Contextual Push-to-Talk: Shortening Voice Dialogs To Improve Driving Performance

Garrett Weinberg, Bret Harsham, Cliff Forlines
Mitsubishi Electric Research Labs
201 Broadway
Cambridge, MA 02139, U.S.A.
+1.617.621.7500
{weinberg, harsham}@merl.com,
forlines@alumni.cmu.edu

Zeljko Medenica
University of New Hampshire
Department of Computer Science
Durham, NH 03824, U.S.A.
+1.603.862.3778
zeljko.medenica@unh.edu

ABSTRACT
We present a driving simulator-based evaluation of a new technique for simplifying in-vehicle device interactions and thereby improving driver safety. We show that the use of multiple, contextually linked push-to-talk buttons (Multi-PTT) shortens voice dialog duration versus the use of a conventional, single push-to-talk button (Single-PTT). This benefit comes without detriment to driving performance or visual attention to the forward roadway. Test subjects also preferred the Multi-PTT approach over the conventional approach, and reported that it imposed a lower cognitive workload.

Categories and Subject Descriptors
H5.2. [**Information Interfaces and Presentation**]: User Interfaces – *Voice I/O; Input devices and strategies (e.g., mouse, touchscreen)*

General Terms
Design, Experimentation, Human Factors

Keywords
Speech recognition, voice dialogs, listen button, push-to-talk, multimodality, driving simulation.

1. INTRODUCTION
Awareness of the dangers of distracted driving appears finally to be permeating the popular press and the public consciousness. The U.S. state of Utah recently passed a law that punishes drivers caught sending text messages just as harshly as drunk drivers [20]. In Gwent, Wales, the police department collaborated on a gory dramatization of the risks of texting that has garnered over 1.9 million views on YouTube [25]. The U.S. Department of Transportation has announced new rules to prohibit commercial truck and bus drivers from texting while driving [9].

Today's "connected car," however, contains more than a mobile phone. Both factory-installed and aftermarket devices used in the car boast a dizzying assortment of mapping and multimedia capabilities whose use cases and potential to cause distraction may differ significantly from basic mobile telephony and messaging.

In terms of driving safety, Voice User Interfaces (VUIs) have shown promise as compared to manual alternatives (buttons, touchscreens, etc.) for tasks ranging from address entry to music retrieval [16, 8, 11]. When drivers operate in-vehicle information systems (IVIS) partially or totally by voice rather than manually, they tend to control their vehicles more carefully, react more quickly to hazards, and exhibit more situational awareness [5].

Despite the indications that voice and multimodal UIs offer potential safety benefits over manual-only UIs, there has been relatively little published work comparing differing VUI implementations *against one another* in terms of suitability for vehicular use. Eager to bring products to market quickly, manufacturers often base their choices in grammar, voice prompt, and dialog design on their best guesses of what *ought* to be suitable for in-car use, rather than what has been empirically proven suitable.

1.1 VUI Issues and Approaches
With the wide array of local and remote connectivity options and rich content offered by a contemporary IVIS, a VUI designed to cover the majority of features can become just as confusing, multi-tiered and complex as a traditional, hierarchical GUI menu. It would seem ideal if one could utter any given natural-language command or free-form search for any type of content from any IVIS application state. Such a design would rid the driver of the need to maintain a mental model of the system's state and available command set. However, the accuracy limitations and CPU constraints of current-generation embedded automatic speech recognition (ASR) engines stand in the way of this design.

Furthermore, it would seem ideal if one could utter a command at any time, perhaps interrupting the stream of conversation with a passenger to address the system. An "always-listening" voice-recognition system (as explained in e.g., [1]) would enable such an interaction by monitoring the recognized speech signal for a wake-up word such as "computer." Drawbacks to this approach include the privacy implications of a constantly active microphone, the relative difficulty spotting the wake-up keyword in high-noise vehicular environments, and the high cost (in user confusion) of misrecognizing a word as the wake-up word. Therefore all current in-car ASR implementations are of the "sometimes-listening" variety: a push-to-talk (PTT) switch is mounted in the car, usually on the steering wheel. Activation of the PTT switch causes the system to interpret subsequent speech by the user as intended system input.

Manufacturers have a variety of techniques for coping with the multiplicity of functions and the diversity of available data addressable by voice. In most modern systems, some voice commands are always available, but others are only active in

MobileHCI'10, September 7–10, 2010, Lisbon, Portugal.
ACM 978-1-60558-835-3/10/09

specific application states. For example, if the current screen depicts traffic conditions or an address book entry and one wants to listen to the album "Abbey Road," one must first use the always-available "music" command, followed by, for example, "search albums" and then "Abbey Road."[1] Whether a press of the PTT button is required before every utterance or only before the first in a sequence of utterances, there will always be at least a short pause after each utterance to allow for recognition and, typically, for the system to audibly confirm the command. This back-and-forth between user and system is called a dialog. Recently, considerable effort has been made to reduce excessive mode-switching and keep the system's state tree as shallow as possible, thanks to emerging research indicating that deep and/or complex dialog design leads to driver inattention and steering errors [12].

This paper concerns a new technique to reduce the duration of voice dialogs through a novel Push-to-Talk technique. We will briefly cover literature relevant to PTT buttons and in-vehicle interfaces. We then present our Multi-PTT approach, along with the results of a usability and suitability evaluation conducted in a driving simulator. We then give a summary of our findings and conclude with a discussion of future work.

2. RELATED WORK

There is ample literature documenting the detrimental effects of in-vehicle devices upon drivers' tactical scanning and lane keeping behavior, especially as the duration of interaction with the device increases. Besides numerous studies on mobile phone dialing and conversation (e.g., [2, 6, 24]), Salvucci showed that selecting songs from an iPod while driving resulted in significantly increased lateral deviation from lane center versus unencumbered driving [21].

As for voice interfaces in particular and their effects on driving performance, Barón and Green note in their widely-cited 2006 survey [5] that voice interfaces to IVIS tend to impact driving performance less than comparable manual-only alternatives, and that they allow for significantly more eyes-on-road time. However they and others note that voice-enabled IVIS are still unacceptably distracting. Their dialog design often encourages long task durations, and when they feature a visual component, users tend to glance down at inopportune times [17] and for longer than dictated by the "2-second rule" of defensive driving [4].

Recent progress has been made in bringing products to the marketplace with shorter dialogs and a comprehensible, consistent command grammar. The Ford Sync offering [9] has been critically and commercially well received in part because nearly every content item can be retrieved in two dialog turns (depending on whether a mode change is required). Honda and IBM have collaborated on a deployment that uses a language model-based recognizer of the sort previously seen only in desktop dictation engines [23]. Though this latter technology does not directly address dialog depth or duration, it allows for a high degree of flexibility in the phrasing of commands and has the potential to reduce user confusion and lower cognitive load.

Other approaches seek to shorten or eliminate voice dialogs by recasting in-vehicle interactions as search tasks rather than browsing or menu-traversal tasks. Divi et al. [8] discuss a paradigm called Speech-In List-Out (SILO) whereby results to a potentially "fuzzy" query are presented in relevance order. They limited their interface to a single domain (a collection of MP3 files), meaning at minimum a button press or spoken mode-change command might be necessary to activate that domain before the advantages of SILO can take effect. Graf et al. extend this paradigm into multiple content domains [13]. They illustrate an effective prototype IVIS that supports cross-domain querying and result presentation, although their input scheme was based on touchpad character-recognition rather than ASR.

There has been less research on the PTT affordance itself, or whether novel PTT controls or sensors can positively impact driving or secondary task performance. Palinko and Kun have instrumented certain surfaces of the steering wheel with sensor strips that expect double-taps as a PTT gesture [18], and found no ill effects on driving. The same authors have also incorporated PTT actuators into a wireless glove as a stand-in for an entirely touch-sensitive steering wheel, and found that subjects glanced down at the wheel less when using this implementation versus a traditional, fixed PTT button.

3. CONTEXTUAL PUSH-TO-TALK

Our design (presented in conceptual form in [27]) proceeds from the realization that voice input need not be an afterthought when considering the physical human-machine interface (HMI) design. If the car or portable device in question is designed to have dedicated buttons for choosing screens or modes, can these buttons somehow be dual-purposed as voice input buttons? Instead of having a unique, single-purpose PTT button, we envision that *any* button or physical control can be a "listen" control when activated in a certain way.

In such a design, the quick press of a mode button might switch to the mode in question, for example Navigation, Music or Contacts. A longer press or a double-press of the button could indicate the user's wish to not only change to the mode, but also to immediately carry out a voice search in the mode; the paradigm in this case is "change to the mode, and find what I say."

Command—rather than mode—controls can also be extended with a voice activation style. For example, the omnipresent green "phone" button might, with an ordinary single-press actuation, bring up a Recent Calls screen. With a double-press actuation, it might cause the system to listen for voice input, in this case the phonebook entry that should immediately be dialed (e.g., "John Doe mobile").

Similarly, "play/pause" and "shuffle" buttons could accept voice modifiers. If the normal actuation acts as a simple toggle (play or pause, random playback on or off), the voice-enabled actuation would listen for the *target* of the operation (play *what*, shuffle *what*).

The multifunction, contextual PTT controls may be placed on the steering wheel or elsewhere (e.g., in the center console). In either case, careful attention must be paid to their placement and their tactile design (e.g., separation distances, surface ridges, button travel and actuation feedback) in order to maximize motor learning and thereby allow the accustomed driver to keep her eyes on the road while comfortably operating the controls.

Whether this multiple-PTT paradigm is applied to command or mode buttons (or indeed other physical controls), its chief advantage is the elimination of one or more turns in a multi-turn voice dialog. In conventional approaches that use a single PTT

[1] In a more advanced implementation, the latter two commands might be combined into "find album Abbey Road."

button, the initial turn or turns are used to extract contextual information—for example information about the search domain of interest—which is then used to activate an ASR grammar appropriate to the next dialog turn. In our Multi-PTT approach, the same contextual information is conveyed *by the user's choice of button*, allowing the system to skip directly to the later dialog turn (see Table 1).

Table 1. Example interaction: Searching for the song "Lucky Star" by Madonna.

Single-PTT		Multi-PTT	
User	**System**	**User**	**System**
<press main PTT button>	<beep>	<press Music-linked PTT button>	*"Searching Music"* <beep>
"Music"	*"Music"*	"Madonna Lucky Star"	<plays the desired song>
"Search"	*"Searching Music"* <beep>	<done>	
"Madonna Lucky Star"	<plays the desired song>		
<done>			

The conventional and contextual-PTT approaches can be fruitfully combined in a single system implementation. Novice users might access a given mode or function via a traditional, multi-turn dialog initiated by pressing a dedicated PTT button. Advanced users would have the additional affordance of using other buttons *besides* the main PTT button to initiate voice input, when they are activated in a special manner. Users may discover this by exploring the interface or by reading documentation. Having gained this knowledge, they are empowered to bypass dialog turns and carry out their tasks more quickly.

4. EXPERIMENT

In comparing a contextual-PTT interaction with a conventional PTT interaction, it initially seems that there should be a natural advantage to using contextual PTT, because it reduces the number of dialog turns required to perform a task. However, in order to initiate a contextual-PTT dialog, the user must first locate and press the correct PTT button, and must use the special button actuation style reserved for the voice modality rather than just pressing the button "normally." Therefore using contextual PTT could potentially distract the driver: she must visually target the button, move one hand off the wheel, and remember the actuation procedure.

In order to evaluate the usability and safety of contextual PTT compared to conventional PTT, we built two separate versions of a prototype IVIS. One version, Single-PTT, used solely the conventional style of PTT interaction. The other version, Multi-PTT, used solely contextual PTT interactions.

Our hypotheses were:

1. Multi-PTT interactions result in shorter task times.
2. Multi-PTT interactions are not more distracting to a driver than Single-PTT interactions.

In order to compare the Multi-PTT and Single-PTT interfaces, we conducted a laboratory study in which participants completed tasks with an in-car system using each technique while operating a driving simulator. Our driving simulator [28] consists of a motion-base cockpit chair facing three displays that render the simulation (see Figure 1). For this study the experimenter sat at a table behind and to the right of the subject, out of her field of view.

4.1 Simulator Hardware

A Windows PC drives three 127-cm (50-inch) DLP rear-projection displays offering a combined resolution of 3072 x 768. A cockpit-style chair [7] offers vibration and two-axis tilt that are coordinated with acceleration, braking and steering/cornering. Bolted to the chair's frame is a Logitech G25 force-feedback wheel and pedal set that affords primary operational input. The simulated engine noise and the sounds/music generated by the IVIS are played through a 5.1-channel speaker system also affixed to the chair. Finally, the simulator includes a commercial eye-tracking system [22] to allow for the analysis of gaze direction and duration.

The left and right displays are angled toward the driver by 20°. With the driving seat slid fully back (as it is for the tallest subjects), this display configuration results in a vertical viewing angle of 30° and a horizontal angle of 107°. With the seat fully forward, the vertical viewing angle is 32° and the horizontal viewing angle is 113°.

4.2 Simulator Software

The commercial driving game rFactor [15] serves as the software platform for our driving simulator. It offers a convincing, realistic driving experience thanks to richly detailed graphics, accurate vehicle physics, and full support of force-feedback steering wheels such as the Logitech G25. It furthermore supports extensive modification and customization of transmission, suspension, and handling settings, which were useful to increase the fidelity of the simulation for the study of standard highway driving (as opposed to race car driving). Most importantly for our purposes, rFactor provides a plug-in API whereby vehicle telemetry (including position, velocity, and acceleration), and user input (steering angle and throttle/brake positions) can be captured at rates up to 90 Hz. For this study we used a gently curving highway course depicting a fictional coastline. This course is available as a free, downloadable add-on to rFactor [19].

Figure 1. Driving Simulator.

4.3 Prototype IVIS

Our prototype IVIS is called SpeakPod. Its visual display is similar in size and placement to built-in navigation system

displays in currently available vehicles. The screen is an LCD measuring 18cm (7in) diagonally. Mounted immediately to the left of the screen is a vertically-oriented Optimus Mini Three keypad [3]. This keypad offers three 20mm x 20mm buttons, each of which houses a full-color 96 x 96 OLED screen. In our experiment, these in-key displays showed graphics corresponding to the three main SpeakPod domains (Navigation, Music, and Contacts), and pressing one of these buttons activated the domain corresponding to the graphic (domain screens shared a thematic color with their corresponding button graphic). The keypad was positioned approximately 13cm (5in) to the right of the steering wheel, with the screen immediately to its right, and both were angled to provide optimal visibility for drivers of average height. Their position and angle was fixed and identical for all study subjects.

The SpeakPod application's main window (see Figure 2) is a vertically-oriented list featuring a selection bar whose position is manipulated by custom-made wireless input device approximately 7cm x 3.5cm x 4cm in size that was attached to the simulator's steering wheel in an unobtrusive position.

At the device's center is a "jog dial" widget that offers unlimited bidirectional rotation and an actuator. This widget controls selection and activation. The button above the jog dial moves up one level in the hierarchy. The lower right button pauses or resumes music playback, and the lower left button activates voice input, if available in that experimental condition (press to talk, release at any time; a short tone indicates the system is listening). The metal switch on the left face of the device is a power toggle.

The font used in the main vertical list is Arial Narrow Bold 30 Point, with an anti-aliasing effect applied to enhance legibility. There are up to seven menu items visible in at a time at the preferred 640 x 480 operating resolution. The window's title bar features centered text in the same font face.

Figure 2. SpeakPod IVIS Setup.

The initial state of the main list depicts items corresponding to each domain (Navigation, Music, and Contacts), and each domain, in turn, has a root menu depicting submenus relevant to that domain (e.g., Artists, Albums and Songs for the Music domain). Each of these submenus offers its own relevant submenus, and so on, in a hierarchical tree structure. The depth and content of this tree is designed to make the number of dialog turns and the duration of the overall retrieval task comparable with current commercial IVIS systems. Activation of a leaf item in the music domain (i.e., an individual track) brings up a Now Playing screen, and activation of leaf items in the Navigation and Contacts domains bring up a point of interest (POI) detail screen and a contact detail screen, respectively.

4.3.1 Speech Interface

The system includes a speech interface which incorporates a noise-robust production ASR engine along with post-processing logic for the management of voice dialogs and searches.

4.3.2 Content Indexing and Searching

In addition to browsing via the hierarchical menus described above, users may find items using short spoken queries [29]. At load time, item metadata is indexed in a domain-dependent way, with the aim of allowing the user to find a given item by saying any combination of relevant words from any combination of indexed fields. For Music-domain items, these fields comprise artist name, album name, and track title. For Navigation-domain items, we used POI name, POI genre (e.g., "hotel"), and town or city name. For Contact-domain items, the fields indexed were first name, last name, and group name (e.g., "colleagues"). The Music database consisted of 103 artists, 86 albums, and 785 songs drawn from a cross-section of popular music. The Navigation database consisted of 4655 POI from the surrounding area. Lastly, the Contacts database consisted of 500 fictional individuals whose names and photographs were harvested and randomized from available Internet data sources. Fictional group names were arbitrarily assigned to 69 of the contacts. The use of a fixed set of retrieval items ensured that all users were equally unfamiliar with the database contents, and that the size of the ASR grammars and vocabulary were the same for all users.

4.3.3 Search Behavior Variants

As mentioned above, we decided to compare conventional PTT with contextual PTT in isolation by creating two separate variants of a prototype IVIS: one whose search feature relies on a conventional, single PTT button, and another where search is carried out using the multiple, domain-linked PTT buttons. Each variant *requires* the use of the jog-dial controller to traverse hierarchies and activate found items, and each *allows* the use of the 3-button OLED keypad to activate domains' root screens.

4.3.3.1 Single-PTT Search

To carry out any interaction with the Single-PTT variant, the user must first press the listen button (the lower left button on the steering wheel-mounted input device). The system then plays a short tone indicating that the microphone is open, at which point the user speaks a command. Each application screen has its own set of screen-dependent commands (displayed in yellow) and a set of always-available commands, which are not explicitly shown on the screen, but are rather implied. These commands consisted of "main menu," "back," "help," "search" and names of each of the domains ("navigation," "music," and "contacts").

The scope of the "search" function is constrained to the currently displayed screen. In other words, search will be performed only within the domain or submenu that is currently being displayed. For example, if a search is executed from the main screen of a particular domain (i.e., the Navigation, Music or Contacts root screen), the search will be performed among all the items in the domain. If the user selects a submenu by either a voice or tactile means before issuing the search command, then the search is constrained to that submenu (for example the Albums submenu within Music). If search is initiated while the result list from a

previous search is showing, a refinement search will be carried out.

Upon recognition of a selection or traversal command (e.g., "back"), the system issues an audible confirmation via text-to-speech. Upon recognition of a "search" command, the system issues an audible reminder of the active search scope (e.g., "searching albums") and then immediately plays the listening tone and enables the microphone.

Each result screen is a numbered list where any item can be selected by either using the jog-dial device or by issuing a voice command consisting of the number of the desired entry.

As a whole, the Single-PTT variant was designed to resemble a contemporary dialog-based IVIS as closely as possible.

4.3.3.2 Multi-PTT Search

In the MultiPTT variant, the PTT button on the jog-dial controller was inactive. Instead, users carried out searches by double-pressing a given key on the Optimus Mini Three keypad. The system would then transition instantly to the state corresponding to a whole-domain search for the chosen domain. For example, if the user double-pressed the Contacts button, the system prompted "searching contacts" and then issued the listening tone and opened the microphone. There was no affordance for sub-domain searches (e.g., searches for POI categories within the navigation domain, or for artists within the music domain).

4.3.3.3 Varying Characteristics

The chief differences between the variants were the addition of the double-tap affordance on the Optimus OLED buttons in the Multi-PTT variant, and the requirement to issue an explicit "search" command in the Single-PTT variant. In addition, the Single-PTT variant afforded sub-domain searches (e.g., Albums within the Music domain, or Categories within the Navigation domain), which, if utilized, could shorten task time for items poorly recognized in global searches (the Multi-PTT variant only supported the latter). In both variants, subjects could use the Optimus buttons to change mode (single-tap), and had to use the steering wheel controller to select from lists.

4.4 Design

We chose a within-subject, repeated measures usability study with *interaction-technique* (interface variant) and *repetition* as independent variables, and a number of dependent variables that measured not only the performance of the user interface (the secondary task), but also the interface's effects on driving performance (the primary task). In order to measure the performance of the user interface, we recorded task time, task errors, and subjective preference. With respect to the interfaces' effects on driving performance, we measured following distance, lateral deviation, driving speed, throttle depression, steering angle, and the number of glances away from the forward roadway. In addition we measured total subjective workload using the NASA-TLX survey [14].

In short, our design was:

18 participants x
2 conditions (Single-PTT, Multi-PTT) x
12 interactions =
432 trials in total

4.5 Protocol

Subjects were recruited from local colleges and universities. Subjects were required to be licensed drivers, to be native speakers of North American English, and to not wear glasses while driving (although contact lenses were acceptable). The second requirement was introduced to reduce variations in the number of ASR errors between subjects, while the third requirement was due to the limitations of our eye-tracking system. In total, eighteen subjects, six female and twelve male (age M=21.75, SD=4.53), participated in this study. Each experimental session was about one hour and thirty minutes long. The subjects were compensated $40 for their participation.

The experiment consisted of three drives: control, Multi-PTT, and Single-PTT. The control drive consisted of driving the simulator without any IVIS interactions. The Single- and Multi-PTT drives incorporated interactions with their respective user interfaces as well. Each drive lasted for ten minutes, and all subjects completed all three drives. All drives took place on the same simulated roadway, a gently curving coastal highway. To combat learning effect, we counterbalanced the order of presentation among subjects. The main task in all three drives was to follow a pace vehicle and to maintain a constant distance behind it, even if the pace vehicle slowed down unpredictably (which it in fact did at certain places on the course). Subjects were instructed that following the vehicle and driving safely had the highest priority, while all other tasks (i.e., listening to the experimenter's prompts and operating the IVIS) had secondary importance. They were also encouraged to delay any interaction with the IVIS if they found it to be too distracting from the main task of driving.

Before starting the experiment itself, subjects had a five minute training period to get accustomed to the driving simulator. Similarly, before each condition, subjects were trained to use the IVIS variant under test (Single-PTT or Multi-PTT) until they became comfortable using it.

During the Single- and Multi-PTT drives, the experimenter prompted subjects to find various randomly-chosen items from the three domains: Music, Navigation, and Contacts. Each retrieval task was considered to be a separate trial for the purposes of our analysis. The time between adjacent tasks was not less than ten seconds. A task was considered to be successfully completed when the sought item was found and activated. In the case of Contacts and Navigation, this entailed opening the person or POI details screen, and in the case of Music, this entailed opening or beginning playback of the sought artist, album, or song. Prompts to the subjects were worded similarly to [21] (e.g., "Please find the album X by Y" or "Please find the contact John Doe").

The subjects had a maximum of 60 seconds to complete a retrieval task. If time elapsed before successful completion, the task was marked as unsuccessful. In addition, as a means of reducing user frustration in cases of poor speech recognition performance, we instructed subjects that they could at any point declare the task to be a failure. In either of these cases, the user was given 10 seconds of recovery time, and then the experimenter began the next task.

Following the completion of each drive, subjects were instructed to complete the NASA-TLX survey considering the driving task alone or the combination of driving and user interface tasks, as appropriate. At the end of the entire experiment, a user interface preference questionnaire was administered.

4.6 Data analysis

Our driving performance metrics (Following Distance, Lateral Position, Speed, Steering Angle, Throttle) were logged by the simulator software in real-time. Eye-tracking measurements were modeled after those discussed in [6]. We divided eyes-off-the-

road glances into short (< 0.5 seconds), medium (0.5 – 2 seconds), and long (> 2 seconds) bins as in [4].

In terms of usability analysis, our primary metrics are retrieval task completion (success/failure) and retrieval task duration. Although Walker expresses doubts about the suitability of duration as a measure of voice dialog usability [26], we feel it is the most reasonable and practical metric for automotive contexts because of the time-critical nature of the primary driving task.

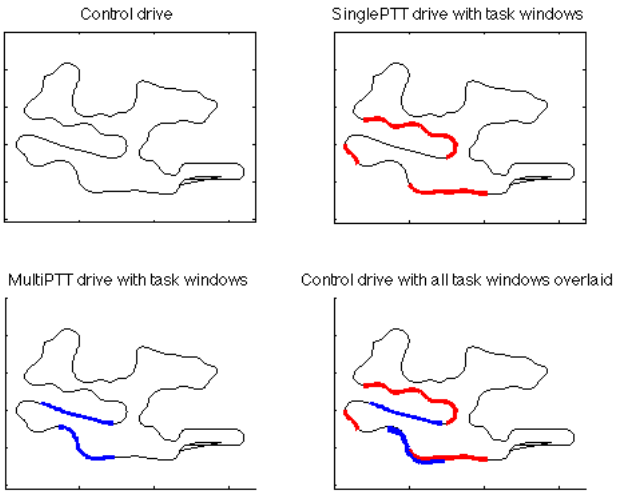

Figure 3. To account for differences in the course from task to task, driving performance during the Single-PTT and Multi-PTT conditions was compared to the relevant section of the course from the control condition.

4.6.1 Window-based Task Analysis

As described above, subjects were asked to begin a new retrieval task approximately 10 seconds after completion of (or failure to complete) the previous task. Task durations were variable (ranging from 5 to 60 seconds). Because of this, tasks were performed at different locations along the route by each user, and in each condition. We call the segment of the course during which a task was performed the "task window."

Our analysis depends on the fact that the route for each drive was identical, even though task windows varied by task. In general, we would expect that in the absence of a distraction, driving performance would be a function of the road (curvature, visibility, etc) and the behavior of the lead vehicle (which the subject was instructed to follow at a constant distance). In this experiment, the lead vehicle was under computer control and behaved identically in every trial (it did not react to changes in the subject's following distance). The short break between tasks gave the subjects time to return to a comfortable following distance. Therefore, we can evaluate the effect of performing a given retrieval task by comparing the driving performance in that task's window to the driving performance over the same section of route during the control drive. Because the route and lead vehicle behavior are identical, any changes in driving performance will be largely due to using the IVIS.

Figure 3 shows an example set of task windows for a subject (the length of task windows is not realistic; it is exaggerated for clarity). Because the subject performed no IVIS tasks during the control drive, there are no task windows shown in the upper left portion of the graphic. In this example, the subject performed three tasks during the Single-PTT drive (upper right). Each corresponding task window is shown as a red overlay on the segment of the route along which it took place (note that the task windows have variable length). During the Multi-PTT drive (lower left graphic), the same example subject performed two tasks. The task windows corresponding to these tasks are shown with blue overlays. The lower right graphic combines the overlays to emphasize that task windows for the same subject across two conditions may either overlap or be completely distinct.

4.6.2 Calculation of Driving Metrics

For each interaction technique (Single-PTT and Multi-PTT), we computed the technique's effect on driving by examining how each subject's driving behavior during content retrieval differed from her driving behavior during the control drive. This gave us two sets of performance data (one for each interaction technique) that characterized the effect of the interaction technique on driving.

In more detail, we found each retrieval task's window (start and endpoints of its corresponding road segment), then calculated the variance in each driving metric (Following Distance, Lateral Position, Speed, Steering Angle, Throttle) over that window. We calculated the corresponding variance in each driving metric over the same road segment for the control drive. For each of these metrics, a higher variance generally represents more erratic driving [6]. In other words: if the IVIS task is distracting, we expect that higher variances in one or more metrics will be observed over the task drive as compared to the control drive for the same road segment. We computed the difference in variance for each metric between the content-retrieval task drive and the control drive. We used this variance difference as our repeated measurement in analyzing the overall behavior within a subject's session.

For example, suppose that task 10 took place on curvy road section X during the Single-PTT drive. We took the start and endpoint of that task (in track coordinates) and calculated a given driving metric (say, variance in lane position) for road section X during both the control drive and the Single-PTT drive. Then we use the delta:

lane variance(SinglePTT[X]) – lane variance(control[X])

as our repeated measure during analysis. The same approach applies to each individual task, on its particular road section—curvy or straight, uphill or downhill.[2]

We consider this task windowing technique to be a contribution in itself. The alternative is to start all tasks at predefined points on the track. When using predefined start points, the distance between start points must be large enough that all subjects can finish a given task prior to reaching the next task start point. The predefined start point method leads to long inter-task downtimes and fewer tasks performed during a session. Our window-based task analysis increased the number of trials that could be performed during a subject's session.

4.7 IVIS Usability Results

In this section, we report results that relate to the usability and performance of the two in-car interfaces.

[2] It should be noted that the items sought during the two content-retrieval conditions were not the same (they were randomly selected at the start of each condition from among all possible items).

4.7.1 ASR Accuracy

The large-vocabulary search grammars were identical for the two IVIS variants. The Single-PTT variant's command grammar was artificially small, offering tens of commands rather than hundreds as found in real contemporary IVIS. Because of this, there were only one or two misrecognized commands in the course of the entire study, across all subjects and tasks.

In other words, command recognition errors did not hinder subjects in advancing through the Single-PTT dialog to reach the large-vocabulary search states. The contextual push-to-talk buttons in the Multi-PTT were shortcuts directly to these search states. Therefore the effective ASR accuracy in the two variants was the same (zero + large-vocabulary search errors), so we can assume ASR errors affected neither IVIS variant more than the other. We therefore omit further analysis of this topic.

4.7.2 Task Time

Task time was defined as the amount of time between the end of the experimenter's task prompt (e.g., "please find the song Lucky Star by Madonna") and the moment when either the participant selected the correct item in the UI, or the task was declared to be a failure. In our analysis, we used a repeated-measures within-participant ANOVA with *interaction-technique* and *repetition* as independent variables. To check for asymmetrical learning effects among the interaction techniques, we also used the *order-of-presentation* of the techniques as a between-participant variable. *Order-of-presentation* had no significant effect on task time ($F_{5,12}$ = 0.53, p = 0.75), or on any other measurement used in this analysis; thus, it is safe to continue with a within-participant design.

We found a significant difference between the two *interaction-techniques* in terms of *task-time* ($F_{1,17}$ = 154.51, p < 0.001), with mean times of 26.5s and 15.8s for Single-PTT and Multi-PTT respectively. Figure 4 shows the mean task times for each *interaction-technique*.

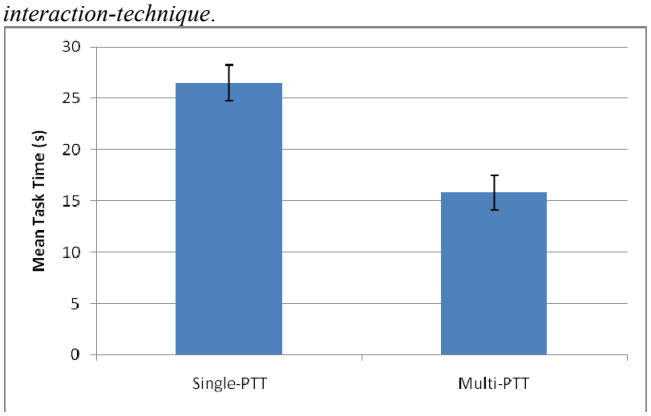

Figure 4. Mean task times for each *interaction-technique*. Error bars represent 95% confidence intervals in all figures.

4.7.3 Task Errors

The Multi-PTT *interaction-technique* resulted in significantly fewer errors (i.e., unsuccessful tasks) than Single-PTT ($F_{1,17}$ = 4.62, p = 0.046). The two *interaction-techniques* had overall error rates of 8.8% and 4.2% for Single-PTT and Multi-PTT respectively. Figure 5 shows the mean error rates for each *interaction-technique*.

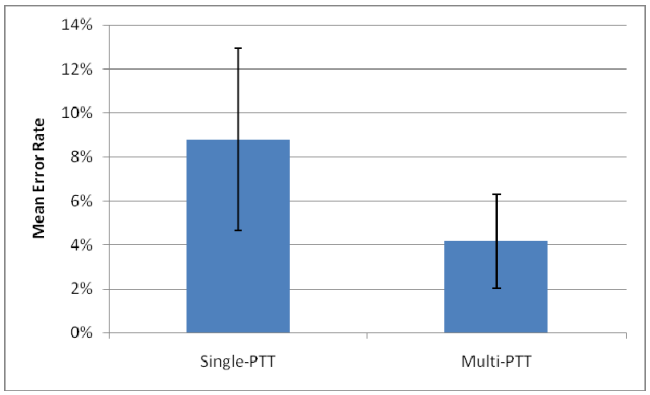

Figure 5. Mean error rates for each *interaction-technique*.

4.7.4 Search Preference

At the end of the study, each participant completed a subjective preference questionnaire that included two questions concerning the *interaction-techniques* themselves. Participants rated their agreement/disagreement with a number of questions on a 5-point Likert scale. Table 2 shows the number of participants who agreed more strongly with the statements on the ease-of-use and desirability of the two techniques, as well as the test statistics for a Wilcoxon Signed Rank Test. As a whole, participants agreed significantly more with the statement "I would use the ___ interface" when applied to the Multi-PTT technique than the Single-PTT technique.

Table 2. The number of participants who agreed more strongly with the preferential statements for each of the two interfaces and the corresponding Wilcoxon Signed Ranks Test statistics. (note: totals do not sum to 18 because some participants agreed equally with both statements)

Statement	Single-PTT	Multi-PTT	Z	Sig.
The ____ interface was easy to use.	3	11	-1.74	0.08
I would use the ____ interface in my car.	3	11	-2.14	0.03

4.8 IVIS Usability Discussion

Overall, the Multi-PTT interaction technique allowed our participants to complete their interactions nearly 40% faster than when using the Single-PTT interface. This increase in task performance occurred without a corresponding increase in error rate. Indeed, the Multi-PTT condition resulted in fewer task errors than the Single-PTT interface, although not significantly so. The difference in task-time is a likely explanation for the observed difference in user preference between the two techniques.

4.9 Driving Performance Results

While the Multi-PTT interface was shown to result in preferential and performance benefits when compared to the Single-PTT technique in terms of interaction with the in-car device itself, the impact of these improvements would be negated if the use of the Multi-PTT interface greatly interfered with the primary driving task. In this section, we report on the results of our experiment as they relate to driving performance.

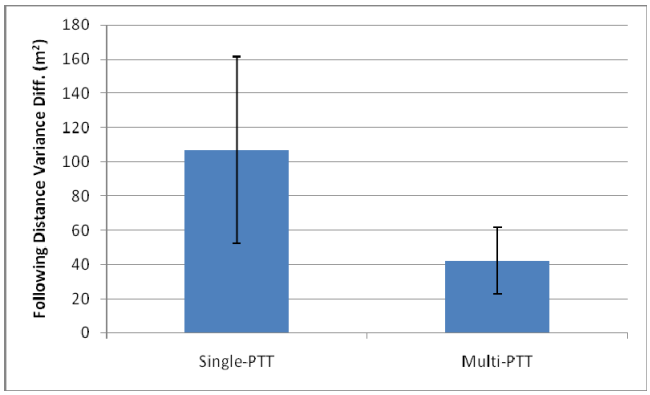

Figure 6. Mean variance differences in following distance for the two *interaction-techniques*.

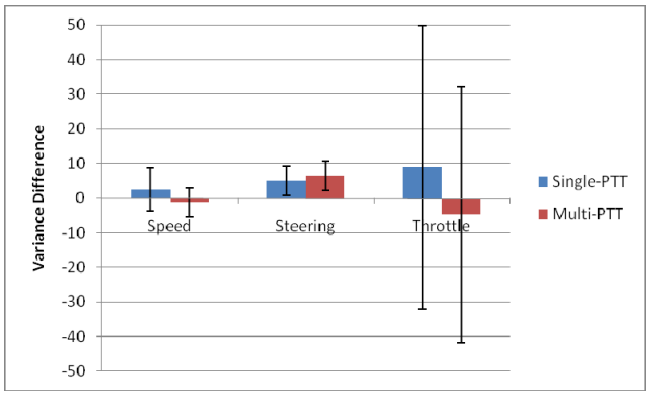

Figure 8. Mean variance difference in speed, steering angle, and throttle percentage for the two *interaction-techniques*.

4.9.1 Following Distance

As with our other measurements, we first checked for asymmetrical learning effects between *interaction-techniques*. The order-of-presentation had no significant effect on any of our measurements ($F_{5,12} = 0.95$, $p = 0.49$); thus, it is safe to continue with a within-participant design.

We found a significant difference in the following distance variances between our two *interaction-techniques* ($F_{1,17} = 7.68$, $p < 0.05$). The mean variance differences in following distance were 106.7 and 42.3 for Single-PTT and Multi-PTT respectively (Figure 6).

4.9.2 Lane Deviation, Speed, Steering, and Throttle

Our analysis found no significant difference between our two *interaction-techniques* in terms of differences of variances of lateral deviation, speed, steering angle, or throttle. Figure 7 shows the mean variance difference in lateral deviation for our two techniques, and Figure 8 shows our results for speed, steering angle, and throttle variance difference.

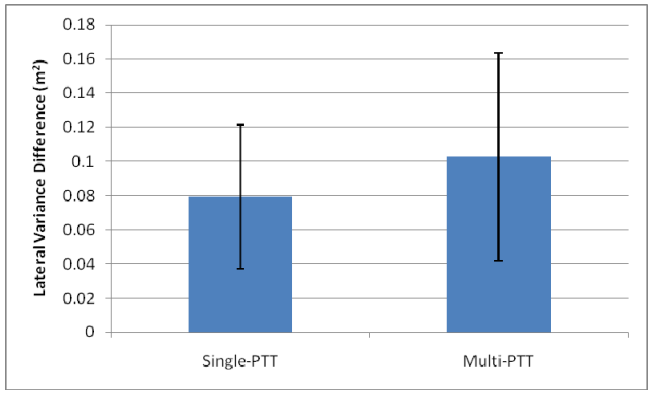

Figure 7. Mean variance difference in lateral deviation for the two *interaction-techniques*.

4.9.3 Eyes-on-the-Road

Our experimental software kept track of the number of times the participants looked away from the forward roadway. We found a significant difference in the number of these glances between our two *interaction-techniques* ($F_{1,17} = 13.50$, $p < 0.005$). On average, participants looked away 11.2 times per trial when using the Single-PTT technique and 7.9 times with the Multi-PTT technique. Figure 9 shows the number of glances away from the forward roadway for each *interaction-technique*, separated by the duration of these glances – with short glances under 0.5s, medium between 0.5s and 2s, and long over 2s.

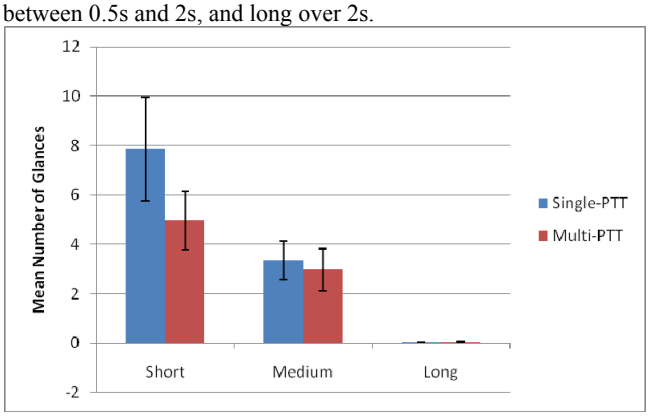

Figure 9. The mean number of glances away from the forward roadway for each *interaction-technique*, grouped by the duration of the glance.

4.9.4 Subjects' Evaluation of Driving Performance

The questionnaire completed by each participant at the end of the study included several questions concerning their impression of the driving distraction that occurred when using each of the *interaction-techniques*. Table 3 shows the number of participants who agreed more strongly with the given statements as applied to each interface. None of the questions resulted in a significant difference between our participants' agreement with one statement over the other.

Table 3. The number of participants who agreed more strongly with the driving performance statements for each of the two interfaces and the Wilcoxon Signed Ranks Test statistics. (note: totals do not sum to 18 because some participants agreed equally with both statements)

Statement	Single-PTT	Multi-PTT	Z	Sig.
The ____ interface distracted me from driving.	4	6	-0.49	0.63
My driving performance degraded while using the ____ interface.	8	2	-1.46	0.15
The ____ interface required me to look away from the road.	5	6	-0.05	0.96

4.9.5 Mental Load

After completing each of the three driving conditions, our participants completed a NASA TLX questionnaire, which is designed to subjectively measure the mental load for an activity. Figure 10 shows the mean values for each of the three conditions. There was a significant difference among these three conditions ($F_{2,34}$ = 26.05, $p < 0.001$), with means of 29.5, 55.2, 49.1 for control drive, Single-PTT, and Multi-PTT respectively (higher values indicate higher mental load). A post-hoc comparison indicated significant differences among all possible pairs.

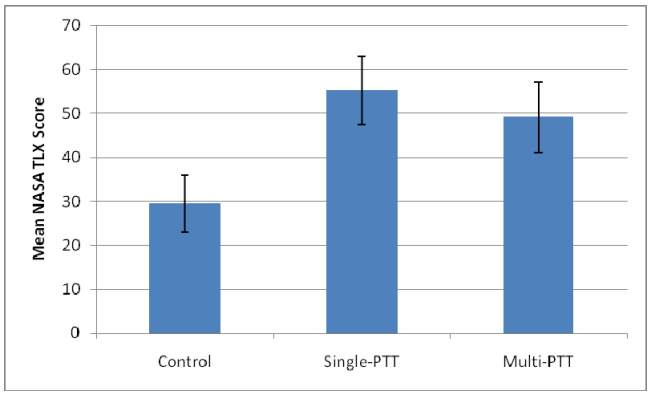

Figure 10. Mean mental load measurements for each of the three driving conditions, as measured by NASA TLX.

4.10 Discussion

The use of the Multi-PTT interface resulted in better driving performance than the Single-PTT interface as measured by following distance, glances away from the road, and mental load. It also offered usability benefits in terms of its significantly lower error (unsuccessful task) rate and shorter task duration. The shorter task duration is unsurprising: Multi-PTT bypasses one or more system states, skipping at least one user utterance and at least one system response. Regarding error rate, recall that tasks were automatically marked failures after 60 seconds. Because subjects generally took longer to complete tasks using Single-PTT, these tasks were more likely to hit the time limit and therefore be marked as errors.

It is tempting to explain the difference in vehicle following distance by attributing this to the difference in glances away from the road, because glances away from the road presumably interfere with one's ability to match the speed of the pace car. However, Figure 9 shows that the difference in glances away from the road is largely due to short-duration glances, which should not interfere with following as much as long or medium glances would. This suggests that the following distance improvements may not be due entirely to visual interference, but rather to a combination of visual demands and the cognitive demands illustrated by Figure 10.

While there was no significant difference between the two *interaction-techniques* in terms of lateral deviation (steering), Figure 7 illustrates an important finding. In both the Single-PTT and Multi-PTT conditions, the difference between the test condition and the control condition is positive and neither confidence interval crosses zero, indicating that there is a difference between driving while using either of the two *interaction-techniques* and driving alone. Subjective workload was also significantly higher for both IVIS than for unencumbered driving, as shown in Figure 10. In other words, operating either of the interfaces while driving is not without a cost in terms of driving safety. This is consistent with previous findings [12, 13].

5. CONCLUSION AND FUTURE WORK

Our hypothesis that the contextual push-to-talk technique shortens task times was confirmed. We also observed that it significantly reduced task error rates.

Our hypothesis that the contextual push-to-talk (Multi-PTT) technique would impose no greater distraction than a traditional voice dialog (Single-PTT) was supported by the lack of significant differences in lane deviation, steering angle, speed or throttle control between the two IVIS conditions. In fact, on the attentionally-demanding task of following a lead vehicle, our test subjects performed better using the contextual push-to-talk technique than they did using the single PTT button and traditional voice dialogs. Furthermore, they preferred the contextual push-to-talk technique and reported that it imposed a lower workload (though still higher than unencumbered driving).

A significant contributor to cognitive load, as well as to task duration, could have been the need to remember the names of the items the experimenter asked the subjects to find. In the future we would like to take a more "naturalistic" approach, wherein subjects are asked to find their own favorite POI from the local area, items from their own media collection, and contacts from their own mobile phones' address books.

Longer-term testing of both single-PTT and multiple-PTT interfaces, if possible using real vehicles, will also help us to understand the advantages and disadvantages of each approach, as well as whether and how to combine them into a road-ready product that is usable, enjoyable, and above all, safe.

6. ACKNOWLEDGEMENTS

The authors wish to thank Yuri Ivanov for editing advice and John Barnwell for help with hardware.

7. REFERENCES

1. Alewine, N., Ruback, H., and Deligne, S. 2004. Pervasive Speech Recognition. IEEE Pervasive Computing 3, 4 (Oct. 2004), 78-81.
2. Alm, H., & Nilsson, L. (1994). Changes in driver behaviour as a function of hands-free mobile phones – A simulator study. Accident Analysis & Prevention, 26, 441-451.
3. Art Lebedev Studios. Optimus Mini Three. http://www.artlebedev.com/everything/optimus-mini, Retrieved 12 May, 2010.

4. Bach, K., Jæger, M., Skov, M. B., and Thomassen, N. 2008. You can touch, but you can't look: interacting with in-vehicle systems. In Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 1139-1148. DOI= http://doi.acm.org/10.1145/1357054.1357233
5. Barón, A. and Green, P. (2006). Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review. Technical Report UMTRI 2006-5. Ann Arbor, MI: University of Michigan Transportation Research Institute.
6. Chittaro, L. and De Marco, L. Driver Distraction Caused by Mobile Devices: Studying and Reducing Safety Risks. In Proceedings of the 1st Int'l Workshop Mobile Technologies and Health: Benefits and Risks (Udine, Italy, 2004).
7. D-Box GP PRO-200 RC, http://www.d-box.com/gaming/en/products/pro-gaming-series/gp-pro-200-rc/, Retrieved 12 May, 2010.
8. Divi, V., Forlines, C., Van Gemert, J., Raj, B., Schmidt-Nielsen, B., Wittenburg, K., Woelfel, J., Wolf, P., and Zhang, F. 2004. A speech-in list-out approach to spoken user interfaces. In Proceedings of HLT-NAACL 2004: Short Papers on XX (Boston, Massachusetts, May 02 - 07, 2004). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 113-116.
9. Federal Register, Vol. 75, No. 17. FR Doc. 2010–1573 Filed 1–22–10. http://www.fmcsa.dot.gov/rules-regulations/administration/rulemakings/notices/Texting-by-Commercial-Motor-Vehicle.pdf. Retrieved 12 May, 2010.
10. Ford Sync, http://www.syncmyride.com/, Retrieved 12 May, 2010.
11. Forlines, C., Schmidt-Nielsen, B., Raj, B., Wittenburg, K., Wolf, P. A Comparison between Spoken Queries and Menu-based Interfaces for In-Car Digital Music Selection. In Proceedings IFIP TC13 International Conference on Human-Computer Interaction (September 12 - 16, 2005). INTERACT '05. Springer, Berlin, Germany, 536-549.
12. Garay-Vega, L., Pradhan, A.K., Weinberg, G., Schmidt-Nielsen, B., Harsham, B., Shen, Y., Divekar, G., Romoser, M., Knodler, M., Fisher, D.L. 2010. Evaluation of Different Speech and Touch Interfaces to In-Vehicle Music Retrieval Systems. Accident Analysis & Prevention, 42, 3 (May 2010), 913-920.
13. Graf, S., Spiessl, W., Schmidt, A., Winter, A., and Rigoll, G. 2008. In-car interaction using search-based user interfaces. In Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 1685-1688. DOI= http://doi.acm.org/10.1145/1357054.1357317
14. Hart, S. G., & Staveland, L. E. 1988. Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human Mental Workload, 139-183. Amsterdam, The Netherlands: Elsevier.
15. Image Space Incoporated, rFactor. http://www.rfactor.net/, Retrieved 12 May, 2010.
16. Itoh, K., Miki, Y., Yoshitsugu, N., Kubo, N., & Mashimo, S. 2004. Evaluation of a Voice-Activated System Using a Driving Simulator. SAE paper 2004-01-0232. Warrendale, PA: SAE.
17. Kun, A. L., Paek, T., Medenica, Ž., Memarović, N., and Palinko, O. 2009. Glancing at personal navigation devices can affect driving: experimental results and design implications. In Proceedings of the 1st international Conference on Automotive User interfaces and interactive Vehicular Applications (Essen, Germany, September 21 - 22, 2009). AutomotiveUI '09. ACM, New York, NY, 129-136. DOI= http://doi.acm.org/10.1145/1620509.1620534
18. Palinko O., Kun A.L. 2008. Steering Wheel Sensor as a Push-To-Talk Solution. In Proceedings of the Fourth IET International Conference on Intelligent Environments (IE08), Seattle, WA, July 21-22, 2008.
19. rFactor Central, Coast Track 1.00, http://www.rfactorcentral.com/detail.cfm?ID=Coast%20Track, Retrieved 12 May, 2010.
20. Richtel, M. (2009). "Utah Gets Tough with Texting Drivers." New York Times, August 29, 2009. http://www.nytimes.com/2009/08/29/technology/29distracted.html, Retrieved 12 May, 2010.
21. Salvucci, D. D., Markley, D., Zuber, M., and Brumby, D. P. 2007. iPod distraction: effects of portable music-player use on driver performance. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 243-250. DOI= http://doi.acm.org/10.1145/1240624.1240665
22. Seeing Machines faceLAB 5, http://www.seeingmachines.com/product/facelab/, Retrieved 12 May, 2010.
23. Sicconi, R., White, K. D., Ruback, H., Viswanathan, M., Eckhart, J., Badt, D., Morita, M., Satomura, M., Nagashima, N., Kondo, K. 2009. Honda Next Generation Speech User Interface. SAE World Congress & Exhibition, April 2009.
24. Strayer D. L., Drews, F. A., Crouch D. J. 2006. A Comparison of the Cell Phone Driver and the Drunk Driver. Journal of the Human Factors and Ergonomics Society, 48, 2 (2006), 381-391.
25. Texting While Driving PSA, http://www.youtube.com/watch?v=DGE8LzRaySk, Retrieved 12 May, 2010.
26. Walker, M. A., Borland J., Kamm, C. A. 1999. The utility of elapsed time as a usability metric for spoken dialogue systems. In Proceedings of the Sixth IEEE Workshop on Automatic Speech Recognition and Understanding. ASRU '99. 317-320.
27. Weinberg, G. 2009. Contextual push-to-talk: a new technique for reducing voice dialog duration. In Proceedings of the 11th international Conference on Human-Computer interaction with Mobile Devices and Services (Bonn, Germany, September 15 - 18, 2009). MobileHCI '09. ACM, New York, NY, 1-2. DOI= http://doi.acm.org/10.1145/1613858.1613960
28. Weinberg, G. and Harsham, B. 2009. Developing a low-cost driving simulator for the evaluation of in-vehicle technologies. In Proceedings of the 1st international Conference on Automotive User Interfaces and Interactive Vehicular Applications (Essen, Germany, September 21 - 22, 2009). AutomotiveUI '09. ACM, New York, NY, 51-54. DOI= http://doi.acm.org/10.1145/1620509.1620519
29. Wolf, P., Woelfel, J., van Gemert, J., Raj, B., and Wong, D. 2004. SpokenQuery: An Alternate Approach to Choosing Items with Speech. In Proceedings of the International Conference on Speech and Language Processing (Jeju Island, South Korea, October 4-8, 2004). ICSLP '04. ISCA, 2004, 221-224.