Effects of an ‘Expert Mode’ Voice Command System on Task Performance, Glance Behavior & Driver Physiology

Bryan Reimer, Bruce Mehler, Jonathan Dobres, Hale McAnulty, Alea Mehler, Daniel Munger & Adrian Rumpold
MIT AgeLab & New England University Transportation Center
77 Massachusetts Avenue, E40-291 - Cambridge, MA 02139
reimer@mit.edu bmehler@mit.edu jdobres@mit.edu mcanulth@mit.edu amehler@mit.edu dmunger@mit.edu rumpold@mit.edu

ABSTRACT
Multi-function in-vehicle interfaces are an increasingly common feature in automobiles. Over the past several years, these interfaces have taken on an ever-greater number of functions and the ways in which drivers interact with information have become more complex. Parallel with these technical developments, interest in ensuring that these systems minimize demand placed upon the driver has also increased. Voice command capability has become a popular and desirable feature, as interacting with a vehicle interface through auditory/vocal interactions is often hypothesized to allow the driver to keep their eyes on the road and hands on the wheel. However, research has shown that production level voice command systems may still impart considerable visual demands on the driver [18]. These demands might be due in part to screen displays associated with extensive confirmatory dialogue and the driver’s desire for visual confirmation that commands were accurately recognized. This study extends this work by comparing the default mode of a production voice system with an “Expert” mode which streamlines tasks by removing several confirmatory steps. We found that, although the use of the Expert mode significantly reduces overall task completion time, it has no appreciable effect on the amount of visual engagement; drivers still glance off the road for durations that are consistent with the Default mode. Implications for interface design and driver safety are discussed.

Author Keywords

ACM Classification Keywords
H.1.2 [User/Machine Systems]: Human Factors; Human Information Processing; H.5.2 [User Interfaces]: Ergonomics; Interaction styles; User-Centered Design; Voice I/O

INTRODUCTION
Drivers continue to demand increased connectivity and more advanced entertainment options while driving. While automotive manufacturers and their suppliers strive to provide drivers with convenient, safe, easy-to-access information to meet this growing demand, there remains no well-established method for optimally achieving this goal. Over the past several years, there has been a shift in automotive driver-vehicle interfaces (DVIs) from purely visual-manual interactions to include options for voice-based or voice-assisted interaction. Voice enabled systems have become a popular feature in many production vehicles as well as on smartphones. It is thought that interacting with a technology in a vehicle through an auditory/vocal modality can occur while the driver’s eyes are oriented towards the road and his/her hands are on the wheel, thus minimizing the need to direct visual attention to the in-vehicle device and physical manipulation of buttons, touch screens or other controllers.

Research has been directed for some time at developing an understanding of and assessing the safety, usability, and demand related aspects of voice-interaction in the vehicle [see 2, 6, 18 for a review]. Much of the literature compares voice systems with handheld visual-manual based tasks (texting and phone dialing), voice-based smartphone applications, “Wizard-of-Oz” simulations, and a range of after-market and research-based voice systems in which there is a wide variation in the amount of detail provided on the functional characteristics of the systems under evaluation. To date, only a small number of studies have addressed the demands of production level voice command interfaces where the operating characteristics of the system can be documented and are available for exact replication in other research. In addition, the overall performance of a system widely available at the consumer level is expected, on average, to be smoother than one developed for
In the broadest field assessments of a production level automotive voice command system to date [10, 18], the cognitive demands placed on the driver, as measured by physiological arousal and self-report, were largely comparable to a visual-manual radio tuning task and below a 1-back level audio-vocal cognitive benchmark task [9, 15]. However, the amount of visual demand, as measured by total off-road glance time, for voice controlled activities such as navigation destination entry significantly exceeded that of the visual-manual radio tuning task, as well as the thresholds set for the National Highway Traffic Safety Administration’s [11; NHTSA] Phase I visual-manual distraction guidelines if they were applied. While significant visual demands were observed with some of the voice command tasks, the popular notion that voice interfaces should reduce visual demand and enable a driver to keep his or her eyes on the road as in-vehicle tasks are performed was supported in a direct comparison of visual-manual radio tuning and an equivalent task completed using the voice command system [10, 18]. In contrast, a comparison of preset selection completed with a traditional visual-manual button press versus the voice command system showed more visual demand in the multi-step voice enabled version of the task. The reports conclude that voice enabled in-vehicle interactions, as implemented in the production level system assessed, are best considered as multi-modal.

In our experience with other production automotive voice systems to date, it seems likely that few, if any, voice-enabled DVI implementations presently involve solely auditory-vocal interaction. At minimum, most current voice-based in-vehicle systems are multi-modal in nature, drawing upon a combination of auditory, vocal, cognitive, manipulative, and visual resources that all need to be considered and optimized to minimize distraction and support driver focus.

The observation of visual demand associated with voice command interactions is not novel. While earlier reports [4, 19] highlighted the advantages of voice command systems, they also present data that illustrates substantive engagement of visual attention associated with their use. In reporting on two small field driving studies, Chiang and colleagues [4] conclude that voice interface based activities resulted in lower off-road visual demands compared to visual manual interfaces, but present findings showing elevated levels over single task driving. Shutko et al. [19] conclude, “As a hands-on-wheel, eyes-on-road system, SYNC provides a hands-free voice interface to safely control cell phones and personal music players while driving.” While the aforementioned conclusions clearly favor the voice system over the hand-held device interactions, a close inspection of the study’s data (see Figure 2 in [19]) suggests that the tasks under investigation would not have met NHTSA’s new guidelines for visual-manual distraction [11], specifically that 85% of the sample should be able to complete a task with less than 12 seconds of total off-road glance time. It does appear that these tasks met the original Alliance of Automotive Manufacturers [5] guideline (criterion 2.1 A) of under 20 seconds of total glance time to the device.

In our initial work [18], it became clear that the default settings for the voice interface under study incorporated a number of confirmatory steps in its dialog model (e.g., “Is this correct?”) with the expectation of a yes/no response. The system also supported an “experienced” or “Expert” mode that eliminated most confirmatory steps and some additional prompting, resulting in a reduction in the amount of listening time required for most tasks. It seemed reasonable to ask to what extent the previously observed visual and other demands might be attributed to the extensive confirmatory dialogue and associated visual displays. In the present study, we investigate what effects, if any, the differences in the Default and Expert mode interaction models have on driver glance behavior, physiology, and vehicle control. Following our previously established methods, equivalent “easy” and “hard” radio tuning tasks were performed to support a direct comparison of the demand between the visual-manual and voice enabled DVIs to complete the same task goal.

METHODS

Participants

Recruitment drew from the greater Boston area. A total of 64 participants are included in this analysis, split equally between the “Default” and “Expert” modes and equally balanced by gender and across four age groups 20-24, 25-39, 40-54, and 55-69 in conformity with NHTSA guidelines on age distribution in the testing of visual-manual distraction associated with DVI use [11] (see Table 1). Participation required presentation of a valid driver’s license, attesting to having had a license for more than three years, driving on average three or more times per week, not having been involved as a driver in a police reported crash in the past year, being in reasonably good health for their age, and not having specified medical conditions or taking medications that might impair safe operation of the vehicle. Compensation was $75.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>40.1 (17.1)</td>
<td>40.1 (16.1)</td>
</tr>
<tr>
<td>Default</td>
<td>39.9 (15.1)</td>
<td>38.9 (15.2)</td>
</tr>
</tbody>
</table>

Table 1: Mean (and standard deviation) of ages for each interface mode and gender.

Apparatus

The study was conducted in a 2010 Lincoln MKS with factory installed voice-command based systems (Ford SYNC™ for voice control of the phone and media connected by USB and the “next-generation navigation system” with Sirius Travel Link). This vehicle was owned by MIT, instrumented for on-road research, and represented...
an example of a widely available production level voice interface when this project was initiated in 2011. The voice interface is engaged using a “push-to-talk” button on the right side of the steering wheel (see Figure 1). When the voice control interface is active, a display screen in the center stack typically supplies supporting information on system status and often provides information on prompts that the driver may use in dialog with the system.

![Figure 1. Interior of the test vehicle. Note the Push-to-Talk button on the right side of the steering wheel that is used to initiate interaction with the voice-command system and the center stack display screen.](image)

The vehicle was instrumented with a customized data acquisition system for time synchronized recording of vehicle information from the controller area network (CAN) bus, a MEDAC System/3 physiology monitoring unit, an Iteris AutoVue® Lane Departure Warning System for assessing lane position, and GPS tracking. CAN bus and lane position data were captured at 10Hz, GPS data at 1Hz, and physiological data at 250Hz to support EKG feature extraction for accurate hear beat interval detection. A set of 6 in-vehicle cameras recorded driver behavior and the surrounding roadway environment. A research associate was seated in the rear of the vehicle and was responsible for ensuring safe vehicle operation.

Subjective workload ratings were obtained using a global rating per task on a scale consisting of 21 equally spaced dots oriented horizontally along a 10cm line. The numbers 0 through 10 were equally spaced below the dots and bookended with the labels “Low” and “High” on the left and right, respectively. Participants were told that 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. Participants were instructed to circle a point along each scale that best corresponds to how much workload they felt was involved in trying to do each task.

**Tasks**

Three in-vehicle task types were assessed: traditional visual-manual control of the radio, voice-command based control of the radio, and voice-command based navigation destination entry and route cancelation. As noted earlier, half of the participants performed the tasks using the voice command system's Default mode and the other half used an Expert mode that removed most of the confirmation steps and some prompts.

An “easy” radio task consisted of tuning the radio by selecting a preset station. With the physical preset buttons utilized, the visual-manual version of this task could be performed in a single step. In contrast, the corresponding voice-command workflow in the Default mode involved 3 steps (1 voice button press and 2 verbal inputs / confirmations, i.e. “preset-1”, “yes”). In the Expert mode, the confirmatory “yes” / “no” response was eliminated, reducing the number of steps involved to two.

The “hard” radio tuning task was modeled on specifics of the NHTSA draft guidelines document [12] provided for the radio tuning reference task for assessing visual-manual DVIs. This task consisted of 3 visual-manual steps (selecting the radio function, selecting the radio band on the touch screen, and then tuning to a specified station by rotating a manual tuning knob). The corresponding Default mode voice-command based interaction also involved 3 steps (1 voice button press and 2 verbal inputs / confirmations, i.e. “100.7”, “yes”). As with the easy voice-based radio task, in the Expert mode the confirmatory “yes” / “no” response was eliminated, reducing this to a two-step activity.

The voice-based interface was also used to input a full destination address into the vehicle’s navigation system. Participants initiated the task after hearing the audio prompt: “Your task is to enter the destination address: 177 Massachusetts Avenue, Cambridge, Massachusetts”. To reduce memory requirements, the address was also provided on a cue card located in the center of the steering wheel. An illustration of the 12 steps (2 voice button presses and 10 verbal inputs / confirmations) that represent the shortest possible dialog for destination entry in the Default mode are detailed below. Italics denote system auditory prompts; expected user vocal input are shown in parenthesis [see 10, 18].

**[Driver presses voice button]**

*Please Say a Command.*

*Destination Street Address“*

*Destination Street Address, is that correct?“*

“Yes“*

*Destination Street Address, please say the city name. “Cambridge“*

*Cambridge, is that correct?“*

“Yes“*

*Cambridge, Please say the street name. “Massachusetts Avenue“*

*Massachusetts Avenue, is that correct?“*

“Yes“*

*Massachusetts Avenue, please say a house number.*
The number you said was 1-7-7, is that correct?
“Yes”
Yes. Press the voice button to say set as destination or back.

[Driver presses voice button]
“Set as destination.”
Set as destination, is that correct?
“Yes”
Set as destination.

As noted earlier, the system included an Expert mode that reduced the amount of audio prompting, resulting in a reduction of the amount of audio content listening time required of the driver. This mode also eliminated 4 of the confirmation steps during address entry. Optimal utilization of the Expert mode for destination entry therefore required a minimum of 8 steps (2 voice button presses and 6 verbal inputs / confirmations); these are detailed below.

[Driver presses voice button]
“Destination Street Address”
Destination Street Address, please say the city name.
“Cambridge”
Cambridge, please say the street name.
“Massachusetts Avenue.”
Massachusetts Avenue, Please say a house number.
“1-7-7”
The number you said was 1-7-7, is that correct?
“Yes”
Yes. Press the voice button to say set as destination or back.

[Driver presses voice button]
“Set as destination.”
Set as destination.

Each destination entry was followed by a cancellation task, in which the driver used the voice system to cancel the route that he/she had just requested. In both the Default and Expert modes this task required 3 steps (1 button press and 2 verbal inputs / confirmations). The command structure in the Expert version of this activity required the same confirmatory “yes”, even though the driver was not prompted for a confirmation. It is unknown whether this was an intentional design feature or an oversight.

Tasks were always conducted twice during periods of highway driving, and task order was counterbalanced across participants. Finally, participants were asked to perform four levels of a delayed digit recall n-back task, e.g. blank back, 0-back, 1-back and 2-back [8, 9] the results of which are not addressed in this paper.

Procedure
Following informed consent and a structured interview to confirm eligibility, physiological sensors were attached, and participants were then escorted to and given an orientation to the research vehicle. The participant was instructed to adjust seat and mirrors and asked to back the vehicle out and then back into the parking space and additional adjustments made as desired. An introduction to the voice command system was provided that included going through the individual voice calibration option to maximize the capacity of the system to correctly recognize commands from each participant. While parked, subjects were provided a structured training on each task that they were expected to perform in the mode of operation that they were assigned. The training included detailed instructions on each activity as well as a minimum of two practice trials of each task.

The driving portion of the study was conducted on roadways in the greater Boston area and divided into three segments. First, a period of approximately 10 minutes of urban driving to reach interstate highway I-93, continuing north on I-93 for an additional 20 minutes to the I-495 intersection. This allowed a total driving adaptation period of approximately 30 minutes prior to the assessment portion of the study. Second, a period driving south on I-495 until the participant was able to complete the first two task periods and then reach a convenient exit point to reverse direction; this portion averaged approximately 30 minutes. Third, the turn-around point to drive north on I-495 to I-93, during which the remaining primary task periods were completed (see Figure 2).

![Figure 2. Experimental Flow.](image-url)

Measures & Data Analysis
Physiological measures, heart rate and skin conductance level (SCL), are reported as percent change values from a 10 minute period of baseline driving drawn from means of 5 two minute periods spread across the experiment. The percent change calculation normalizes these measures to address high individual variability in base physiological values that could impact a between-subject design. Task completion time was calculated as the time between the end of the task prompt and successful task completion or failure. Errors in task performance were coded according to the taxonomy detailed in [18 Appendix G] and summarized as per the legend in Figure 3.

Glance data were manually coded based on video from an in-vehicle camera directed at the driver’s face according to the taxonomy and procedures outlined in [18 Appendix F]. The coding process involved the use of custom software [16] that was developed (and subsequently open sourced) to allow for rapid frame-by-frame annotation of videos. Although this method lacks the assumed granularity of an automated eye tracking solution, it allows us to robustly
differentiate between gross glance targets (i.e., road, instrument cluster, center stack, left window, right window, etc.) and ensures that nearly all task periods can be coded with a high degree of accuracy. Each task period of interest was coded twice, by two different members of the research staff. Any discrepancies between the two coders—the identification of conflicting glance targets, missed glances, or glance timings that differed by more than 200ms—were mediated by a third staff member [see 20 for arguments on the need for two or more independent raters]. The resulting dataset allowed for the calculation of several measures of glance behavior: mean off-road glance time, percentage of glances greater than 2s, total off-road glance time, number of off-road glances, and percent of time glancing off-road.

Data from the vehicle’s CAN bus was used to compute several measures of driving behavior, including the number of accelerations in excess of 0.1g [17], mean and standard deviation of vehicle velocity, and wheel reversals (angular gap of either 3° for major or 0.1° for minor events) and the standard deviation of wheel position [for details on the computation of wheel reversals see 13].

Statistical analyses were performed in R [14]. Owing to non-normality of the data and/or the use of ratio data (percentages) for several dependent measures, in most 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 selected analyses, repeated-measures ANOVA by ranks are presented.

RESULTS

![Figure 3. Task performance by participant.](image)

Figure 3 illustrates performance differences between participants in the Default and Expert mode groups. Perhaps the most overt difference between the Default group and the Expert group is the decrease in single-error task performances during the navigation cancel task (9 in Expert mode vs. 17 in Default). While the navigation cancel task is only a 3-step activity, its high error rate illustrates that participants appear to have difficulty with this task. The errors appear to be linked to a difficulty remembering the command “Navigation Cancel Route”. A similar study [18] in which participants were provided an explicit prompt on the command prior to the task (i.e. “Your task is to cancel the route using the command ‘Navigation Cancel Route’”) show nearly flawless performance on this activity. The overall success rate for the 8-12 step navigation entry task is particularly notable, as is the observation that task performance improves for almost all tasks during the second trial.

Mean values for dependent measures were computed across voice-based and visual-manual based task periods for the two groups (Default and Expert modes). While no main effects of the interface mode were observed for the two radio manual tasks (Wilcoxon test for main effect of mode, all p-values > .05), a main effect of interface mode did appear for several metrics during voice tasks. A summary of those statistical tests is presented in Table 2. From a control perspective, drivers’ completion of tasks using the Expert mode was associated with fewer micro-accelerations (0.89 vs. 0.96 accelerations/min), as well as lower variability of velocity (SD of 2.50 km/hr vs. 2.93 km/hr) and variability of wheel angle (1.38º vs. 1.53º).

<table>
<thead>
<tr>
<th>Measure</th>
<th>V</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of Accelerations &gt; 0.1g</td>
<td>365</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Mean Velocity</td>
<td>503</td>
<td>p = 0.91</td>
</tr>
<tr>
<td>Variability of Wheel Position</td>
<td>693</td>
<td>p = 0.01</td>
</tr>
<tr>
<td>Variability of Velocity</td>
<td>666</td>
<td>p = 0.04</td>
</tr>
<tr>
<td>Count of Large Wheel Reversals</td>
<td>562</td>
<td>p = 0.51</td>
</tr>
<tr>
<td>Count of Small Wheel Reversals</td>
<td>569</td>
<td>p = 0.45</td>
</tr>
<tr>
<td>Percent of Off-Road Glances &gt; 2s</td>
<td>480.5</td>
<td>p = 0.66</td>
</tr>
<tr>
<td>Mean Off-Road Glance Duration</td>
<td>516</td>
<td>p = 0.96</td>
</tr>
<tr>
<td>Number of Off-Road Glances</td>
<td>494</td>
<td>p = 0.81</td>
</tr>
<tr>
<td>Total Off-Road Glance Time</td>
<td>513</td>
<td>p = 0.99</td>
</tr>
<tr>
<td>Task Completion Time</td>
<td>814</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Percent of Off-Road Glance Time</td>
<td>293</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Percent Change in Heart Rate</td>
<td>574</td>
<td>p = 0.41</td>
</tr>
<tr>
<td>Percent Change in SCL</td>
<td>437</td>
<td>p = 0.75</td>
</tr>
<tr>
<td>Self-Reported Workload Rating</td>
<td>638.5</td>
<td>p = 0.09</td>
</tr>
</tbody>
</table>

Table 2. Wilcoxon signed-rank tests for a main effect of interface mode among the voice command based interactions. Significant differences are highlighted in bold.

As expected, given that the amount of prompting and listening time required of the driver is reduced, mean task completion time for the two replications of the four voice command tasks in Expert group was 22.8% lower (5.4 vs. 7.0 minutes). As illustrated in Figure 4, the lower task completion time was consistent across all four activities.
(pairwise p-values all < .05). At a task level, this corresponded to 26.1%, 34.5%, 16.8% and 28.0% lower mean task completion times for the radio easy, radio hard, navigation entry and navigation cancel tasks respectively.

Figure 4. Mean task completion time for each task and interface where (M) indicates visual-manual and (V) indicates voice command. Error bars represent one mean-adjusted standard error.

Interestingly, while Expert mode significantly reduces total task completion time, it does not affect the total amount of time spent glancing off-road (13.3s vs. 13.5s). Figure 5, illustrates a high degree of similarity between the mean total off-road glance times observed in the two groups. With the lower task completion time, but no corresponding reduction in glance time off the road, this results in a higher percentage value for time spent glancing off-road during voice command tasks for the Expert mode group (29.6% vs. 21.5%). As illustrated in Figure 6, there is a consistent nominal difference in the percentage of time spent glancing off-road for all voice tasks. However, pairwise differences by mode only reaches statistical significance for the radio easy, radio hard and navigation cancel tasks completed through the voice command interface (p-values < .05).

Figure 6. Percent off-road glance time for each task and interface. Labeling as in Figure 1.

Although it is often assumed that increased demand on the driver is less safe, the relationship between the overt demands of an activity and safety is hard to quantify. Nonetheless, a complete picture of the allocation of visual attention should include long duration glances, which are often considered to be the most risky [7]. Interestingly, as illustrated in Figure 7, the voice command tasks all consistently show lower percentages of glances greater than 2 seconds away from the road than the visual-manual radio hard (NHTSA benchmark) tuning task. While the 85th
percentile for the two groups’ radio tuning tasks (blue line segments in Figure 7) appears to suggest a nominal divergence between the groups, these rankings represent a subtle difference involving one more participant in the Default group failing to complete the visual-manual radio hard tuning task with 85% or more of the glances shorter than 2 seconds. Not surprisingly, this suggests that NHTSA’s proposed visual-manual guidelines are sensitive to slight variations in a participant sample studied under field testing conditions with two task replications.

A number of differences appear across various measures among the visual-manual and voice-command tasks. A full discussion of these differences is beyond the scope of this paper. However, key differences related to the measures detailed in Figures 4-6 deserve acknowledgement. Concerning Figure 4, a significant difference in task completion time appears across the different interface tasks ($X^2(5) = 244.5, p < .001$). The effect appears to be driven by voice-command navigation entry. Mean completion time was 102.7 seconds (4 times longer than visual-manual radio hard tuning) and the voice-command navigation cancel task averages 40.6 seconds (1.6 times visual-manual radio hard tuning). Navigation entry and navigation cancel took significantly longer to complete than the hard radio manual task (both $p < .001$, Wilcoxon test). A main effect of task type also appears for total off-road glance time ($X^2(5) = 238.6, p < .001$). As illustrated in Figure 5, and consistent with total task completion time, the effect is in part driven by navigation entry, which drew on average 30.8 seconds (SD 21.7s) of attention away from the roadway. The mean off-road glance time for the voice-command navigation entry is on the order of 2.7 times greater than the attentional demand of the visual-manual radio hard tuning task. While a main effect of task type appears for the percentage off-road glance time ($X^2(5) = 169.1, p < .001$), in contrast to the measures discussed earlier, the effect is driven by the visual-manual radio hard tuning task. All of the voice-command tasks draw about half as much attention away from the road per unit time.

**DISCUSSION**

The Expert mode effectively reduces total task completion time through a streamlining of the task workflow. However, the reduced workflow and reduction of visual support screens associated with confirmatory response steps does not substantially reduce the total eyes off-road time (visual demand). It appears that simple yes/no response interactions are not pulling glances off the roadway as is the case with other aspects of the DVI. In other words, the shorter Expert mode does not provide the overt advantage in the visual domain that might have been expected. While not likely a negative, one consequence of considering the Expert mode is an increase in the relative percentage of time spent glancing off-road during an activity.

The strength of the observed differences between the Default and Expert mode groups are supported through the consistency of observations between the groups on the two visual-manual tasks not affected by the mode of interaction. The consistency between measures on the radio easy and radio hard visual-manual tasks illustrate a degree of equivalence between the groups, suggesting that the effects observed during the voice-command based interface tasks are predominately based upon the operating mode, and not an artifact of the sample.

It is difficult, and beyond the scope of this experimental work, to assess whether either the Default or Expert mode places a safety-relevant level of demand on the driver. Observed reductions in micro-accelerations and the variability of velocity suggest that Expert mode was associated with smoother vehicular control. It is not clear if these changes are suggestive of slightly lighter demands on the driver, i.e. improved control, or a by-product of gaze concentration [17], an effect that would be more illustrative of increased demand. What is clear, however, is that a complete understanding of the demands of these two interaction modes will require a more detailed understanding of how task completion time, total off-road glance time, the percentage of off-road glance time, and other factors come together to impact the overall task demands and moment-to-moment demands placed upon the driver. While it has been argued for some time that task time is an important factor [3], one must consider how shortening the task time in a voice command based interface influences the ability of a driver to effectively allocate off-road glances across an activity. This work illustrates how the temporal characteristics of voice command based interactions are fundamentally different from those of “classic” visual-manual interactions. Percent off-road glance time may in some ways be a useful and parsimonious metric, as it normalizes glance time against total task time. Percentage of off-road glance time for all voice-command activities fell well below that of the visual-manual radio tuning (NHTSA benchmark) tuning task.

As detailed in [10, 18], the results presented here illustrate that voice command based interactions are not purely a cognitively oriented interaction or “an eyes free way of communicating with a vehicle interface”. The interface under study and others of a similar class are best described as auditory-vocal-visual-manual mixed-mode interfaces. The optimization of demand across the four modalities impacted by interactions with these types of interfaces will require a deeper understanding of how drivers balance the input and output of information across demands, clearly a complex undertaking that will require new research. With other facets of voice command based interfaces, such as the ability to interrupt the task flow, long-term use of the system, and error recovery still sparsely studied, the overall demands of these multi-modal systems are not yet well understood. It remains an open question of how previous studies, which have primarily focused on auditory-vocal (cognitive) and traditional visual-manual interactions, are applicable to these newer technologies.
In terms of the total eyes off-road time metric (Figure 5), a number of results markedly exceeded NHTSA’s 12-second criterion, if it were applied. Alternatively, if The Alliance’s 20-second threshold and the glance-to-device metric are employed, the manual radio tuning reference task and Navigation cancel task would likely meet this criterion if participants were given additional on-road practice trials. However, the Navigation entry task would likely remain over either threshold. As suggested for the single long glance duration metric, it is possible that providing drivers with additional on-road practice trials might well have resulted in our sample meeting The Alliance guideline. On the other hand, it is questionable whether additional practice for this sample, or consideration of “error-free” cases would have produced on-road performance within the total eyes off-road time metric NHTSA has established for simulation testing. This assessment was carried out on-road as opposed to under the simulation conditions specified by NHTSA, and the 12-second criterion is meant to reflect a boundary defined by the actual demands of the manual radio tuning reference task. The results here do suggest that the 12-second criterion may be difficult to meet for many of the auditory-voice-visual-manual interactions (mixed mode task) that would also fall under the auspice of the NHTSA guidelines [1]. Clearly, some thought needs to be given to a more holistic approach to assessing the benefits and limitations of multi-modal interfaces.

LIMITATIONS
This work excluded individuals who showed overt difficulty communicating with the voice system and is based upon a relatively large sample of 64 to minimize the impact of outliers. The analysis did not attempt to control for cases where participants deviated from the optimal command structures or failed to adequately complete selected activities, as would be required when assessing “error free” engagements per NHTSA’s proposed guidelines [11]. An analysis comparing “error free” tasks is undertaken in [18] Appendix B and may provide some reference as to how these results may be interpreted in that context. Further, this assessment is limited to a single production level system. Based upon a review of voice command systems with similar stepwise command structures, it is likely that the results presented here do have a reasonable level of generalizability among this class of voice command interface. However, the specific system studied may not be fully representative of current technologies. Therefore, additional research will be needed to verify the extent of concurrence. The incorporation of glances to cue cards utilized to provide drivers with addresses for the navigation entry task (an attempt to reduce cognitive memory requirements) into off-road glances likely artificially elevates reported levels of visual demand somewhat. The degree to which this impacts these results is unknown. The results should also be considered in light of the highway operating environment utilized and the experimental nature of the data collection. Drivers in situ would be expected to frequently self-regulate interactions with respect to the demands of the operating environment and such interactions may differ from the context of the experimentally presented tasks. How this field data correlates with findings that would be obtained under the simulation procedures specified in the Phase I visual-manual guidelines is unknown. Finally, the demands observed were not clearly linked to overt safety issues, and while conceptual concerns about increased demand can be raised and are worthy of study, they should be interpreted cautiously.

ACKNOWLEDGMENTS
Acknowledgement is extended to The Santos Family Foundation and US DOT’s Region I New England University Transportation Center at MIT for providing support for the exploration of the expert mode of the SYNC system and to Toyota’s CSRC for funding for the study of the default system [see 10 for details] that was conducted in parallel with the evacuation of the expert mode.

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


