface), and a single turn voice interface (i.e. one-shot voice interface) during simulated driving. The task took longer to complete using the multiple turn voice interface. Furthermore, only the single turn voice interface reduced the average time that drivers had their eyes off the road compared with the iPod™. The multiple turn voice interface was also perceived to be more demanding than the single turn interface.

In sum, naturalistic driving research indicates increased risk of safety-critical events from the visual and manual demands of in-vehicle secondary tasks (Fitch et al. 2013; Klauer et al. 2006; Victor et al. 2014). Research also indicates that voice interfaces reduce workload and visual attentional demand relative to visual–manual interfaces (e.g. Chiang, Brooks, and Weir 2005; Haigney, Taylor, and Westerman 2000; Mehler et al. 2014; Owens, McLaughlin, and Sudweeks 2010; Reimer et al. 2013; Shutko et al. 2009). However, recent research indicates that voice-based interactions may introduce noticeable visual demand (e.g. Mehler et al. 2014; Reimer et al. 2013) and that some voiced interface designs can increase perceived workload and visual demand when driving in a simulator relative to others (e.g. Garay-Vega et al. 2010). These findings support concerns that distraction can remain (depending upon implementation), despite the use of voice-based systems and led to the task analysis by Reagan and Kidd (2013), described above, and to the study reported here.

1.3. Objectives and approach

A primary objective of the current study was to compare the relative demands of production implementations of primarily visual–manual vs. voice-involved human–machine interfaces intended to allow completion of the same end-goal task while driving by considering the effects on driving performance, visual demand and indices of mental workload (heart rate, skin conductance and subjective ratings). Of equal interest was an exploration of the significance of differing design approaches to voice-based systems (e.g. a one-shot vs. multi-step entry). A 2013 Chevrolet Equinox equipped with the MyLink system and a 2013 Volvo XC60 equipped with Sensus served as test case exemplars of these two system designs in the research reported here.

Volunteer drivers drove either the Chevrolet or Volvo on a highway while initiating calls through a phone contact list using voice and manual inputs and entering addresses into the navigation system using voice input with the vehicle's embedded system and a mounted smartphone. In the case of phone calling, using voice inputs of the embedded systems was expected to degrade driving performance less, reduce visual demand and lower workload levels compared with performing these tasks manually. Based on the task analysis by Reagan and Kidd (2013), the relative benefits of using voice input compared with manual input were expected to be greater for drivers using the Chevrolet MyLink. However, the absence of verification steps with MyLink was expected to increase the number of errors using voice inputs for complex tasks such as address entry.

2. Methods

2.1. Participants

Participants were identified primarily using online and newspaper advertisements in the greater Boston area. Recruitment was directed at obtaining a sample of relatively healthy and

Table 1. Mean age (and SD) of participants by age group, gender and vehicle.

	Chevrolet (n = 40)		Volvo (n = 40)		Combined
Age group	Female (n = 20)	Male (n = 20)	Female (n = 20)	Male (n = 20)	n = 80
20–24 (n = 20)	21.4 (0.9)	22.4 (1.8)	23.2 (0.8)	21.2 (1.3)	22.1 (1.4)
25–39 (n = 20)	33.0 (3.4)	31.2 (4.9)	28.8 (3.2)	28.9 (4.0)	30.5 (4.3)
40–54 (n = 20)	45.4 (4.0)	47.6 (3.9)	48.6 (5.0)	49.8 (3.7)	47.9 (4.2)
55–69 (n = 20)	62.4 (2.7)	59.0 (2.6)	59.8 (4.0)	62.4 (4.3)	60.9 (3.6)
Combined	40.6 (15.8)	40.1 (14.9)	40.2 (16.0)	40.6 (17.1)	40.3 (15.6)

experienced drivers. Participants were required to be between the ages of 20 and 69, have been licensed for a minimum of 3 years, and self-report driving at least 3 times a week and being in relatively good health for their age. Also based on self-report, individuals were excluded if they were a driver in a police-reported crash in the past year, were positive for any of a range of serious medical conditions (e.g. a major illness resulting in hospitalisation in the past 6 months, a diagnosis of Parkinson's disease, a history of stroke) or were taking medications that might impair their ability to drive safely under the study conditions (e.g. anti-convulsants, anti-psychotics, medications causing drowsiness).

Recruitment continued until a sample of 80 participants with usable driving performance, glance and physiological data, and equally balanced across the two vehicles by gender and age was obtained. The target age distribution was in general conformity with the recommendations for device assessment in NHTSA's (2013) *Visual–Manual Driver Distraction Guidelines for In-Vehicle Electronic Devices*, which call for an equal number of participants across four age groups (18–24, 25–39, 40–54, 55 and older). In line with recruitment goals, participant age did not vary significantly by gender or vehicle (both $F(1,79) = 0.949$) (see Table 1); the sample ranged in age from 20 to 66 years. Recruitment procedures and the overall experimental protocol were approved by MIT's institutional review board, and compensation of $75 was provided.

2.2. Apparatus

Participants drove one of two standard production vehicles, a 2013 Chevrolet Equinox equipped with the MyLink system and a 2013 Volvo XC60 equipped with the Sensus system. Both vehicles were equipped with forward collision warning and lane departure warning safety systems. Phone connectivity was supported by pairing a Samsung Galaxy S4 smartphone (model SCH-1545) to each vehicle's embedded system via the vehicle's Bluetooth wireless interface. Both vehicles were instrumented with a custom data acquisition system for time synchronised recording of vehicle information from the controller area network (CAN) bus, a Garmin 18X Global Positioning System (GPS) unit, a MEDAC System/3™ physiological monitoring unit to provide EKG and skin conductance level (SCL) signals, video cameras and a wide-area microphone to capture driver speech and audio from the vehicle's speech system. The five video cameras provided views intended to capture the driver's face for primary glance behaviour analysis, the driver's interactions with the vehicle's steering wheel and centre console, the forward roadway (narrow and wide-angle images) and a rear roadway view. Data were captured at 10 Hz for the CAN bus and GPS, 30 Hz for the face and narrow forward roadway cameras, 15 Hz for the remaining cameras and 250 Hz for the physiological signals to support EKG feature extraction for heartbeat interval detection.

For EKG recordings, the skin was cleaned with isopropyl alcohol and standard pre-gelled silver/silver chloride disposable electrodes (Vermed A10005, 7% chloride wet gel) applied using a modified lead II configuration that placed the negative lead just under the right clavicle, the ground just under the left clavicle and the positive lead over the lowest left rib. Skin conductance was measured utilising a constant current configuration and non-polarising, low impedance gold-plated electrodes that allowed electrodermal recording without the use of conductive gel. Sensors were attached with medical grade paper tape on the underside of the outer segments of the middle fingers of the left hand to leave the right hand free for engaging the push-to-talk button on the steering wheel of each vehicle and controls on the instrument cluster. The thin surface design of the electrodermal sensors minimised interference with a natural grip of the steering wheel associated with the use of more traditional cup style electrodes.

2.3. Secondary tasks

2.3.1. Calling a phone contact

A phone list of 108 contacts was used for all phone calling tasks. Characteristics of how each system organised information were taken into consideration so that neither system was disadvantaged. The list was ordered by first name and entries started with A and ranged through R so that all target selections could be reached through a comparable number of manual actions in each system. There were 18 names in each of 6 alphanumeric ranges corresponding to the bin organisation used in the Chevrolet MyLink manual interface (ABC, DEF, GHI, JKL, MNO, PQRS, TUV and WXYZ).

Calling a phone contact was presented as a sequence of two 'easy' and two 'hard' tasks. The easy tasks were calling a contact with only one phone number entry (Mary Sanders and Carol Harris). The hard tasks were calling a contact with two phone numbers (e.g. home and mobile). For these contacts (Pat Griffin on mobile and Frank Scott at work), the target phone was never the first listing so that simply requesting the contact name alone would not dial the correct number. The form of the easy task prompt was 'Your task is to call Mary Sanders. Begin'. The form of the hard task prompt was 'Your task is to call Frank Scott at work. Begin'. The contacts were the same across the manual and voice interface interactions so that any aspect/characteristic of a particular contact name

that might influence the relative difficulty was constant (e.g. alphabetic location). As previously noted, a detailed description of the operations and resources required to dial a contact using the MyLink and Sensus systems is provided in Reagan and Kidd (2013). The key elements of each approach as they relate to the tasks used in this study follow.

Calling a contact using the MyLink visual–manual interface had the most discrete steps and began by locating and selecting the phone subsystem, followed by selecting the alphanumeric bin (e.g. ABC, DEF) containing the target contact. For contacts with a single phone number (easy case), the contact name was then selected from the list. In the case of two numbers for a contact name (hard case), both numbers appeared sequentially in the same list (i.e. Frank Scott work, Frank Scott home) and the target option was selected. Calling a contact using the Sensus visual–manual interface required the user to select the phone subsystem and then scroll through the upper level of the contact list to the appropriate contact name using a rotary knob on the centre console. In the case of contacts with a single phone number (easy case), pressing an 'OK' button initiated the call. For contacts with multiple numbers (hard case), pressing the button brought up a submenu listing the phone numbers for that contact. The rotary dial was again used to locate the desired selection and the 'OK' button was pressed.

In contrast to manual calling, the MyLink voice interface required few steps. After pressing the push-to-talk button on the steering wheel, the driver could initiate both the easy and hard tasks in a single command string (i.e. 'call Mary Sanders', 'call Pat Griffin on mobile'). No confirmation step was required if the system had confidence in the identification of the selection. This kept the interaction brief but meant the driver had to interrupt call initiation if a recognition error occurred. The voice implementation in the Sensus system more closely mirrored the multi-level menu structure used in the manual interface and asked for confirmation of selections. In specific, after pressing the push-to-talk button on the steering wheel, the driver could issue the compound command 'Phone call contact' to access the phone list and then say 'Mary Sanders'. The entry list would then appear on the display screen, and the driver was asked to say a line number and then asked to confirm the selection. In the case of multiple phone numbers for the contact, a second level menu would appear showing the options. The driver selected from this listing verbally and then confirmed the selection.

Each phone number associated with a target contact connected with a voicemail recording that confirmed the contact identity and stated that the phone call could now be disconnected. If the target contact was not reached, the call connected to a voicemail indicating that the MIT AgeLab had been reached and the phone call could now be disconnected. This provided auditory confirmation to the participant and research associate as to whether the target contact had been correctly selected or not.

2.3.2. Entering an address into the navigation system

Participants were asked to enter three addresses into each navigation system: (1) 177 Massachusetts Avenue, Cambridge, Massachusetts; (2) 293 Beacon Street, Boston, Massachusetts; and (3) their home address. The form of the prompt was 'Your task is to enter the destination address: 177 Massachusetts Avenue, Cambridge, Massachusetts. Begin'. The first two addresses also were printed in large black text on a white card attached to the centre of the steering wheel to minimise any cognitive load of needing to memorise and hold the address in memory during the duration of the interaction with the navigation system. This card was in place throughout the drive so that participants were exposed to the addresses for a minimum of 40 min prior to being asked to enter them into the system.

Address entry with MyLink was initiated by pressing the 'push-to-talk' button and saying the command 'navigation'. After prompting the driver for a navigation command, the system was flexible in terms of command syntax, accepting variations including 'destination address', 'enter address' and simply 'address'. The full address could then be entered as a single verbal string in the form '177 Massachusetts Avenue, Cambridge, Massachusetts'. If the system was able to parse the string into component parts that was interpreted as a unique address at a high confidence level, there was no confirmation step and navigation instructions were initiated. If multiple potential targets were identified, they were presented auditorily and visually to be selected by verbalising an option number.

Address entry using Sensus supported the compound but specific command, 'navigate go to address' to select address entry. In contrast to the single string 'one shot' approach of MyLink, Sensus prompted the user for the component parts of the address in steps, i.e. city name, street name and street number were entered separately. Recovery from a user error or system misidentification at each step required little familiarity with the system as the prompt for each step offered the option of returning to a previous step, e.g. 'please say the house number in single digits or say correction'. If the street number was correctly identified, the driver was prompted to say 'finish'. An additional confirmation step prompted the driver to say 'enter destination' to proceed with initiating navigation. If the system identified multiple potential targets, a list of options was shown on the centre stack display screen and the system prompted the driver to 'say a line number or say not on list'.

2.3.3. Instructions on task prioritisation

Participants were instructed several times (in the written consent form, by recorded instructions and through direct prompting by the research associate in the vehicle) that priority should be given to safe driving. Recorded instructions presented just prior to starting the drive stated the following:

> When you reach I-495 and have had a few minutes of driving on that highway, short recorded prompts will tell you what task we would like you to consider trying. When you hear

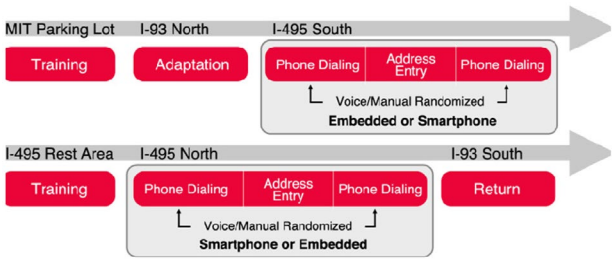

Figure 1. Schematic representation of the experimental design. Half of the participants interacted with the embedded vehicle system (Chevrolet MyLink or Volvo Sensus) during the first half of the drive and half during the second.

these prompts, please do not start a task until you hear the word 'begin'. While we would like you to consider doing each task, you should always give priority to safe driving. If you feel for any reason that a task will interfere with your ability to drive safely, delay starting the task until you feel it is safe to do so or skip the task entirely if you feel that is the best thing to do. Your safety, and the safety of other people around you, is the highest priority. Please also be aware that you will be responsible for paying for any citations that you might be issued for traffic regulation violations. If at any time you feel uncomfortable driving the vehicle or in your ability to drive safely, please let the research associate know how you are feeling and they will confer with you about pulling off the roadway at the nearest safe location.

2.4. Experimental design

Gender- and age-balanced samples were distributed across the Chevrolet and Volvo vehicles (see Table 1). As represented schematically in Figure 1 and further detailed in Section 2.6 on procedure, participants were presented with the phone calling tasks to be undertaken using a voice-based interface and a visual–manual interface and presented with addresses to enter into a voice-based navigation interface. While the present report focuses on the use of embedded vehicle systems to engage in these tasks, participants were also presented with the same tasks to be accomplished using a smartphone. Within each vehicle group, random assignment was made to either an 'embedded vehicle system' or a 'smartphone' first condition. Within each condition, random assignment determined whether voice-based or manual phone calling was presented first. Consequently, any advantage of being presented with the same contact to dial a second time was balanced across the modalities. The address entry tasks were always presented between the two methods of phone calling.

In sum, across 6 distinct task periods, each participant was presented a total of 22 secondary tasks, 11 during the southbound segment using either the embedded or smartphone device (4 manual phone calling trials, 3 address entry trials and 4 voice calling trials) and then repeated the same 11 tasks during the northbound segment using the alternate device. As noted above, this study focuses on the participants' interactions with the embedded systems. Findings considering interactions with the Smartphone are presented in a companion report (Reimer et al. 2015); however, a description of when participants were trained and assessed on both the smartphone and embedded systems is included in the following summary of the study procedures to provide a complete overview of the larger study.

2.5. Procedure

Participants reviewed and signed an informed consent, and a structured face-to-face interview confirmed eligibility that was initially established via phone or online screening. A questionnaire covering demographic information, attitudes towards driving, and technology experience was completed, an explanation of the workload rating scale provided, and physiological sensors were attached.

Participants were escorted to the research vehicle, instructed on how to adjust the seat and mirrors, asked to back up the vehicle and then return to the parking space to obtain an initial feel for the vehicle, and encouraged to make additional adjustments for comfort and visibility as desired. Participants were trained in the parking lot in the use of the system (embedded or smartphone) to which they were assigned for the first half of the drive. Training began with manual phone calling, followed by voice phone calling and then by voice destination address entry. Participants were trained to use the relevant rotary knobs on each vehicle's centre stack. The Chevrolet had a touch screen interface that could be used to complete selection tasks as an alternative to pressing the rotary knob; similarly, the Volvo had a thumb wheel on the steering wheel that could be used as an alternative method for scrolling through lists and making selections. Although they were not trained to use these alternative methods, participants were allowed to use the method if they discovered and preferred its use over the rotary centre stack knob. For the embedded vehicle systems, following the approach taken in Reimer et al. (2013) and Mehler et al. (2014), the default factory settings for the vehicle voice interfaces were used. Moreover, participants were given guidance on the use of shortcut command options to reduce the number of steps required to complete tasks. As an example, to use the voice interface in the Sensus system, calls could be placed by saying the upper level command 'Phone', waiting for a response and then saying 'Call Contact'. During training, participants were told 'Calls can be placed by speaking the command "Phone Call Contact"; you can also use the shorter command, "Call Contact"'. The remainder of the training interaction then focused on the shorter version.

As just described, training focused on providing participants with guidance on an efficient method of completing the tasks, while not constraining them to a fixed set of steps (beyond using the voice or manual interface at specified times) if the interface allowed multiple ways of accomplishing the task. Allowing participants to select an alternate method

(e.g. touching a selection on-screen rather than pressing the rotary control) if it felt more comfortable to them was seen as an approach that made task performance more naturalistic. This approach does diverge from an assessment that is specifically aimed at quantifying the demand associated with an exact sequence of steps as might be done in testing whether a specific method of completing a task meets the NHTSA (2013) voluntary guidelines.

Participants with the Volvo Sensus system were taken through the voice calibration procedure, which is intended to tune the voice recognition system to participants' pronunciation based on a set of command relevant words; the Chevrolet MyLink system did not offer this feature. Orientation and training consisted of recorded instructions to provide consistency, supplemented with guidance by a research associate to clarify details and answer questions as needed. Participants were encouraged to repeat tasks until they felt comfortable to proceed. The orientation/training period typically ranged between 15 and 30 min, with a mean of approximately 20 min.

Then participants drove the vehicle on actual roadways in and around the greater Boston area. A driving adaptation period of approximately 30 min took place prior to the start of the formal assessment. The route consisted of approximately 10 min of urban driving from MIT to interstate highway I-93 and approximately 20 min north on I-93 to I-495. For the portions used in this study, I-495 is a divided interstate that is largely surrounded by forest with three traffic lanes in each direction with lane widths of 15 feet (3.62 m). The posted speed limit is 65 mph (104.6 kph).

Presentation of the secondary tasks with the first assigned system interface (smartphone or embedded system) occurred while driving south on I-495 (see Figure 1). At the end of this southbound segment, a break was taken at a highway rest stop where participants completed workload and other ratings for the tasks just completed. They were then trained on the alternate interface (smartphone or embedded) for the same set of secondary tasks. Assessment of the alternate interface then took place during the second half of the drive as participants proceeded north on I-495. Most participants took approximately 35–40 min to drive each segment (north and south) (70–80 min combined).

The difficulty of the phone tasks was presented within each voice or manual period in the following order: easy, easy, hard, hard (see Section 2.3.1). This was intended to provide participants additional familiarity with the interface before assessing the harder task trials. Between individual trials, there was an interval of 30 s after the research associate recorded the completion of a task and the recorded instructions began for the next. A separation period of at least 3 min was provided following the end of one group of related tasks and the next period (e.g. between phone calling and address entry). Workload ratings for the second segment of the drive were completed along with a post-experimental questionnaire after completing the entire route. Total contact time for the study including intake and debrief was typically about 4 h.

2.6. Dependent measures

2.6.1. Self-reported workload

Subjective workload ratings were obtained using a single global rating per secondary task type on a scale consisting of 21 equally spaced dots oriented horizontally along a 10-cm line with the numbers 0 through 10 equally spaced below the dots and end points labelled 'Low' and 'High' on the left and right, respectively. The rating scales for all the secondary tasks were presented on one sheet, allowing participants to rate the tasks relative to each other. Participants were instructed to 'circle a point along each scale that best corresponds to how much workload you felt was involved in trying to do each task. Workload is best defined by the person doing the task and may involve *mental effort*, the amount of *attention required, physical effort, time pressure, distraction, or frustration* associated with trying to do the task while continuing to drive safely'. Previous work (Beckers et al. 2014; Dopart et al. 2013) using this approach produced ratings across user interface tasks that were consistent with relative rankings obtained concurrently using the NASA-Task Load Index (Hart 2006; Hart and Staveland 1988), one of the most widely used subjective workload assessment scales. In this regard, Hendy, Hamilton, and Landry (1993) provide a useful consideration of the sensitivity of simple univariate workload scales relative to multifactor scales when the goal is to obtain an overall workload rating.

2.6.2. Task completion time

The time it takes to complete a task has often been used as one measure of demand/usability (e.g. Hornbæk and Law 2007; Nielsen and Levy 1994). Task duration has been considered in the automotive literature, particularly within the context of navigation entry tasks in evaluation of driver-interface usability and safety (e.g. Green 1994, 1999a, 1999b). Existing industry guidance in the form of SAE standard J2364 (SAE 2004) recommends a maximum total task time of 15 s under static testing conditions for driver information systems that incorporate manual controls and visual displays; the guidelines explicitly state that this does not apply to voice-activated controls. The manner in which task duration is best interpreted in the context of voice-involved systems is largely an open question. In the current study, task completion time was calculated as the time between the end of a prompt to begin a task and completion of the task, which could involve successful completion (e.g. participant uttered the command to initiate a call to the specified contact), unsuccessful completion (e.g. participant uttered the command to initiate a call when it was not the specified contact) or failure at the point where the experimenter halted the trial (e.g. participant attempted to restart the full task for a third time, participant continued to

pronounce an entry the same way multiple times and voice recognition could not interpret or misinterpreted).

2.6.3. Physiological metrics

Physiological metrics have long been used as objective measures of task demand in fields such as aviation (Kramer 1991; Mulder 1992; Roscoe 1992; Veltman and Gaillard 1998; Wilson 2002) and have increasingly been employed in automotive-related research (Brookhuis and de Waard 2001; Collet, Salvia, and Petit-Boulanger 2014; Engström, Johansson, and Östlund 2005; Lenneman and Backs 2010; Mehler, Reimer, and Coughlin 2012; Solovey et al. 2014). Mehler et al. (2009) explored a range of peripheral physiological measures for differentiating objectively scaled levels of cognitive demand under driving simulation and found heart rate and SCL to be sensitive to task demand and practical to record. The same pattern of response was later observed for these two measures during actual highway driving (Reimer and Mehler 2011) and their sensitivity was further characterised (Mehler, Reimer, and Coughlin 2012). Heart rate and SCL were thus selected for inclusion in the current study.

The locations of R-wave peaks in the EKG signal were used to determine inter-beat intervals and calculate instantaneous heart rate using software developed at the MIT AgeLab. In line with existing standards (Task Force 1996), automated detection results were visually reviewed and misidentified and irregular intervals manually corrected. Another MIT-developed data processing package removed high-frequency noise in the skin conductance signal, following Reimer and Mehler (2011), and identified motion artefacts were manually edited.

2.6.4. Visual metrics

The mean duration of individual (single) glances, the percentage of glances per participant greater than 2.0 s and the total time a participant glanced away from the forward road scene during a task were used as the primary glance metrics. These are the metrics specified by NHTSA (2013) in its recommended guidelines for evaluating the visual distraction associated with in-vehicle visual–manual electronic devices. The 'away from the forward road scene' definition means that glances to other driving-relevant locations such as the instrument cluster and the rear and side mirrors are counted as looking away from the forward road scene. Prior to the release of NHTSA's guidelines, a more typical approach to task demand assessment had been to consider only glances to locations functionally relevant to the task under evaluation, such as a display or controls on the instrument cluster (e.g. Driver Focus-Telematics Working Group 2006). Glance behaviour was categorised as part of this study to support both characterisations. While the values obtained with each method differ and are worthy of consideration (e.g. Dopart et al. 2013; Reimer et al. 2013), using one or the other did not appreciably change the relative pattern of results in this data set comparing systems and modalities. Given the potential relevance to ongoing guideline development, the metrics recommended by NHTSA are presented.

Eye orientation measures were quantified following ISO standards (ISO 15007–1 2002; ISO 15007–2 2001) with a glance to a region of interest defined to include the transition time to that object. In the manual coding of video images, the timing of glance is labelled from the first video frame illustrating movement to a 'new' location of interest to the last video frame prior to movement to a 'new' location. A recent analysis of driver glances to the centre stack and other low-angle glances collected under the variable lighting conditions of real-world highway driving compared automated region of interest classification from commercial eye tracking equipment with video recordings (Reimer et al. 2013). The comparison found a high percentage of missing or inaccurate classifications in the automated output. Therefore, glance data for the current study were manually coded based on video of the driver following the taxonomy and procedures outlined in Reimer et al. (2013, Appendix G). Software, now available as open source (Reimer, Gruevski, and Couglin 2014), allowed for rapid frame-by-fame review and coding. Each task period of interest was independently coded by two evaluators. Discrepancies between the evaluators (the identification of conflicting glance targets, missed glances or glance timings that differed by more than 200 ms) were mediated by a third researcher (see Smith et al. (2005) for a discussion of the importance of multiple coders).

2.6.5. Vehicle control metrics

The vehicle control metrics were changes in vehicle speed, standard deviation of speed and steering wheel reversal rates.

2.6.5.1. Vehicle speed. A reduction in mean vehicle speed (forward velocity) has frequently been observed during periods of increased task demand and is often interpreted either as an attempt to increase safety margins or to reduce/manage the concurrent demands of the primary driving task and a secondary task (Angell et al. 2006). Speed reduction during secondary tasks tends to be more apparent when the task requires drivers to take their eyes off the forward roadway and/or actively involves taking a hand off the steering wheel, such as occurs when a driver holds a mobile phone to the ear (Brookhuis, de Vries, and de Ward 1991; Engström, Johansson, and Östlund, 2005; Green 1994; Patten et al. 2004). Nominal increases in mean speed have sometimes been observed during pure auditory-vocal tasks such as conversing on a hands-free phone (Patten et al. 2004).

2.6.5.2. Standard deviation of vehicle speed. Variability in speed can be influenced by a range of factors, such as changes in the roadway environment and interactions with other drivers. To the extent that road conditions remain relatively constant, increased variability in speed can be interpreted as a reduction in direct attention to vehicle control and has been used at various times as a measure of control and/or changes in driver workload associated with secondary tasks (Green 1994; Noy 1990; Östlund et al. 2005).

2.6.5.3. Steering wheel reversal rates. During normal driving conditions, drivers typically make small steering wheel corrections to adjust vehicle heading for variations in roadway conditions (Liu, Schreiner, and Dingus 1999). When visual attention is diverted from the roadway ahead, a driver's ability to make modest tracking responses is generally suspended until visual orientation to the roadway is regained. This results in periods of fixed steering wheel angle (Godthelp, Milgram, and Blaauw 1984) and the need to make larger corrections upon return of the eyes to the roadway. Similarly, taking a hand off the steering wheel to operate a secondary control can result in more marked adjustments in steering. Östlund et al. (2004) found that visually demanding secondary tasks often invoke relatively large steering reversals of 2°–6°, findings that were replicated in Engström, Johansson, and Östlund (2005). It is appropriate to note that it has been argued that steering wheel reversal rate is not a simple function of secondary task demand, but rather involves a complex interaction between primary task demand, secondary task demand(s) and the effort invested in the different tasks. McDonald and Hoffman (1980) suggested that steering frequency measures such as steering wheel reversal rate can reflect control effort and are not just a measure of tracking performance.

For purposes of this evaluation, *major steering wheel reversals* were considered as a control metric and classified as proposed in the final report of the European Union AIDE project (deliverable D2.2.5, Section 7.12) (Östlund et al. 2005). This metric captures the number of steering wheel inputs exceeding an angular reversal gap of 3°. The rate of steering wheel reversals per minute was obtained by dividing the raw reversal rate by the task trial duration.

2.7. Data reduction and analysis

2.7.1. Subjective workload, behavioural and physiological measures

Baseline driving reference periods consisted of 2 min of just driving prior to a recorded audio message indicating that a new task period was about to start (see Figure 1). There were six such baseline periods per participant on the I-495 portion of the drive, and a seventh 2-min reference was recorded on I-93 south on the return to MIT (14 min total). Values for relevant metrics were calculated, and the mean values across the baseline periods were used as an overall baseline/'just driving' reference. As already described, two trials of each type of phone calling and three trials of address entry using an embedded vehicle interface were presented to each participant. Values for each dependent measure were calculated per trial and mean values across trials were used for analytic purposes. All trials with usable data were included regardless of whether user or system errors occurred (see Section 2.7.2 for more detail on error states and how they were handled). Trials with errors were included in the analysis as this was seen as more representative of the actual user experience than only considering error-free trials.

An evaluation of mean speed based on vehicle CAN data indicated a significant overall difference in speed between the two vehicles (Volvo $M = 107.5$ km/h; Chevrolet $M = 111.6$ km/h; $F(1,78) = 5.4$, $p < 0.023$); this difference was apparent even during baseline just driving periods (109.4 and 113.1 km/h, respectively; $F(1,78) = 8.9$, $p = 0.004$). However, there was no significant difference in overall speed when considering data from the GPS units installed in each vehicle (Volvo $M = 110.2$ km/h; Chevrolet $M = 108.5$ km/h; $F(1,78) = 0.53$, $p = 0.468$). This suggests that the CAN values may have systematically underestimated actual vehicle speed in the Volvo and overestimated speed in the Chevrolet. As a result, speed data from the GPS units were normalised as percentage changes relative to baseline driving periods for purposes of comparing the effects of interaction with the embedded systems in each vehicle. For consistency, GPS-based values were used for considering changes in standard deviation of speed as well.

Major steering wheel reversal rates were markedly higher during baseline driving in the Chevrolet ($M = 20.39$ per minute, SE = 0.9) vs. the Volvo ($M = 3.29$, SE = 0.2) ($F(1,78) = 234.1$, $p < 0.001$). This could reflect basic tuning of the steering, other handling characteristics of the two vehicles and/or differences in the quantification of steering wheel angle on the respective CAN buses. Consequently, comparisons of steering wheel metrics are reported considering per cent changes relative to baseline driving.

Statistical analyses were performed in R (R Core Team 2014) and an alpha level of 0.05 was used for assessing statistical significance. Owing to the non-normal distribution of the data and/or the use of ratio data (percentages) for several dependent measures, in many cases, non-parametric statistics such as the Wilcoxon signed rank test and the Friedman test were used (similar to the *t*-test and repeated-measures ANOVA, respectively). For multi-factorial analyses, repeated-measures ANOVA by ranks are presented. These tests have been shown to be more robust against Type I error in cases where data are non-normal (Conover and Iman 1981; Friedman 1937).

There were substantive differences between the contact calling and address entry tasks. For example, independent periods of contact calling with voice commands and manual entry were considered, while address entry was undertaken with voice commands only. Consequently, separate analyses were conducted for the two types of secondary tasks. The design for the contact calling was a mixed design with vehicle and the associated embedded system as a between-subject variable (MyLink or Sensus). There were two within-subject factors, modality (manual entry or voice entry) and task difficulty (easy or hard), resulting in a $2 \times 2 \times 2$ mixed design. The full model was used in the analysis of the self-reported workload and task completion time data where effects for the easy and hard categorisations were of most interest for characterising system implementation differences. Task

difficulty was dropped from the model for analysis of physiology, eye glance and driving performance metrics as typical use of the technologies would likely involve a mix of the easy and hard categories of contact calling. The analysis for the address entry task considered only embedded system (MyLink or Sensus) as a between-subject factor. Where applicable, tests comparing differences on selected variables between baseline driving and periods with the phone calling and navigation tasks are presented.

2.7.2. Error analysis and interaction characterisation

A multi-step analysis of participants' interactions with the vehicle systems was carried out. The first analysis considered for each individual task trial whether it was error free or whether a system or user-based error occurred. An example of a user error is a participant speaking an incorrect command during a voice-entry task that resulted in the task moving forward incorrectly or not moving forward at all. An example of a system error is the system misinterpreting a voice command that was in the correct form and understandable to the research associate present in the vehicle or a staff member listening to an audio recording of the interaction. Two members of the research staff independently evaluated each trial for errors (the research associate observing the participant during the drive and a second staff member who reviewed video and audio recordings of the interaction). A third staff member mediated any discrepancies. For the binary classification of whether a user or system error occurred during a trial, it was decided that if a user error and system error occurred in the same trial, the trial would be coded as a user error regardless of the number of each error type in the same trial. Thus, it is likely that the rate of system errors is under-represented in this analysis.

The second error analysis was a more fine grained characterisation of the extent to which participants experienced any difficulty in completing a task. Individual trials were classified as: (1) being completed without error or backtracking, (2) completed with backtracking, (3) completed with one instance of the research associate providing a prompt to assist the participant, (4) completed with more than one prompt from the research associate, or (5) failure to complete the task. The 'backtracking' category covered situations where, for example, the system did not recognise or misinterpreted a street name, but the system dialog asked for confirmation and allowed for another opportunity for entry without exiting and requiring the participant to begin the entire task from the start. In other words, a backtracking classification indicates that the system successfully supported error recovery (arising from either user error or system recognition error). Backtracking could also occur because a participant recognised that they made an error (such as giving a wrong street name) and used an option provided by the system to correct the error. If the research associate judged that a participant was not going to progress through a task on their own, one or more limited prompts were provided to the participant. The intent here was largely to provide the participant with further assistance in learning how to use the system so that they might gain additional familiarity and become more successful on subsequent trials. If a participant had to restart a task more than twice or otherwise failed to progress at a point in the interaction despite support from the research associate, then the research associate guided the participant through terminating the trial and moved on. Failure to progress could be due to either user or system errors. Trials that were terminated or that failed to progress due to either user or system errors were categorised as a failure.

3. Results

Beyond the 80 participants considered in the age- and gender-balanced analysis sample, there were a number of task relevant exclusions. These included eight individuals who were not taken on-road: two who experienced consistent voice recognition problems with a vehicle voice system (both with the Chevrolet MyLink); two who expressed discomfort with the idea of engaging in one or more of the tasks while driving after being exposed to them during training (both female, 64 years of age); and four who experienced significant difficulty trying to learn tasks in the parking lot (all male, 45–64 years). Of individuals taken on-road, exclusions included one (63-year-old male) who was consistently unable to recall the actions necessary to complete tasks, requiring continuous prompting by the research associate; two (56- and 65-year-old males) for whom the research associate discontinued presentation of one or more task sets due to concerns on the research associate's part regarding the participant's ability to engage in tasks safely while driving. Other non-task-relevant exclusions included three participants who were withdrawn during the drive due to broader safety concerns (one expressed drowsiness while driving, one frequently drifted out of lane and one with other unsafe driving habits) and four cases where weather and/or traffic conditions precluded continuing.

Findings for the analysis sample are presented first for the phone calling tasks followed by results for destination address entry tasks. Each section considers participants' subjective assessment of the workload associated with each task followed by objective data that include task duration, physiological measures, glance behaviour, vehicle control metrics and secondary task errors.

3.1. General sensitivity of physiological and driving metrics to secondary task periods

Changes in physiological arousal are characterised for analysis purposes as percentage changes relative to baseline driving to account for the different base values of individual participants. As expected, engaging with the secondary tasks while driving was associated with a higher state of arousal. Relative to baseline driving, there was on average an increase during

the phone tasks across modalities and systems in heart rate of 2.2% (SE = 0.8) ($W = 2516$, $p < 0.001$) and an increase in SCL of 11.3% (SE = 3.1) ($W = 2397$, $p < 0.001$). During the voice command-based destination address entry across both systems, heart rate increased on average 1.5% (SE = 0.5) ($W = 2112$, $p = 0.018$) and SCLs increased 7.3% (SE = 2.4) ($W = 1956$, $p = 0.002$).

Mean speed decreased significantly across the combined manual and voice-based phone calling tasks periods ($M = -2.5\%$, SE = 1.1; $W = 680$, $p < 0.001$) although not during the voice command-based address entry task periods ($M = -0.4\%$, SE = 0.6; $W = 1510$, $p = 0.559$). Standard deviation of speed decreased across the manual and voice-based phone calling tasks ($M = -37.6\%$, SE = 5.2; $W = 76$, $p < 0.001$) and the voice-based destination entry tasks ($M = -19.9\%$, SE = 5.0; $W = 584$, $p < 0.001$). The rate of major steering wheel reversals increased significantly across the combined manual and voice-based phone calling tasks ($M = 31.9\%$, SE = 5.0; $W = 2585$, $p < 0.001$) but not during voice-based address entry ($M = -0.49\%$, SE = 4.2; $W = 1494$, $p = 0.547$).

3.2. Phone contact calling

In considering the phone calling tasks, 'modality' refers to the overt method of interface interaction (manual or voice) and 'difficulty' refers to the easy or hard form of the task. Table 2 provides the means and standard errors of the measures used for analysis of the contact calling tasks presented by modality and embedded system type (Chevrolet MyLink or Volvo Sensus). An expanded set of tables providing details on measures not directly used in the analysis, such as alternate glance metrics, is provided in Tables A1 and A2 in Appendix 1.

3.2.1. Self-reported workload

A full breakdown of means and standard errors for self-reported workload by modality (manual or voice), system (Chevrolet MyLink or Volvo Sensus) and task difficulty (easy vs. hard) are presented in Figure 2 and Table 2. Self-reported workload for phone calling differed significantly by modality ($F(1,76) = 144.1$, $p < 0.001$) and difficulty level ($F(1,76) = 32.9$, $p < 0.001$). Mean ratings were higher for manual phone calling ($M = 5.3$, SE = 0.40) than for voice-based calling ($M = 2.1$, SE = 0.28). On average, the hard phone calling task had higher workload ratings than the easy phone calling task (easy $M = 3.4$, SE = 0.33; hard $M = 3.9$, SE = 0.35). On average, subjective workload ratings were lower with MyLink ($M = 3.3$, SE = 0.35) compared with the ratings with Sensus ($M = 4.0$, SE = 0.33); however, this difference only approached statistical significance ($F(1,76) = 3.31$, $p = 0.07$). There were no significant interactions between embedded system, modality or task difficulty ($p > 0.16$).

3.2.2. Task completion time

In contrast with self-reported workload, there was no main effect of modality on task completion time ($F(1,78) = 1.14$, $p = 0.288$), but there was a significant difference between the systems ($F(1,78) = 89.9$, $p < 0.001$) and a significant interaction between modality and system ($F(1,78) = 37.6$, $p < 0.001$) (see Figure 3). On average, participants using MyLink took longer to complete the phone calling task using the manual interface ($M = 26.2$ s, SE = 1.5) than the voice interface ($M = 21.6$ s, SE = 1.6). Conversely, participants using Sensus took longer using the voice interface ($M = 38.2$ s, SE = 1.5) than the manual interface ($M = 32.9$ s, SE = 2.3).

Significant interactions between modality and difficulty ($F(1,78) = 15.71$, $p < 0.001$) and between system and difficulty ($F(1,78) = 12.4$, $p < 0.001$) were present (see Table 2). Compared with the easy phone calling task, the hard phone calling task took longer to complete when participants used the voice interfaces (easy $M = 27.5$ s, SE = 1.3; hard $M = 32.3$ s, SE = 1.7) but took less time when using the manual interfaces (easy $M = 30.3$ s, SE = 2.0; hard $M = 28.8$ s, SE = 1.8). The task completion times for the easy phone calling and hard phone calling tasks were similar with MyLink ($M = 20.5$ s, SE = 1.3 and $M = 22.7$ s, SE = 1.8, respectively), but the hard phone calling task took 21% longer to complete than the easy phone calling task with Sensus (41.9 and 34.5 s, respectively).

3.2.3. Physiological measures

There were no significant main effects of modality, system or modality by system interaction across the tasks in terms of percentage change in either heart rate or SCL during task periods relative to baseline (all $p > 0.05$). There was a three-way interaction between modality, system and task difficulty for percentage change in heart rate ($F(1,78) = 15.2$, $p = 0.002$). The percentage change in heart rate among participants who used the MyLink voice interface to complete the hard contact calling task ($M = 3.75\%$, SE = 0.8) was greater than that for participants who used the voice interface in Sensus ($M = 0.97\%$, SE = 0.6). The difference in percentage change in heart rate between MyLink and Sensus was not observed for the easy contact calling task using the voice interfaces or the easy or hard contact calling tasks with the manual interfaces.

3.2.4. Glance behaviour

There were significant main effects of modality ($F(1,78) = 204.8$, $p < 0.001$) and system $F(1,78) = 10.6$, $p = 0.002$), and a significant interaction between modality and system ($F(1,78) = 24.5$, $p < 0.001$) for mean single glance duration. Mean single glance duration for off-road glances was shorter when phone calling was performed using a voice interface ($M = 0.69$ s, SE = 0.02) compared with the manual interface ($M = 0.93$ s, SE = 0.02) in both vehicles; however, the reduction in single glance duration during voice interaction compared with manual interaction was greater with MyLink than Sensus (see Figure 4).

Table 2. Means (and standard errors) by phone calling task and embedded vehicle system (Chevrolet MyLink or Volvo Sensus) for measures used for analysis.

	Vehicle	Phone easy (manual)	Phone hard (manual)	Phone easy (voice)	Phone hard (voice)
Self-reported workload	Chevrolet	4.28 (0.4)	5.20 (0.4)	1.81 (0.3)	1.90 (0.3)
	Volvo	5.49 (0.4)	6.12 (0.4)	2.05 (0.2)	2.55 (0.3)
Task completion time	Chevrolet	29.18 (2.0)	23.30 (0.9)	20.48 (1.3)	22.74 (1.8)
	Volvo	31.43 (2.0)	34.36 (2.6)	34.48 (1.3)	41.87 (1.6)
% change in heart rate	Chevrolet	2.54 (0.9)	1.12 (0.8)	2.46 (0.8)	3.75 (0.8)
	Volvo	2.07 (1.0)	2.10 (0.7)	2.47 (0.7)	0.97 (0.6)
% change in SCL	Chevrolet	15.06 (3.8)	13.15 (3.0)	13.66 (3.3)	12.22 (3.2)
	Volvo	13.09 (3.0)	12.75 (2.9)	7.62 (2.8)	3.63 (2.7)
Mean off-road glance duration	Chevrolet	0.89 (0.0)	0.94 (0.0)	0.60 (0.0)	0.61 (0.0)
	Volvo	0.94 (0.0)	0.95 (0.0)	0.79 (0.0)	0.79 (0.0)
% of off-road glances > 2.0 s	Chevrolet	1.73 (0.6)	3.12 (0.8)	0.09 (0.1)	0.49 (0.5)
	Volvo	2.98 (1.0)	3.80 (1.0)	0.94 (0.4)	0.57 (0.2)
Total off-road glance time	Chevrolet	15.16 (1.2)	11.97 (0.7)	3.42 (0.5)	3.23 (0.4)
	Volvo	15.95 (1.1)	16.82 (1.4)	9.78 (0.7)	10.65 (0.9)
Number of off-road glances	Chevrolet	16.74 (1.1)	12.82 (0.6)	5.26 (0.7)	5.05 (0.6)
	Volvo	16.96 (1.0)	17.70 (1.3)	12.44 (1.0)	13.46 (1.0)
% change speed (GPS)	Chevrolet	−3.64 (1.6)	−4.63 (1.9)	−1.09 (1.0)	−0.13 (1.1)
	Volvo	−4.03 (0.9)	−2.14 (0.6)	−1.72 (0.7)	−2.56 (1.1)
% change in SD of speed (GPS)	Chevrolet	−33.48 (5.6)	−41.58 (4.3)	−54.76 (4.3)	−53.90 (3.6)
	Volvo	−27.36 (9.4)	−36.24 (3.8)	−30.57 (4.2)	−22.80 (6.3)
% change in major wheel reversals	Chevrolet	23.26 (8.4)	27.58 (7.2)	28.63 (9.2)	16.80 (10.5)
	Volvo	51.41 (13.4)	38.11 (16.0)	31.31 (17.0)	37.84 (14.8)

Note: Time metrics in seconds.

On average, only a small percentage of participants' glances were longer than 2 s (M = 1.4%, SE = 0.3) (see Figure 5). Nonetheless, there was a significant main effect of modality ($F(1,78) = 39.0$, $p < 0.001$). On average, the percentage of glances that were longer than 2 s for each participant was smaller when using the voice interfaces (M = 0.5%, SE = 0.2) compared with using the manual interfaces (M = 2.9%, SE = 0.5). There was no significant main effect of system type ($F(1,78) = 2.1$, $p = 0.149$) or interaction between modality and system ($F(1,78) = 0.100$, $p = 0.768$).

For total eyes-off-road time, there were significant main effects of modality ($F(1,78) = 266.8$, $p < 0.001$) and system ($F(1,78) = 35.3$, $p < 0.001$), and a significant interaction between modality and system ($F(1,78) = 30.6$, $p < 0.001$). The mean values for total eyes-off-road time were less when participants completed the phone calling task using the voice interfaces (M = 6.8 s, SE = 0.6) than when using the manual interfaces (M = 15.0 s, SE = 0.7); however, the reduction in total eyes-off-road time associated with using the voice interface relative to the manual interface was much greater for participants using MyLink than participants using Sensus (Figure 6). There was no significant main effect of task difficulty on total eyes-off-road time ($F(1,78) = 0.78$, $p = 0.379$).

3.2.5. Vehicle control metrics

As previously noted, on average, participants decreased their speed somewhat during the phone task periods. There was a greater percentage reduction in speed relative to baseline during manual phone calling (M = 3.6%, SE = 1.3) compared with voice phone calling (M = 1.4%, SE = 1.0) ($F(1,78) = 10.5$, $p = 0.002$). No significant main effect of system ($F(1,78) = 1.65$, $p = 0.202$) or interaction appeared ($F(1,78) = 0.22$, $p = 0.643$).

The was no overall main effect of modality on the percentage change in standard deviation of speed ($F(1,78) = 2.02$, $p = 0.159$). However, a significant main effect of system ($F(1,78) = 12.01$, $p = 0.001$) was present. On average, the reduction in standard deviation of speed during phone calling was greater for participants who used MyLink (M = −45.9%, SE = 4.5) than for those who used Sensus (M = −29.2%, SE = 5.9). In addition, a system by modality interaction was observed ($F(1,78) = 21.55$, $p < 0.001$) (see Figure 7). Detailing the interaction, there were no significant differences in the percentage reduction in standard deviation of speed between manual and voice calling with the Sensus (manual M = −31.8%, SE = 5; voice M = −26.7%, SE = 5.3) or between the two manual interfaces. In contrast, there were on average greater reductions in the standard deviation of speed with MyLink voice (M = −54.3%, SE = 4.0) than the manual (M = −37.5%, SE = 5.0) mode, and the reduction in the MyLink voice calling condition was greater than both Sensus conditions.

Overall, using the percentage change from baseline driving metric, the relative increase in major steering wheel reversals was nominally higher during manual calling (M = 35.1%, SE = 6.5) than during voice command-based calling (M = 28.6%, SE = 7.6); however, the difference was not statistically significant ($F(1,78) = 2.14$, $p = 0.148$). There was no significant main effect of system on major steering wheel reversal rates ($F(1,78) = 0.58$, $p = 0.45$) and no significant interaction between system and modality ($F(1,78) = 0.16$, $p = 0.69$).

3.3. Destination address entry into a navigation system

Descriptive statistics and analytic results considering the extent to which significant differences appeared between participant groups using the two voice command-based systems to enter destination addresses are presented in Table 3. An expanded listing including alternate eye glance metrics and absolute values for measures prior to conversion to percentage change scores appear in Table A3 in Appendix 1.

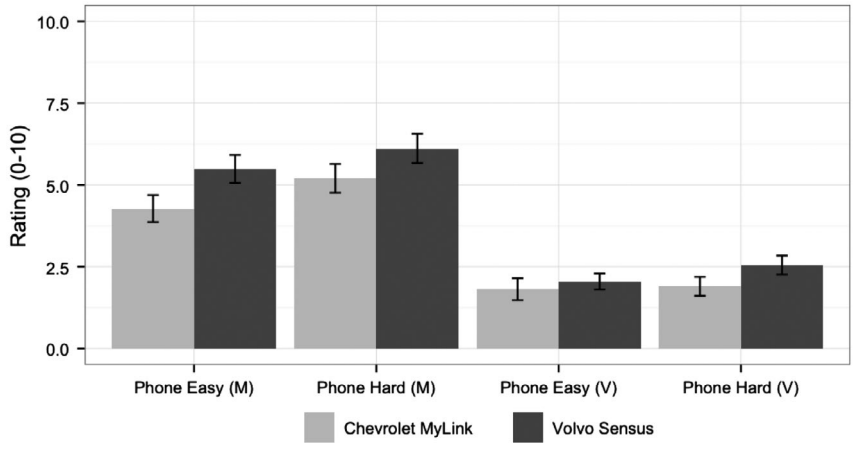

Figure 2. Mean self-reported workload ratings for all phone tasks by modality (manual or voice) and embedded system type (Chevrolet MyLink or Volvo Sensus) on a 0 (low) to 10 (high) scale.
Note: Error bars represent ±1 standard error.

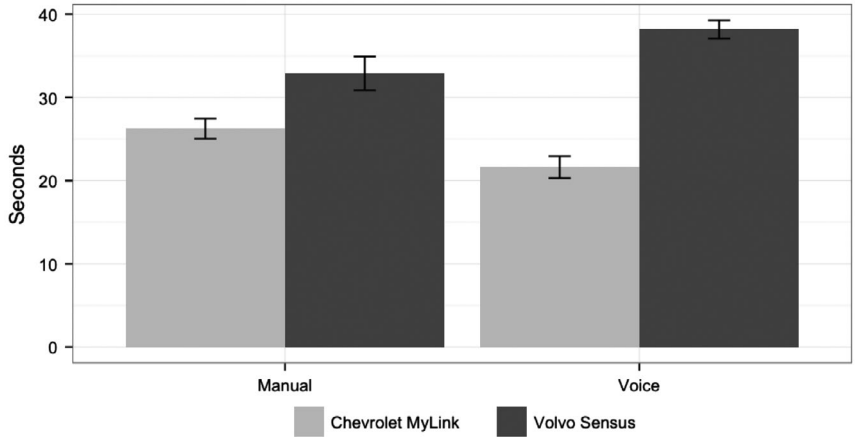

Figure 3. Mean completion time for phone calling by modality and type of embedded system.
Note: Error bars represent ±1 standard error.

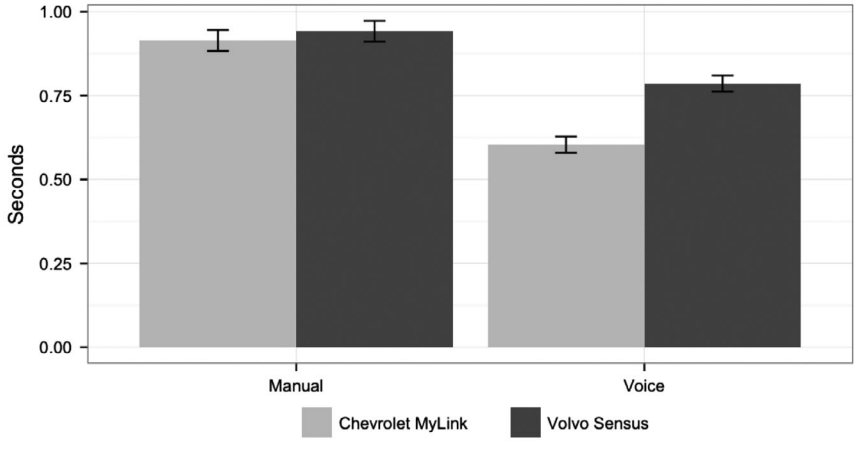

Figure 4. Mean single glance duration for all off-road glances during task periods by interface type and embedded system type.
Note: Error bars represent ±1 standard error.

3.3.1. Subjective workload

Mean self-reported workload for navigation address entry was nominally higher for the MyLink system (M = 3.59; SE = 0.44) than for Sensus (M = 2.54; SE = 0.28); however, this difference did not reach statistical significance (W = 925, p = 0.15).

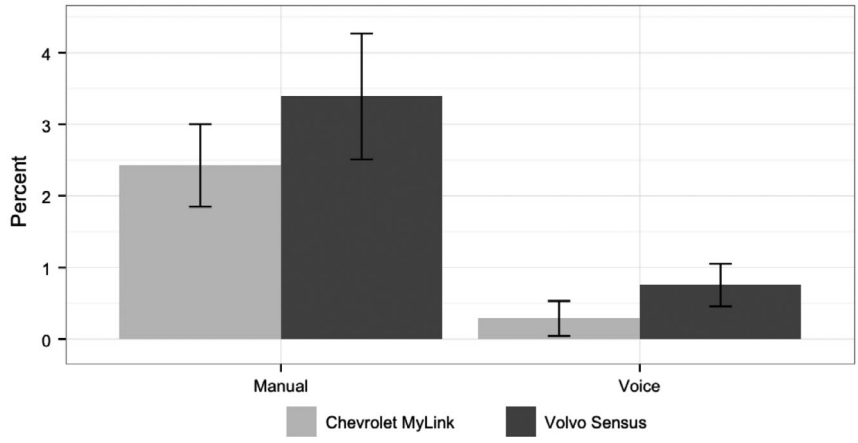

Figure 5. Per cent of off-road glances greater than 2 s in duration.
Note: Error bars represent ±1 standard error.

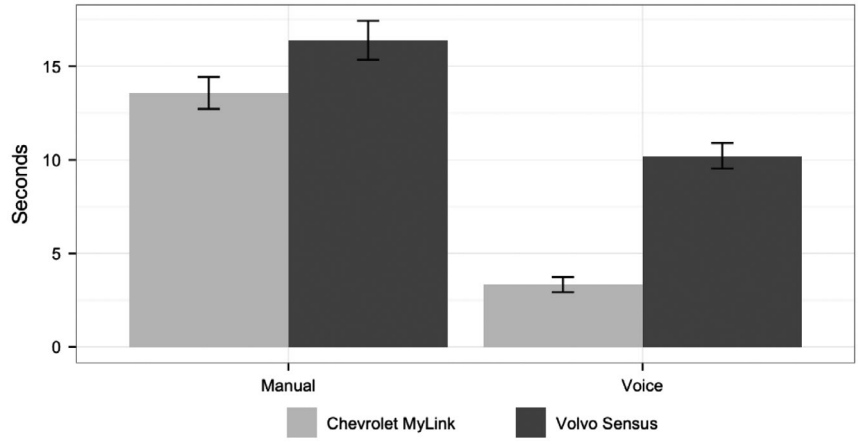

Figure 6. Total eyes-off-road time.
Note: Error bars represent ±1 standard error.

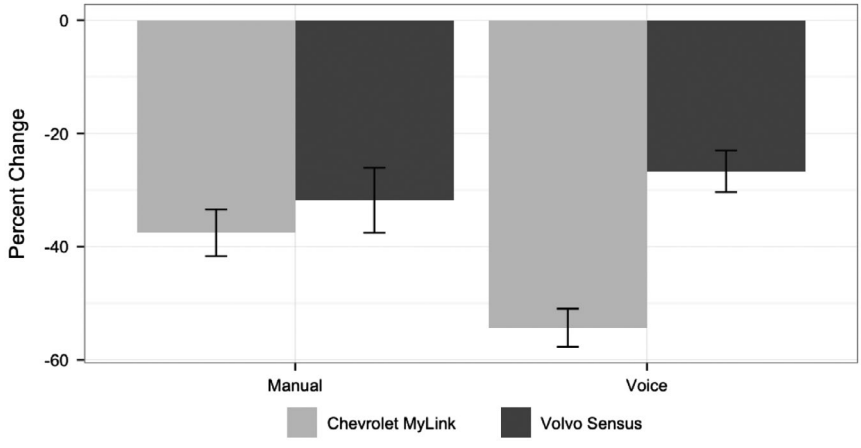

Figure 7. Mean per cent change in standard deviation of speed (GPS) during phone task periods relative to baseline.
Note: Error bars represent ±1 standard error.

3.3.2. Task completion time

There was a significant main effect of system on the time it took to complete the navigation address entry task ($W = 408$, $p < 0.001$). On average, participants using MyLink ($M = 66.7$ s, SE = 2.85) completed the address entry task in less time than participants using Sensus ($M = 80.6$ s, SE = 1.71).

Table 3. Means (and standard errors) and results of Wilcoxon signed rank tests for the destination address entry tasks.

	Chevrolet	Volvo	*W*	*p*-value
Self-reported workload	3.59 (0.44)	2.54 (0.28)	924.5	0.154
Task completion time	66.68 (2.85)	80.60 (1.71)	408	<0.001*
% change in heart rate	1.66 (0.87)	1.25 (0.67)	801	0.996
% change in skin conductance level	11.59 (3.77)	3.29 (2.44)	811	0.172
Mean off-road glance duration	0.74 (0.02)	0.82 (0.02)	562	0.022*
% of off-road glances > 2.0 s	1.02 (0.29)	1.27 (0.36)	777.5	0.813
Total off-road glance time	14.28 (1.22)	22.56 (1.43)	367	<0.001*
Number of off-road glances	18.65 (1.52)	27.77 (1.75)	397	<0.001*
% change in speed (GPS)	0.60 (0.62)	−0.98 (0.55)	990	0.068
% change in SD of speed	−29.53 (4.58)	−10.35 (5.46)	550	0.016*
% change in major wheel reversals	7.34 (6.10)	−8.32 (6.82)	1003	0.051

Change scores represent the percentage (%) change from baseline just driving.
*$p < 0.05$. Time metrics in seconds.

Table 4. Number of trials with errors and breakdown by type of error.

System	Task	Trials	Error free	System errors	User errors	Total errors
Chevrolet MyLink	Calling manual	160	153	0	7	7
Volvo Sensus	Calling manual	160	144	0	16	16
Chevrolet MyLink	Calling voice	160	147	5	8	13
Volvo Sensus	Calling voice	160	149	2	9	11
Chevrolet MyLink	Address entry	120	59	38	23	61
Volvo Sensus	Address entry	120	107	5	8	13

3.3.3. Physiological measures

While heart rate and SCL were higher during address entry than during baseline driving (see Section 3.1), there was no significant effect of system for the percentage change in heart rate during address entry relative to baseline driving (MyLink $M = 1.7\%$, SE = 0.9; Sensus $M = 1.3\%$, SE = 1.7; $W = 801$, $p = 0.996$) or the percentage change in SCL (MyLink $M = 11.6\%$, SE = 3.8; Sensus $M = 3.3\%$, SE = 2.4; $W = 811$, $p = 0.172$).

3.3.4. Glance behaviour

Mean single off-road glance duration during navigation address entry was significantly shorter for participants using MyLink ($M = 0.74$ s, SE = 0.02) compared with participants using Sensus ($M = 0.82$ s, SE = 0.02) ($W = 562$, $p = 0.022$). Similarly, the average total off-road glance time was significantly shorter for participants using MyLink ($M = 14.3$ s, SE = 1.2) than participants using Sensus ($M = 22.6$ s, SE = 1.4) ($W = 367$, $p < 0.001$). The overall number of long-duration glances was low, and there was no significant main effect of system on the percentage of glances made by a participant that were longer than 2 s ($W = 777.5$, $p = 0.81$).

3.3.5. Vehicle control metrics

The main effect of system on the percentage change in mean speed during navigation address entry relative to baseline approached statistical significance ($W = 990$, $p = 0.068$). Speed nominally increased among participants who used MyLink ($M = 0.6\%$, SE = 0.62) but decreased for participants who used Sensus ($M = -1.0\%$, SE = 0.55). Participants using MyLink showed a significantly greater reduction in their standard deviation of speed relative to baseline ($M = -29.5\%$, SE = 4.6) than participants using Sensus ($M = -10.4\%$, SE = 5.5) ($W = 550$, $p = 0.016$). In terms of the percentage change in major steering wheel reversal rate relative to baseline driving, there was a nominal difference associated with system type during address entry ($W = 1003$, $p = 0.051$). The percentage change in major steering wheel reversal rate was 7.34% (SE = 6.1) for participants using MyLink and −8.3% (SE = 6.8) for participants using Sensus.

3.4. Errors and interaction characterisation

Errors occurred in 7.3% of the phone calling trials (47 of 640 trials). As can be observed in Table 4, errors attributable to a system were virtually non-existent for manual contact calling (1 trial) and were present 2% of the time (7 trials) for voice command entry. Considering both modalities together, user errors when attempting to call a contact were more prominent than system errors, occurring in 6.1% of the trials ($W = 477.0$, $p < 0.001$). If user and system errors are combined as a measure of usability and the two systems are considered together, no generalised advantage in frequency of trails with error appears by modality (manual: 23 trials; voice: 24 trials).

The overall rate of errors for voice command-based entry of a destination address was markedly higher (30.8%) than the rate for voice-based phone calling (7.5%) (see Table 4). A significant difference by system was also apparent ($W = 1128.5$, $p = 0.001$). An error occurred in more than half of the address entry trials (51%) for MyLink compared with 10.1% for Sensus. Comparing error types, system-based errors represented a larger percentage of the errors for MyLink (38 of 61 trials with errors; 62.3%) than Sensus (5 of 13 trials with errors; 38.5%).

Given the much higher error rates for address entry, Figure 8 provides a characterisation of the relative degree of difficulty participants experienced with each embedded navigation system in each of the three trials. It can be observed that only two outright failures to input the correct address occurred among participants using Sensus vs. 24 failures experienced with MyLink. It can also be seen in Figure 8 that the address for trial 2 was more challenging to enter in both vehicles. Trial 3 consisted of the entry of each driver's own home address. It can be observed that during trial 3, 38 of 40 participants

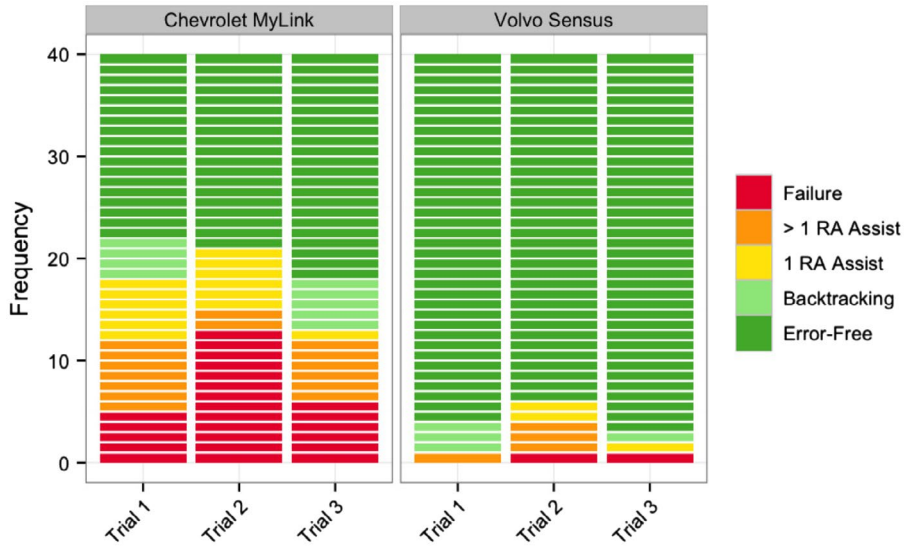

Figure 8. Characterisation of participant experience by trial for destination address entry. The stacked scaling represents individual drivers sorted by their experience for an individual trial (i.e. 40 drivers per vehicle). 'RA Assist' refers to prompting support provided by a research assistant as detailed in the methods.

were able to successfully enter their home address without external support using Sensus, while a more modest 27 of 40 were successful using MyLink.

4. Discussion

The findings for the embedded phone calling tasks add to previous research indicating that using voice interfaces to interact with an 'infotainment' system can significantly reduce subjective workload and visual demand compared with using a manual interface. With both the Chevrolet MyLink and the Volvo Sensus embedded systems, participants reported significantly lower levels of subjective workload, had shorter mean single off-road glance durations, had fewer off-road glances longer than 2 s, and spent less time looking away from the forward roadway during voice command phone calling compared with manual phone contact calling.

While participants assigned to both vehicles experienced a number of apparent advantages using voice commands relative to manual input for the embedded phone tasks, there are still potential trade-offs to be considered in evaluating net benefits and method of interaction more generally. For example, depending on the nature of the task and the implementation, voice command interactions can take longer than using a manual interface. In an examination of radio tuning (Mehler et al. 2014; Reimer et al. 2013), manually pressing a radio preset button took less time than depressing a press-to-talk button and then verbally requesting a preset. In contrast, verbally requesting a specific station was faster and resulted in less diversion of the eyes from the roadway than making multiple button presses to change modes and frequency band and then manually rotating a tuning knob. Thus, a traditional manual interface seems to be more advantageous in the first case and the voice command option more advantageous in the latter. This study extends upon this level of detail by characterising the extent to which system implementation differences can impact various variables. Consistent with the hypotheses that stemmed from Reagan and Kidd (2013), manual phone contact calling took more time than voice contact calling with the MyLink interface, whereas the opposite was true with the Sensus interface.

As is evident in the task completion time results, design philosophy and implementation differences in the voice command-based systems can significantly impact objective metrics. Overall, the Sensus approach broke the task into discrete steps; this was most evident in the navigation system, which dealt with city, street name and street number independently. In contrast, MyLink employed an initial 'one-shot' approach in which the full address was presented in a single vocal string. With phone calling, the vocal string could be relatively simple (e.g. 'Call Frank Scott at work'.), and this approach worked well for almost all participants. For address entry, however, results were quite different. When MyLink successfully parsed and decoded a one-shot full address string, the task was completed relatively quickly. However, a trade-off appears in a higher failure rate due to voice recognition errors by the system and user input errors. Only two outright failures in address entry occurred using Sensus, while 24 were recorded for MyLink in the analysis sample.

It is also worth considering the extent to which implementation characteristics outside of the fundamental voice recognition system design and capabilities might play a role in observed recognition errors. As detailed in a companion report (Reimer et al. 2015), voice recognition tasks in a dash-mounted smartphone also were evaluated in both vehicles. Although the same smartphone was being used, voice recognition errors were found to be higher in the Chevrolet Equinox than in the Volvo XC60. *Post hoc* sound level readings taken while

the vehicles were travelling at 65 mph found that the Chevrolet had louder ambient noise levels than the Volvo in the 250 Hz (Chevrolet: 65 dBA; Volvo: 62 dBA) and 1000 Hz bands (Chevrolet: 62.6 dBA; Volvo: 60.1 dBA). Thus, one hypothesis to explain some of the variance in voice recognition errors might be the impact of ambient noise levels. This highlights the broader issue of system integration in automotive and other contexts, e.g. considering the optimisation of a voice system in the overall vehicle environment.

4.1. Training and mental models

In addition to voice recognition errors, some level of research staff prompting was required in a much higher percentage of cases during address entry while underway with the MyLink system, in spite of the fact that everyone was trained on the interface in the parking lot prior to going on-road. During the third trial, where participants were entering their home address, only one driver using Sensus required prompting to successfully complete the entry. For drivers working with MyLink, 7 needed prompting assistance to successfully complete the task.

It is possible that some of this performance differential may disappear if a user gains additional experience with a system. According to the subjective impressions of the research staff, a significant challenge for participants using the one-shot interface was learning to speak full addresses relatively rapidly and in a continuous stream, i.e. without pauses between a street name and the city name or long enunciation of individual digits of a street number. It appeared that 'trying to help the system' by speaking slowly and with pauses between elements was, in fact, not a good strategy with this system. Designing systems that work with speech spoken in a relatively natural, continuous stream without pauses should ultimately benefit the consumer. However, this exemplifies the challenge of how to communicate functional design characteristics to novice users when they do not have a mental model for system operation or where their existing mental model does not match the implementation. It is likely that a frustrated user may limit, or discontinue altogether, use of a system that proves difficult to use initially. It is also plausible that better understanding of a system's model of operation would lead to more use and increase the potential advantages of using voice-based interfaces over manual interfaces. Further research could assess such a hypothesis in a longitudinal study.

4.2. Visual demands of voice interfaces

While the embedded voice command-based interfaces studied here were associated with lower visual demand than the embedded manual interfaces for phone calling, they were still highly multi-modal, including manual interactions and involving measurable time looking off the roadway. Visual engagement associated with a voice–command interface can vary markedly depending on the system design approach and the type of task. Total eyes-off-road time during voice-based phone calling was relatively brief, with a mean of around 3.3 s for MyLink and a notably higher 10.3 s with Sensus. During voice-based address entry, the mean total eyes-off-road time was 14.3 s with MyLink and significantly longer at 22.6 s with Sensus. Relatively long total eyes-off-road times were also observed during address entry in a 2010 Lincoln MKS system which employed a menu-based approach similar to the Sensus (Mehler et al. 2014; Reimer et al. 2013).

4.3. Cognitive demands

In addition to the visual demands documented here, the question of the extent to which cognitive demands are an issue in voice command systems remains a valid and challenging question. Reagan and Kidd (2013) specifically note the concern that although voice interfaces reduce visual demand, secondary activities, regardless of input modality, may produce levels of cognitive demand that may reduce road users' safety compared with just driving. Studies have shown that increased cognitive demands result in more constrained visual scanning patterns (Recarte and Nunes 2000; Reimer et al. 2012), suppression of brain activity in visual processing areas (Just, Keller, and Cynkar 2008) and degradation of vehicle control on the test track (Owens, McLaughlin, and Sudweeks 2011). Likewise, Lo and Green (2013) observed that voice interfaces have been shown to offer various advantages, but still require cognitive demand, which can interfere with the primary driving task. Strayer et al. (2013) provide a particularly extensive review of reasons why cognitive demands arising from auditory–vocal interactions with technologies could be problematic when driving.

Viewed broadly, the voice tasks studied here did not appear to produce high cognitive workloads compared with other secondary tasks studied previously (Mehler et al. 2014; Reimer and Mehler 2013). Self-reported workload was lower for both voice-based phone calling and destination address entry than what was reported for manual phone calling. Considering physiological arousal as an indicator of workload, increases were present during all voice and manual tasks, but did not differentiate between modalities. Compared with data collected in Mehler et al. (2014), elevations in heart rate were in the same general range as that induced by the 0-back level of the n-back surrogate working memory task (generally considered a very low cognitive demand task) and skin conductance values were nominally below the 1-back level (generally considered a moderately demanding cognitive task). Thus, while demands with voice interaction were present in the current study, the findings may not warrant the degree of concern raised in recent evaluations of embedded voice systems (e.g. Strayer et al. 2014), particularly when considering the several

measures that indicate lower demand for the two embedded voice systems tested here relative to their manual counterparts. The present work provides additional evidence in two different vehicle implementations that voice-based interfaces are multi-modal in nature, drawing upon auditory, vocal, visual, manipulative and cognitive resources. At a minimum, the consideration of visual demand, a well-established key correlate to safety, must be taken into account in developing a comprehensive assessment of voice interfaces. It is clear that providing a voice interface does not inherently mean that drivers will or can keep their eyes continuously on the road.

Nevertheless, the data collected in the current study do not exhaustively explore the extent to which drivers might become so absorbed in a secondary task that look-but-do-not-see events become an issue or that frustration over problematic voice recognition might divert attention. Well-developed work to better understand the extent of these issues is needed. In this context, comprehensive assessment of cognitive absorption in voice-involved interactions should be considered relative to purely visual–manual alternatives in addition to 'just driving'. For example, two simulation studies of smartphone interactions found that drivers took longer to notice a light stimulus and missed more of the stimuli overall when using voice-based entry of a destination address compared with baseline driving (Beckers et al. 2014; Munger et al. 2014). At the same time, response rates and miss values were significantly lower for the voice-based interactions than for interactions with the visual–manual interface. Thus, while it is important to recognise that voice interfaces are not free of demands on attention, it is also important to better understand the relative risks of different types of interactions while driving and to communicate this understanding to the public.

4.4. Limitations

The data presented characterise the behaviour of drivers who were trained on the use of the information systems tested. Compared with actual owners of a vehicle who use such systems regularly, the population of study had limited experience. Furthermore, their interaction with the systems was evaluated at designated points during a structured drive. It is unknown how such an experimental evaluation mirrors the manner in which drivers generally use such systems and the self-regulatory patterns that accompany secondary task engagement. It might reasonably be expected that driver performance and comfort could improve with additional experience and greater self-selection of the points at which they engage with the systems. The extent to which this would impact the relative demand profiles across the interface models and the systems observed here is unknown. On the other hand, compared with other novice users, participants were given an in-depth introduction to the systems, which included guidance on shortcut methods to accomplish the tasks, and participants who were taken on-road practiced with the systems in a parking lot until they indicated they understood how to use them. The extent to which other novice users would attempt to actually use the technologies on-road without similar training and context is unknown.

Throughout the research protocol, multiple instructions presented in written form, recorded audio and verbal reinforcement by a research associate emphasised that participants need not engage in a secondary task if they felt unsafe or if they would not typically engage in the tasks during their personal driving. As previously detailed, two older participants expressed reservations during training about engaging with the tasks while driving and did not go on-road; four additional older participants had sufficient difficulty learning the tasks that the research associate declined to proceed to on-road assessment. No participants who went on-road declined to engage with a task. However, one older participant was unable to recall the training sufficiently or deduce operation of most of the tasks while underway. In the case of two other participants, task presentation was discontinued due to a research associate's concern over the participants' ability to engage with the tasks safely while underway. These cases were relatively equally distributed across the two vehicles and not included in the analysis set, but should be kept in mind in terms of broader usability considerations of the self-reported workload data and other variables presented.

While the data presented here show that interaction with voice interfaces can involve substantial visual engagement, a direct connection to driving safety risk is difficult to establish. The type of data presented here is informative concerning the attentional demand characteristics of the interface tasks, rather than necessarily being predictive of risk to drivers who are operating their own vehicles. Additional naturalistic and/or epidemiological research will be required to evaluate the extent to which interactions with these embedded vehicle systems present any significant elevation in risk.

In the current study, the measures of cognitive demand were not exhaustive, and different measures might provide an alternate perspective. It is also possible that other voice command implementations or interactions with the systems under study (e.g. without awareness of shortcuts) might be associated with greater or lesser overt levels of cognitive or visual demand.

The presentation sequence used for the easy and hard phone tasks could be seen as a methodological limitation. The hard tasks were intentionally presented last to provide participants with maximum exposure to the contact calling interfaces prior to assessing the hard tasks so as to reduce the effect of novelty on the most challenging task. While not considered in detail in the results presented here, several measures suggested that some learning took place over the initial trials of basic phone calling such that, in some instances, a presumed hard task appeared less demanding than the easy task. For example, total task time for manual phone calling in the hard phone task was lower than that observed for the earlier easy trials in the MyLink system.

5. Conclusions

The comparison of manual and voice phone calling with the 2013 Chevrolet Equinox MyLink system and the 2013 Volvo XC60 Sensus system indicates that auditory–vocal interfaces can provide drivers with a means to decrease the time that their eyes are drawn away from the forward roadway when engaging in this type of secondary task. As was found in previous on-road research with actual production systems (Chiang, Brooks, and Weir 2005; Mehler et al. 2014; Reimer et al. 2013; Reimer et al. 2014), the embedded voice interfaces studied here significantly reduced mean single glance time, the percentage of long-duration glances (>2 s) and total off-road glance time relative to embedded manual interfaces.

An important extension in the current study compared with prior research is the comparison of the impact of differing system design approaches. As anticipated based on the task analysis of Reagan and Kidd (2013), the streamlined 'one-shot' approach of MyLink showed a distinct advantage with regard to total task time and several visual demand metrics compared with the layered menu-based approach of Sensus. However, limitations in voice recognition and parsing technology were more apparent with MyLink, where one-shot entry of a full address much more frequently resulted in voice recognition errors by the system, as well as more user input errors and difficulty using the system without assistance. These errors may result in increased workload from frustration associated with repeated engagement or cessation of an engagement in favour of an alternative approach such as visual–manual interaction and use of a smartphone. Similarly, the recent report by Strayer et al. (2014) suggests that significant differences in voice system demand can be observed across vehicles, while Reimer et al. (2014) and Munger et al. (2014) showed in an embedded vehicle system and smartphone that differences in demand can occur based upon system settings.

Taken as a whole, these findings suggest both support for and caution in the development of auditory–vocal interfaces for use by drivers. While a properly designed and used interface can significantly reduce eyes-off-road time, neither of the interfaces studied here eliminated visual demand. Further, overall task duration and visual engagement were quite extensive in the case of destination address entry when compared with the traditional reference of manual radio tuning (NHTSA 2013). The complex relationship between the observed levels of visual engagement and the time involved with voice-based interactions requires further study. It is unclear to what extent risk-based guidance developed for visual demand during manual interactions is directly applicable to voice-based interfaces. This study and previous work (e.g. Mehler et al. 2014; Reimer et al. 2013) suggest that the voice interfaces of current embedded systems are highly multi-modal and the full range of potential demands (auditory, vocal, visual, manipulative, cognitive, tactile, etc.) need to be taken into account. Clearly, evaluations that ignore the complex intertwining of resource demands placed upon the driver paint an incomplete picture of the benefits and limitations associated with various interface design approaches and implementations. Future work needs to further investigate how different interface designs manage the transitions between the auditory–vocal and visual–manual subcomponents of a voice-based activity.

Acknowledgements

Acknowledgement is gratefully extended to Peter Hamscher, Alex Hruska, Martin Lavallière, Alea Mehler, Hale McAnulty, Mauricio Muñoz, Lauren Parikhal, Anthony Pettinato, Adrian Rumpold and Andrew Sipperley for their contributions in protocol development, data collection, reduction and manual scoring. Appreciation is also extended to Adrian Lund and David Zuby for review and comment on the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Insurance Institute for Highway Safety and a grant from United States Department of Transportation's Region One University Transportation Center at MIT.

References

Angell, L., J. Auflick, P. A. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger. 2006. *Driver Workload Metrics Task 2 Final Report & Appendices*. Final Research Report: DOT HS 810 635. Washington, DC: NHTSA/USDOT.

Barón, A., and P. Green. 2006. *Safety and Usability of Speech Interfaces for In-vehicle Tasks While Driving: A Brief Literature Review*. Ann Arbor, MI: The University of Michigan Transportation Research Institute (UMTRI).

Beckers, N., S. Schreiner, P. Bertrand, B. Reimer, B. Mehler, D. Munger, and J. Dobres. 2014. "Comparing the Demands of Destination Entry Using Google Glass and the Samsung Galaxy S4." *Proceedings of the 58th Annual Meeting of the Human Factors and Ergonomics Society*, Chicago, IL, October 27–31.

Blanco, M., W. J. Biever, J. P. Gallagher, and T. A. Dingus. 2006. "The Impact of Secondary Task Cognitive Processing Demand on Driving Performance." *Accident Analysis and Prevention* 38: 895–906.

Brookhuis, K. A., G. de Vries, and D. de Waard. 1991. "The Effects of Mobile Telephoning on Driving Performance." *Accident Analysis and Prevention* 23 (4): 309–316.

Brookhuis, K. A., and D. de Waard. 2001. "Assessment of Drivers' Workload: Performance and Subjective and Physiological Indexes." In *Stress, Workload, and Fatigue*, edited by P. A. Hancock and P. A. Desmond, 321–333. Mahwah, NJ: Lawrence Erlbaum.

Chiang, D. P., A. M. Brooks, and D. H. Weir. 2005. "Comparison of Visual-manual and Voice Interaction with Contemporary Navigation System HMIs (Paper No. 2005-01-0433). *SAE 2005 World Congress & Exhibition*. Warrendale, PA: SAE International.

Collet, C., E. Salvia, and C. Petit-Boulanger. 2014. "Measuring Workload with Electrodermal Activity during Common Braking Actions." *Ergonomics* 57 (6): 886–896.

Conover, W. J., and R. L. Iman. 1981. "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics." *American Statistician* 35 (3): 124–129.

Dopart, C., A. Häggman, C. Thornberry, B. Mehler, J. Dobres, and B. Reimer. 2013. "A Driving Simulation Study Examining Destination Entry with iPhone iOS 5 Google Maps and a Garmin Portable GPS System." *Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society*, San Diego, CA, September 30–October 4.

Driver Focus-telematics Working Group. 2006. *Statement of Principles, Criteria, and Verification Procedures on Driver-interactions with Advanced In-vehicle Information and Communication Systems*. Washington, DC: Alliance of Automobile Manufacturers.

Engström, J., E. Johansson, and J. Östlund. 2005. "Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving." *Transportation Research Part F: Traffic Psychology and Behaviour* 8: 97–120.

Friedman, M. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (200): 675–701.

Fitch, G. M., S. A. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. L. Olson, M. A. Perez, R. J. Hanowski, J. M. Hankey, and T. A. Dingus. 2013. *The Impact of Hand-held and Hands-free Cell Phone Use on Driving Performance and Safety-critical Event Risk*. Report No. DOT HS 811 757. Washington, DC: National Highway Traffic Safety Administration.

Garay-Vega, L., A. K. Pradhan, G. Weinberg, B. Schmidt-Nielsen, B. Harsham, Y. Shen, G. Divekar, M. Romoser, M. Knodler, and D. L. Fisher. 2010. "Evaluation of Different Speech and Touch Interfaces to In-vehicle Music Retrieval Systems." *Accident Analysis and Prevention* 42: 913–920.

Godthelp, H., P. Milgram, and J. Blaauw. 1984. "The Development of a Time-related Measure to Describe Driving Strategy." *Human Factors* 26 (3): 257–268.

Graham, R., and C. Carter. 2000. "Comparison of Speech Input and Manual Control of In-car Devices While on the Move." *Personal Technologies* 4: 155–164.

Green, P. 1994. *Measures and Methods Used to Assess the Safety and Usability of Driver Information Systems*. Report No. UMTRI-93-12./FHWA-RD-94-088.

Green, P. 1999a. "The 15-second rule for driver information systems." In *Proceedings of the ITS America Ninth Annual Meeting*, Washington, DC, CD-ROM.

Green, P. 1999b. "Estimating Compliance with the 15-second Rule for Driver-interface Usability and Safety." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 43, No. 18, 987–991. Sage.

Haigney, D. E., R. G. Taylor, and S. J. Westerman. 2000. "Concurrent Mobile (Cellular) Phone Use and Driving Performance: Task Demand Characteristics and Compensatory Processes." *Transportation Research Part F: Traffic Psychology and Behaviour* 3: 113–121.

Harbluk, J. L., P. C. Burns, M. Lochner, and P. L. Trbovich. 2007. "Using the Lane-change Test (LCT) to Assess Distraction: Tests of Visual-manual and Speech-based Operation of Navigation System Interfaces." *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 16–22. Lowa City, IA: University of Iowa Public Policy Center.

Hart, S. G. 2006. "NASA-task Load Index (NASA-TLX); 20 years Later." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 50, No. 9, 904–908. Sage.

Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, edited by P. A. Hancock and N. Meshkati , 139–183. Amsterdam: North Holland Press.

Hendy, K. C., K. M. Hamilton, and L. N. Landry. 1993. "Measuring Subjective Workload: When is One Scale Better than Many?" *Human Factors* 35 (4): 579–601.

Hornbæk, K., and E. L. C. Law. 2007. "Meta-analysis of Correlations among Usability Measures." In *Proceedings of the ACM SIGCHI Conference on Human Factors*, 617–626. New York: ACM Press.

ISO 15007-1. 2002. *Road Vehicles – Measurement of Driver Visual Behaviour with Respect to Transport Information and Control Systems – Part 1: Definitions and Parameters*. Geneva: International Standards Organization.

ISO 15007-2. 2001. *Road Vehicles – Measurement of Driver Visual Behaviour with Respect to Transport Information and Control Systems – Part 2: Equipment and Procedures*. Geneva: International Standards Organization.

Just, M. A., T. A. Keller, and J. Cynkar. 2008. "A Decrease in Brain Activation Associated with Driving When Listening to Someone Speak." *Brain Research* 1205: 70–80.

Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. 2006. *The Impact of Driver Inattention on Near-crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data* (DOT HS-810-594). Washington, DC: National Highway Traffic Safety Administration.

Kramer, A. E. 1991. "Physiological Metrics of Mental Workload: A Review of Recent Progress." In *Multiple-task Performance*, edited by D. L. Damos, 279–328. London: Taylor & Francis.

Kubose, T. T., K. Bock, G. S. Dell, S. M. Garnsey, A. F. Kramer, and J. Mayhugh. 2006. "The Effects of Speech Production and Speech Comprehension on Simulated Driving Performance." *Applied Cognitive Psychology* 20: 43–63.

Lee, J. D., B. Caven, S. Haake, and T. L. Brown. 2001. "Speech-based Interaction with In-vehicle Computers: The Effect of Speech-based E-Mail on Drivers' Attention to the Roadway." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43 (4): 631–640.

Lenneman, J. K., and R. W. Backs. 2010. "Enhancing Assessment of In-vehicle Technology Attention Demands with Cardiac Measures." In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 20–21. New York: ACM Press.

Liu, Y. C., C. S. Schreiner, and T. A. Dingus. 1999. *Development of Human Factors Guidelines for Advanced Traveler Information Systems (ATIS) and Commercial Vehicle Operations (CVO): Human Factors Evaluation of the Effectiveness of Multi-modality Displays in Advanced Traveler Information Systems* (No. FHWA-RD-96-150).

Lo, V. E.-W., and P.A. Green. 2013. "Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature." *International Journal of Vehicular Technology* 2013, 13 p. doi:10.1155/2013/924170.

Maciej, J., and M. Vollrath. 2009. "Comparison of Manual vs. Speech-based Interaction with In-vehicle Information Systems." *Accident Analysis and Prevention* 41: 924–930.

McDonald, W. A., and E. R. Hoffman. 1980. "Review of Relationships between Steering Wheel Reversal Rate and Driving Task Demand." *Human Factors* 22 (6): 733–739.

Mehler, B., B. Reimer, and J. F. Coughlin. 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: The Journal of the Human Factors and Ergonomics Society* 54 (3): 396–412.

Mehler, B., B. Reimer, J. F. Coughlin, and J. A. Dusek. 2009. "Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers." *Transportation Research Record: Journal of the Transportation Research Board* 2138: 6–12.

Mehler, B., B. Reimer, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and J. F. Coughlin. 2014. *Further Evaluation of the Effects of a Production Level "voice-command" Interface on Driver Behavior: Replication and a Consideration of the Significance of Training Method* (MIT AgeLab Technical Report No. 2014-2). Cambridge, MA: Massachusetts Institute of Technology.

Mulder, L. J. 1992. "Measurement and Analysis Methods of Heart Rate and Respiration for Use in Applied Environments." *Biological Psychology* 34 (2–3): 205–236.

Munger, D., B. Mehler, B. Reimer, J. Dobres, A. Pettinato, B. Pugh, and J. F. Coughlin. 2014. "A Simulation Study Examining Smartphone Destination Entry While Driving." *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicle Applications* (AutoUI 2014), Seattle, WA, September 17–19.

NHTSA (National Highway Traffic Safety Administration). 2013. *Visual-manual NHTSA Driver Distraction Guidelines for In-vehicle Electronic Devices (Docket No. NHTSA-2010-0053)*. Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA).

Nielsen, J., and J. Levy. 1994. "Measuring Usability: Preference vs. Performance." *Communications of the ACM* 37 (4): 66–75.

Noy, Y. I. 1990. *Attention and Performance While Driving with Auxiliary In-vehicle Displays* (Transport Canada Publication TP 10727 (E)). Ottawa: Transport Canada, Traffic Safety Standards and Research, Ergonomics Division.

Östlund, J., L. Nilsson, O. Carsten, N. Merat, S. Jamson, W. Janssen, S. Mouta, J. Carvalhais, J. Santos, V. Anttila, H. Sandberg, D. Luoma, de Waard, K. Brookhuis, E. Johansson, J. Engström, T. Victor, J. Harbluk, W. Janssen, and R. Brouwer. 2004. *Deliverable 2—HMI and Safety-related Driver Performance. Human Machine Interface and the Safety of Traffic in Europe (HASTE) Project*. Report No. GRD1/2000/25361 S12.319626.

Östlund, J., B. Peters, B. Thorslund, J. Engström, G. Markkula, A. Keinath, D. Horst, S. Juch, S. Mattes, and U. Foehl. 2005. *Adaptive Integrated Driver-vehicle Interface (AIDE): Driving Performance Assessment – Methods and Metrics*. Report No. IST-1-507674-IP. Gothenburg: Information Society Technologies (IST) Programme.

Owens, J. M., S. B. McLaughlin, and J. Sudweeks. 2010. "On-road Comparison of Driving Performance Measures When Using Handheld and Voice-control Interfaces for Mobile Phones and Portable Music Players (Paper No. 2010-01-1036). *SAE 2010 World Congress & Exhibition*. Warrendale, PA: SAE International.

Owens, J. M., S. B. McLaughlin, and J. Sudweeks. 2011. "Driver Performance While Text Messaging Using Handheld and In-vehicle Systems." *Accident Analysis & Prevention* 43: 939–947.

Patten, C. J. D., A. Kircher, J. Östlund, and L. Nilsson. 2004. "Using Mobile Telephones: Cognitive Workload and Attention Resource Allocation." *Accident Analysis and Prevention* 36: 341–350.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ranney, T. A., J. L. Harbluk, and Y. I. Ian Noy. 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.

Reagan, I. J., and D. G. Kidd. 2013. "Using Heirarchical Task Analysis to Compare Four Vehicle Manufacturers' Infotainment Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1495–2599, Santa Monica, CA: Human Factors and Ergonomics Society.

Recarte, M. A., and L. M. Nunes. 2000. "Effects of Verbal and Spatial-imagery Tasks on Eye Fixations While Driving." *Journal of Experimental Psychology Applied* 6 (1): 31–43.

Reimer, B., P. Gruevski, and J. F. Coughlin. 2014. *MIT AgeLab Video Annotator*. Cambridge, MA. https://bitbucket.org/agelab/annotator.

Reimer, B., and B. Mehler. 2011. "The Impact of Cognitive Workload on Physiological Arousal in Young Adult Drivers: A Field Study and Simulation Validation." *Ergonomics* 54 (10): 932–942.

Reimer, B., and B. Mehler. 2013. *The Effects of a Production Level "voice-command" Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance* (White Paper 2013-18A). Cambridge, MA: Massachusetts Institute of Technology AgeLab. http://web.mit.edu/reimer/www/pdfs/MIT_AgeLab_White_Paper_2013-18A_(Voice_Interfaces).pdf.

Reimer, B., B. Mehler, J. Dobres, and J. F. Coughlin. 2013. *The Effects of a Production Level "voice-command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance*. MIT AgeLab Technical Report No. 2013-17A. Cambridge, MA: Massachusetts Institute of Technology.

Reimer, B., B. Mehler, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and A. Rumpold. 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 Vehicular Applications* (AutoUI 2014), Seattle, WA, September 17–19.

Reimer, B., B. Mehler, I. Reagan, D. Kidd, and J. Dobres. 2015. "Multimodal Demands of a Smartphone Used to Place Calls and Enter Addresses during Highway Driving Relative to Two Embedded Systems." Manuscript under review. *Ergonomics*.

Reimer, B., B. Mehler, Y. Wang, and J. F. Coughlin. 2012. "A Field Study on the Impact of Variations in Short-term Memory Demands on Drivers' Visual Attention and Driving Performance across Three Age Groups." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54 (3): 454–468.

Roscoe, A. H. 1992. "Assessing Pilot Workload. Why Measure Heart Rate, HRV and Respiration?" *Biological Psychology* 34 (2–3): 259–287.

SAE. 2004. *Navigation and Route Guidance Function Accessibility While Driving (SAE J2364)*. Warrendale, PA: SAE International.

Schreiner, C. S., M. Blanco, and J. Hankey. 2004. "Investigating the Effect of Performing Voice Recognition Tasks on the Detection of Forward and Peripheral Events." *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, Santa Monica, CA, September, 2004, 2354–2358.

Shutko, J., K. Mayer, E. Laansoo, and L. Tijerina. 2009. "Driver Workload Effects of Cell Phone, Music Player, and Text Messaging Tasks with the Ford SYNC Voice Interface versus Handheld Visual-manual Interfaces." (Paper No. 2009-01-0786). *SAE 2009 World Congress & Exhibition*. Warrendale, PA: SAE International.

Smith, D. L., J. Chang, R. Glassco, J. Foley, and D. Cohen. 2005. "Methodology for Capturing Driver Eye Glance Behavior During In-vehicle Secondary Tasks." *Transportation Research Record: Journal of the Transportation Research Board* 1937 (1): 61–65.

Solovey, E., M. Zeck, E. A. Garcia-Perez, B. Reimer, and B. Mehler. 2014. "Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies." *Proceedings of the 32nd Annual Conference on Human Factors in Computing Systems (CHI 2014)*, Toronto, Canada, April 26–May 1, 4057–4066.

Strayer, D. L., J. M. Cooper, J. Turrill, J. Coleman, N. Medeiros-Ward, and F. Biondi. 2013. *Measuring Cognitive Distraction in the Automobile*. Washington, DC: AAA Foundation for Traffic Safety.

Strayer, D. L., F. A. Drews, and W. A. Johnston. 2003. "Cell Phone-induced failures of Visual Attention during Simulated Driving."

Journal of Experimental Psychology: Applied 9 (1): 23–32.

Strayer, D. L., J. Turrill, J. R. Coleman, E. V. Ortiz, and J. M. Cooper. 2014. *Measuring Cognitive Distraction in the Automobile II: Assessing In-vehicle Voice-based Interactive Technologies*. Washington, DC: AAA Foundation for Traffic Safety.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. 1996. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use." *European Heart Journal* 17: 354–381.

Tsimhoni, O., D. Smith, and P. Green. 2004. "Address Entry While Driving: Speech Recognition versus a Touch-screen Keyboard." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (4): 600–610.

Veltman, J. A., and A. W. K. Gaillard. 1998. "Physiological Workload Reactions to Increasing Levels of Task Difficulty." *Ergonomics* 41 (5): 656–669.

Victor, T., J. Bärgman, C. Boda, M. Dozza, J. Engström, C. Flannagan, J. D. Lee, and G. Markkula. 2014. *Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk*. (SHRP 2 Safety Project S08A Prepublication Draft). Gothenburg: Safer Vehicle and Traffic Safety Centre at Chalmers.

Wilson, G. F. 2002. "Psychophysiological Test Methods and Procedures." In *Handbook of Human Factors Testing and Evaluation*, edited by S. G. Charlton and T. G. O'Brien, 127–156. Mahwah, NJ: Lawrence Erlbaum.

Appendix

Table A1. Means (and standard errors) by phone calling task and embedded vehicle system (Chevrolet MyLink or Volvo Sensus). SD = standard deviation; kph = km/hr.

	Vehicle	Phone easy (manual)	Phone hard (manual)	Phone easy (voice)	Phone hard (voice)
Self-reported workload	Chevrolet	4.28 (0.4)	5.20 (0.4)	1.81 (0.3)	1.90 (0.3)
	Volvo	5.49 (0.4)	6.12 (0.4)	2.05 (0.2)	2.55 (0.3)
Task completion time	Chevrolet	29.18 (2.0)	23.30 (0.9)	20.48 (1.3)	22.74 (1.8)
	Volvo	31.43 (2.0)	34.36 (2.6)	34.48 (1.3)	41.87 (1.6)
% change in heart rate	Chevrolet	2.54 (0.9)	1.12 (0.8)	2.46 (0.8)	3.75 (0.8)
	Volvo	2.07 (1.0)	2.10 (0.7)	2.47 (0.7)	0.97 (0.6)
% change in SCL	Chevrolet	15.06 (3.8)	13.15 (3.0)	13.66 (3.3)	12.22 (3.2)
	Volvo	13.09 (3.0)	12.75 (2.9)	7.62 (2.8)	3.63 (2.7)
Mean off-road glance duration	Chevrolet	0.89 (0.0)	0.94 (0.0)	0.60 (0.0)	0.61 (0.0)
	Volvo	0.94 (0.0)	0.95 (0.0)	0.79 (0.0)	0.79 (0.0)
Mean glance to device duration	Chevrolet	0.91 (0.0)	0.97 (0.0)	0.37 (0.0)	0.37 (0.1)
	Volvo	0.97 (0.0)	0.98 (0.0)	0.85 (0.0)	0.86 (0.0)
% of off-road glances > 2.0 s	Chevrolet	1.73 (0.6)	3.12 (0.8)	0.09 (0.1)	0.49 (0.5)
	Volvo	2.98 (1.0)	3.80 (1.0)	0.94 (0.4)	0.57 (0.2)
% of glances to device > 2.0 s	Chevrolet	1.77 (0.6)	3.40 (0.9)	0.05 (0.1)	1.46 (1.5)
	Volvo	3.33 (1.2)	4.12 (1.1)	1.13 (0.6)	0.90 (0.4)
Total off-road glance time	Chevrolet	15.16 (1.2)	11.97 (0.7)	3.42 (0.5)	3.23 (0.4)
	Volvo	15.95 (1.1)	16.82 (1.4)	9.78 (0.7)	10.65 (0.9)
Total to device glance time	Chevrolet	14.41 (1.2)	11.44 (0.6)	1.61 (0.4)	1.26 (0.2)
	Volvo	15.35 (1.1)	15.99 (1.4)	7.53 (0.6)	7.99 (0.8)
Number off-road glances	Chevrolet	16.74 (1.1)	12.82 (0.6)	5.26 (0.7)	5.05 (0.6)
	Volvo	16.96 (1.0)	17.70 (1.3)	12.44 (1.0)	13.46 (1.0)
Number glances to device	Chevrolet	15.39 (1.0)	11.79 (0.5)	2.19 (0.4)	1.76 (0.3)
	Volvo	15.89 (1.0)	16.24 (1.2)	8.84 (0.8)	9.31 (0.8)
Speed (CAN – kph)	Chevrolet	108.86 (2.0)	107.82 (2.4)	111.71 (1.4)	112.89 (1.5)
	Volvo	105.06 (1.2)	105.83 (1.6)	107.53 (1.1)	106.72 (1.3)
Speed (GPS – KPH)	Chevrolet	105.90 (2.0)	104.83 (2.3)	108.65 (1.4)	109.71 (1.5)
	Volvo	107.45 (1.3)	109.58 (1.0)	110.05 (1.1)	109.02 (1.4)
% change in speed (CAN)	Chevrolet	−3.77 (1.6)	−4.71 (1.9)	−1.19 (1.0)	−0.17 (1.1)
	Volvo	−3.94 (0.9)	−3.30 (1.3)	−1.70 (0.7)	−2.37 (1.1)
% change speed (GPS)	Chevrolet	−3.64 (1.6)	−4.63 (1.9)	−1.09 (1.0)	−0.13 (1.1)
	Volvo	−4.03 (0.9)	−2.14 (0.6)	−1.72 (0.7)	−2.56 (1.1)
SD of speed (GPS - kph)	Chevrolet	2.99 (0.3)	2.58 (0.2)	1.93 (0.2)	1.95 (0.2)
	Volvo	3.04 (0.5)	2.63 (0.2)	2.85 (0.2)	2.87 (0.2)
% change in SD of speed (GPS)	Chevrolet	−33.48 (5.6)	−41.58(4.3)	−54.76 (4.3)	−53.90 (3.6)
	Volvo	−27.36 (9.4)	−36.24 (3.8)	−30.57 (4.2)	−22.80 (6.3)
Major wheel reversals per minute	Chevrolet	25.65 (1.9)	26.79 (1.8)	26.37 (1.8)	23.22 (1.9)
	Volvo	4.78 (0.4)	4.10 (0.5)	3.74 (0.4)	4.05 (0.3)
% change in major wheel reversals	Chevrolet	23.26 (8.4)	27.58 (7.2)	28.63 (9.2)	16.80 (10.5)
	Volvo	51.41 (13.4)	38.11 (16.0)	31.31 (17.0)	37.84 (14.8)

Note: All time metrics in Tables are in seconds unless otherwise indicated.

Table A2. Summary of ANOVA by ranks on the phone tasks for variables of vehicle, modality, and vehicle × modality. SD = standard deviation; kph = km/hr.

Variable	Modality	System	Modality × system
Self-reported workload	***	NS	NS
Task completion time	NS	***	***
% change in heart rate	NS	NS	+
% change in SCL	+	NS	NS
Mean off-road glance duration	***	**	***
Mean glance to device duration	***	***	***
% of off-road glances > 2.0 s	***	NS	NS
% of glances to device > 2.0 s	***	NS	NS
Total off-road glance time	***	***	***
Total glance to device time	***	***	***
Number of off-road glances	***	***	***
Number of glances to device	***	***	***
Speed (CAN)	**	***	NS
Speed (GPS)	**	NS	NS
% change in speed (CAN)	**	NS	NS
%change in speed (GPS)	**	NS	NS
SD of speed (GPS)	*	**	***
% change in SD of speed (GPS)	*	***	***
Major wheel reversals per minute	NS	***	NS
% change in major wheel reversals	NS	NS	NS

Notes: + = borderline effect ($p < 0.10$); NS = not significant.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table A3 Means (and standard errors) and results of Wilcoxon signed rank tests for variable measured during the destination address entry task periods. SD = standard deviation; kph = km/hr.

	Chevrolet	Volvo	*W*	*p*-value
Self-reported workload	3.59 (0.44)	2.54 (0.28)	924.5	0.154
Task completion time	66.68 (2.85)	80.60 (1.71)	408	<0.001*
% change in heart rate	1.66 (0.87)	1.25 (0.67)	801	0.996
% change in SCL	11.59 (3.77)	3.29 (2.44)	811	0.172
Mean off-road glance duration	0.74 (0.02)	0.82 (0.02)	562	0.022*
Mean glance to device duration	0.81 (0.04)	0.91 (0.03)	575	0.03*
% of off-road glances > 2.0 s	1.02 (0.29)	1.27 (0.36)	777.5	0.813
% of glances to device > 2.0 s	1.48 (0.48)	1.83 (0.54)	747.5	0.569
Total off-road glance time	14.28 (1.22)	22.56 (1.43)	367	<0.001*
Total glance to device time	8.25 (0.89)	15.80 (1.16)	305	<0.001*
Number of off-road glances	18.65 (1.52)	27.77 (1.75)	397	<0.001*
Number of to device glances	9.16 (0.98)	17.43 (1.25)	286.5	<0.001*
Speed (CAN – kph)	113.68 (1.06)	108.26 (0.88)	1186	<0.001*
Speed (GPS – kph)	110.47 (1.04)	110.85 (0.92)	797	0.981
% change in speed (CAN)	0.57 (0.62)	−1.01 (0.54)	997	0.058
% change in speed (GPS)	0.60 (0.62)	−0.98 (0.55)	990	0.068
SD of speed (GPS - kph)	3.24 (0.18)	3.85 (0.25)	649	0.148
% change in SD of speed (GPS)	−29.53 (4.58)	−10.35 (5.46)	550	0.016*
Major wheel reversals per minute	21.78 (1.22)	2.88 (0.23)	1598	<0.001*
% change in major wheel reversals	7.34 (6.10)	−8.32 (6.82)	1003	0.051

Change scores represent the per cent (%) change from baseline just driving. Where presented, per cent change values are likely to provide a more accurate representation of relative change for a particular variable as discussed in the body of the paper.
*$p < 0.05$.