MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Display Style Considerations for In-Car Multimodal Music Search

Garrett Weinberg, Dhimiter Kondili

TR2008-038 August 2008

Abstract

In this pilot study, the authors employed a basic driving simulator to examine both driving behavior and task performance as subjects performed music retrieval tasks using one of three variants of the ”SpeakPod” voice search prototype. The variants shared speech and manual interface designs but differed in visual output capabilities. Preliminary data indicate that the chosen variant and the presence or absence of a music search task had little impact on the chosen driving metric. Post-drive NASA-TLX survey results do not show any of the three variants to be any more cognitively demanding than any other. There was also no clear winner in terms of task success rate.

IHCI 2008

DISPLAY STYLE CONSIDERATIONS FOR IN-CAR MULTIMODAL MUSIC SEARCH

Garrett Weinberg
Mitsubishi Electric Research Labs
201 Broadway 8th floor
Cambridge, Massachusetts 02139, U.S.A.
weinberg@merl.com

Dhimiter Kondili
Tufts University
161 College Ave
Medford, Massachusetts 02155, U.S.A.
dhimiter.kondili@tufts.edu

ABSTRACT

In this pilot study, the authors employed a basic driving simulator to examine both driving behavior and task performance as subjects performed music retrieval tasks using one of three variants of the "SpeakPod" voice search prototype. The variants shared speech and manual interface designs but differed in visual output capabilities. Preliminary data indicate that the chosen variant and the presence or absence of a music search task had little impact on the chosen driving metric. Post-drive NASA-TLX survey results do not show any of the three variants to be any more cognitively demanding than any other. There was also no clear winner in terms of task success rate.

KEYWORDS

Speech recognition, driving simulation, screen size

1. INTRODUCTION AND RELATED WORK

1.1 Motivation

Leaving aside concerns about display positioning raised by Lamble (1999) and others, a valid argument could be made that either a minimal or a maximal display is more appropriate for multimodal (speech/manual) side tasks undertaken in combination with the primary task of driving. Larger, high-resolution displays offer better text and icon clarity, but automotive HMI designers often succumb to the temptation to fill the available screen space with drop shadows, animations, and other potentially distracting "eye candy." Smaller, text-only displays, on the other hand, offer high contrast and fewer distracting bells and whistles, but their low pixel counts and font aliasing can make at-a-glance comprehension difficult.

The following paper presents the results of an initial investigation into the advantages and disadvantages of minimal versus maximal displays in the context of in-car multimodal search.

1.2 Speech, In-Vehicle Technologies, and Driver Distraction

The Society of Automotive Engineers has introduced a so-called "15-second rule" governing the use of in-vehicle technologies (IVTs). This widely-adopted guideline states that while the vehicle is in motion, tasks carried out using an IVT should last no longer than 15 seconds (SAE 2004). Although "voice-activated controls" are granted an explicit exemption from the SAE's guideline, it may the case that some voice-activated interfaces distract the driver more than others. Of particular concern would be those interfaces that require significant hands-off-the-wheel and/or eyes-off-the-road time.

The preponderance of research in this area compares the distraction caused by speech/multimodal interfaces to that caused by their manual-only alternatives (Baron & Green 2006). Less attention has been paid to variations *among* different voice user interface (VUI) designs or implementations. Schmidt-Nielsen et al. suggest (2008) that such factors as poor recognition accuracy, inflexible dialog pacing, and inconsistent command structure can significantly decrease usability and consequently increase the cognitive load imposed upon the driver, with potentially dangerous consequences.

The central role of the search result list in the VUI style called Speech-In, List-Out, or SILO (Divi 2004), may call for a different set of constraints and guidelines as compared to more traditional, command-oriented speech applications such as those cited in the SAE's 15-second rule.

While it is true that SILO-style applications in an automotive context allow for users to attend to the roadway while speaking search terms or listening to audible feedback, what if the sought item is not the top match in the result list? In this case one must try again—possibly reformulating one's search phrase—or one must somehow browse through the result list to find the sought item, all while maintaining focus on the primary task of safely operating the vehicle. A wheel-mounted input device is a helpful affordance during this browsing process, but what role does result display style and legibility play?

By conducting an experiment in which the vocal and wheel-mounted tactile aspects of a SILO voice search application remained constant and only the visual aspect varied, we hoped to learn more about how best to apply the SILO paradigm to the automotive environment. Our intention was furthermore to test the fidelity of our driving simulation setup* and the applicability of our analysis methodologies, in hopes of arriving at suitable, repeatable metrics for future work on both driver distraction and multimodal usability.

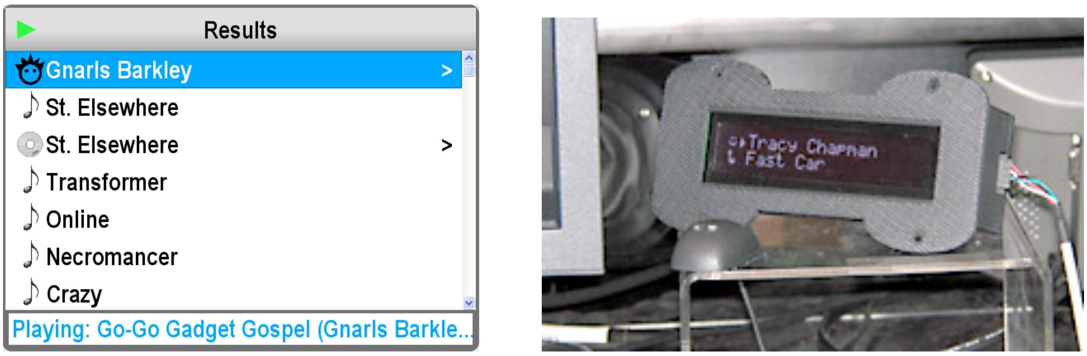
2. EXPERIMENT

2.1 Software and Hardware Employed

SpeakPod is an application prototype based on the SpokenQuery (Wolf et al. 2004) voice search engine. In addition to traditional hierarchical browsing and playback of a digital music collection using a GUI similar to the iPod's, SpeakPod allows for the retrieval of music using simple spoken phrases formulated in a manner similar to Google queries for Web pages.

In the current iteration of SpeakPod, various result types are grouped together into a single match list and displayed in relevance order. The list appears automatically when speech recognition and SpokenQuery lookup are complete, and the top match immediately begins playing (see Figure 1).

Figure 1. Left: Top seven results for query "gnarls barkley saint elsewhere" on interface variant A. Right: Top two results for query "tracy chapman fast car" on interface variant B.



Results

Gnarls Barkley >

St. Elsewhere

St. Elsewhere >

Transformer

Online

Necromancer

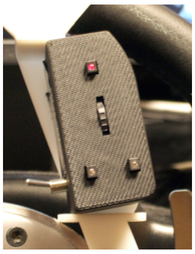
Crazy

Playing: Go-Go Gadget Gospel (Gnarls Barkle...

* Testing in real vehicles is not feasible at this time due to insurance and liability constraints.

Manual interaction with SpeakPod takes place using a custom-made wireless input device (Figure 2) approximately 7x3.5x4 cm in size that is attached to the simulator's steering wheel in an unobtrusive position.

Figure 2. Input device



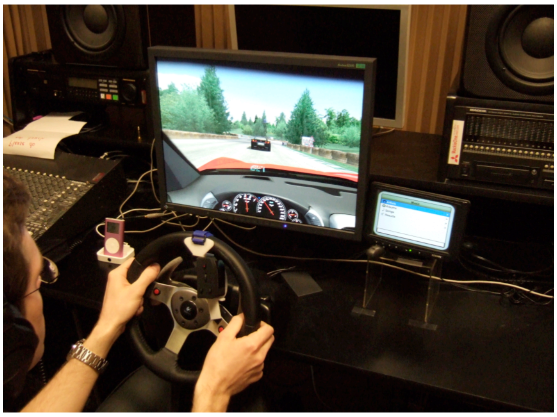
At the device's center is a "jog dial" widget that offers unlimited bidirectional rotation and an actuator. This widget serves to change the currently selected node and to activate/open that node. The button above the jog dial returns to the next uppermost level in the hierarchy. The lower right button pauses or resumes playback, and the lower left button activates voice input (press to talk, release at any time; a short tone indicates the system is listening). The metal switch protruding from the left face of the device is its power toggle.

Upon selection changes (jog dial rotation) and state changes (e.g. completion of a voice query), the type and name of the newly selected item are read aloud by a state-of-the-art commercial text-to-speech (TTS) engine. For example, as one moved through a result list for the query "david bowie," one might hear "Artist, David Bowie," "Album, Hunky Dory," "Song, Changes" "Song, Life on Mars," etc.

The commercial racing game rFactor (Image Space Incorporated, 2008) was chosen 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. A game plugin was written to capture the subject's input at rates up to 90 Hz.

A Windows XP PC with a high-end graphics card served as the primary hardware platform, running both rFactor and SpeakPod simultaneously with no perceptible scene rendering or speech recognition lag (SpeakPod has also been ported to more limited Windows CE-based platforms, which were not used in this experiment). A Logitech G25 force-feedback steering wheel and pedal set provided driving input (an automatic transmission setting was used). A 21-inch (53 cm) LCD was placed immediately behind the steering wheel's base and was adjusted to a height such that the real-world steering wheel was at the appropriate level relative to the virtual dashboard depicted in the game (see Figure 3).

Figure 3. Simulator with variant A shown



As shown in Figure 3, for SpeakPod variants that used a display, the display was placed on a small stand immediately to the right of the primary LCD. Voice input was captured by a standard AKG Q400 automotive microphone affixed to the same stand.

2.2 Interface Variants

We prepared three variants of SpeakPod for use in this experiment. Variant A (Figure 1, left) approximates the "look and feel" of the iPod, with icons added and the font size adjusted for legibility. In variant B, two lines of 20 characters each constitute the entire GUI, with the top row representing the selected item. Variant C has no screen at all. Users find their way within the item lists by manipulating the wireless input device and listening to the synthesized speech readout.

2.3 Experimental Protocol

22 licensed drivers ranging in age from 19 to 34 (median age 22.5) were paid to take part in the study. 10 were female. One male and one female subject drove extremely erratically and seemed not to take seriously the experimenter's instructions to treat the game more like normal driving than racing. These subjects' data were excluded. All subjects were required to own iPods that contained at least 400 songs, under the assumption that voice search among that many songs would be more efficient than manual search. Upon arrival, a subject connected his or her iPod and the experimenter generated a music search task list based on the iPod's contents.

Subjects were arbitrarily assigned to use one of the three SpeakPod variants, and were given a brief explanation/demonstration of the chosen variant's capabilities. They then received approximately five minutes training time using the simulator (a gently curving highway course was used during training).

The test sessions themselves took place on a course consisting of a mix of town and highway sections (approx. 30%/70% town/highway split). Our rFactor plugin logged driving behavior continuously throughout 20 minutes of drive time. *In-task* (IT) and *out-of-task* (OOT) periods alternated, with OOT period durations varying dynamically depending on accumulated IT time for the drive (however, the minimum OOT period duration was 20 seconds, and maximum was one minute). The experimenter used a separate device to prompt tasks, time them, and mark them as successful or failed. Tasks lasting more than two minutes were automatically marked as failures. At the start of each task, the experimenter verbally announced the task objective using standardized wording.

Subjects wore closed-cup headphones in which a mix of SpeakPod sounds (listening tone, TTS readout, audio playback), simulator effects (engine and road noise), and a sidetone (their own and the experimenter's voices) could be heard.

Immediately following each 20-minute test session, an electronic version of the NASA-TLX (Task Load indeX) survey (Hart & Staveland 1988) was administered. With the experimenter not present in the room, subjects rated the combined mental and physical demands imposed by operating the simulator and using SpeakPod simultaneously.

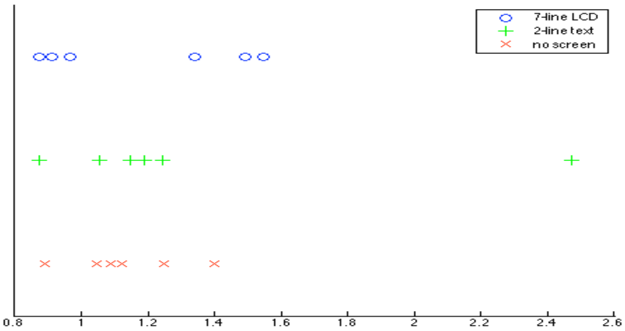
RESULTS

As Baron & Green note (2006, pp. 25-28), most investigations of side task performance during actual or simulated driving focus on one or more objective driving performance measures, such as lane deviation or following distance; one or more objective usability measures, such as task completion rates; and one or more subjective assessments, such as the NASA-TLX workload survey mentioned above.

As our simulator is still in the design and testing phase, we decided against having an explicit control group of subjects who performed no music retrieval tasks. With a mature, high-fidelity simulator calibrated to behave similarly to an actual vehicle, one could expect any licensed driver to be capable of operating the simulator in a relatively standard manner on any course with which they are presented. With our current simulator, on the other hand, subjects with higher levels of video gaming experience tended to drive less erratically than subjects with less video gaming experience—especially on the more technical town-driving portion of the course. For this reason, the driving data gathered during a subject's own out-of-task (OOT) periods is the only legitimate control data to which to compare his in-task (IT) periods.

A single driving metric was chosen for simplicity at this stage of the research. The metric chosen was average rate of change in steering wheel angle—a measure of how erratic steering input is over a given span of time. As shown in Figure 4, the ratios of IT to OOT means for steering angle values were in most cases greater than one, meaning that subjects tended to steer more erratically during music retrieval tasks.

Figure 4. Mean rate of change in steering angle: in-task to out-of-task ratios for each study subject



However, when separate single-factor ANOVAs were performed on each subject's OOT and IT steering angle averages, the p value obtained was below the coincidence threshold ($\alpha = 0.05$) for only one among the 20 subjects, meaning that the null hypothesis that there is no difference between in-task and out-of-task behavior cannot confidently be rejected.

Subjects' own assessments of their combined mental and physical workload paint a similarly indistinct picture. While it appears from the data in Table 1 that variant C (no screen) is the most cognitively demanding, a single-factor ANOVA on the raw data yields low confidence ($p = .64$) in this outcome, probably due to the small size of the dataset.

Table 1. NASA-TLX results

Interface	**# Subjects**	**Median Total Workload**
A (7-line LCD)	7	54.00
B (2-line text)	6	48.50
C (no screen)	7	59.33

It would be reasonable to expect, however, that variant C might still be judged the most cognitively demanding in a study with more statistical power. Although it poses no risk from distraction due to the "eye candy" effect, the lack of visual cues results in the need to maintain a more detailed mental model of selection position and/or menu hierarchy. Past work (Muttart et al. 2007, Strayer et al. 2001) has demonstrated decreased driver awareness due to the cognitive load involved in semantic understanding tasks. The load involved in mentally modeling variant C's application state may have a similar effect.

Usability metrics included the number of music retrieval tasks completed during a subject's drive, as well as the number of these tasks that were successful. These data are presented in Table 2.

Table 2. Music retrieval task breakdown

Subject group	**Mean tasks completed per 20-minute drive**	**Mean successful tasks per 20-minute drive**	**Percentage successful**
A (7-line LCD)	21.71	19.29	88.8%
B (2-line text)	25.00	23.83	95.3%
C (no screen)	22.86	21.43	93.8%

Variant B seems to have a slight edge here, but another single-factor ANOVA performed on the raw data shows that the null hypotheses—that there is effectively no difference between the three variants in terms of task completion or task success rates—cannot be rejected ($p = .71$ for task completion and $p = .63$ for task success).

CONCLUSIONS AND FUTURE WORK

In neither objective driving behavior measurement, objective task performance measurement, nor subjective workload assessment is there a statistically significant indication that any of these three multimodal music-retrieval interface variants caused more driver distraction or was less usable than any other. Nor is there an indication that being in-task versus out-of-task using any of the variants caused driving behavior to degrade to a significant degree.

It is possible to draw either of two conclusions from these findings. Potentially any of the three variants causes little enough distraction to merit future study as an in-car music-retrieval solution, since none of the variants caused steering behavior significantly more erratic while in-task than while out-of-task. Alternatively, because of the relatively small size of this study and the paucity of driving metrics used, we could simply lack the discriminative power to identify one of the variants as less distracting and/or more usable than the others.

In order to better elucidate differences among these interface variants in future work, the driving simulator we employed must be further standardized and calibrated so that any licensed driver with any level of video gaming experience can perform within acceptable norms on the virtual roadway. In other words, the simulator's fidelity must be increased. In addition to course design and vehicle telemetry considerations, this process may involve adding motion cues such as chair rumble and tilt, as well as introducing larger, more immersive primary displays.

Another major area of work lies in the expansion and refinement of the driving behavior metrics used. It is particularly vital to use gaze-tracking equipment such as that employed by Muttart et al. (2007) in order to examine the ways in which differences in systems' visual modalities or their lack of a visual modality contribute to both road-scanning behavior and subjects' perceived or measured cognitive load.

ACKNOWLEDGMENTS

Thanks to Bret Harsham, Bent Schmidt-Nielsen, John Barnwell, Janet McAndless, and Cliff Forlines for vital assistance with various aspects of this work.

REFERENCES

Baron, 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*. University of Michigan Transportation Research Institute at Ann Arbor, Michigan, U.S.A.

Divi, V. et al., 2004. A speech-in-list-out approach to spoken user interfaces. In *Proceedings of the Human Language Technology Conference*. Boston, Massachusetts, U.S.A. Association for Computational Linguistics, pp. 113-116.

Hart, S.G. and Staveland, L.E. 1988. Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*. Hancock, P.A. and Meshkati, N., eds. North-Holland: Elsevier, pp. 139-183.

Image Space Incorporated, 2008. http://www.rFactor.net

Lamble, D. et al., 1999. Detection thresholds in car following situations and peripheral vision: implications for positioning of visually demanding in-car displays. *Ergonomics,* Vol. 42, pp. 807-815.

Muttart, J.W. et al., 2007. Driving Without a Clue: Evaluation of Driver Simulator Performance During Hands-Free Cell Phone Operation in a Work Zone. *Journal of the Transportation Research Board,* Vol. 2018, pp. 9-24.

Schmidt-Nielsen, B. et al., 2008. Speech Based UI Design for the Automobile. In *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*. Lumsden, J., ed. National Research Council of Canada, Ottowa, Ontario, Canada. Vol. 1, Ch. 15, pp. 237-252.

Strayer, D.L. et al., 2001. Cellphone-induced perceptual impairments during simulated driving. In *Proceedings of the International Driving Symposium on the Human Factors in Driver Assessment, Training, and Vehicle Design.* Aspen, Colorado, U.S.A.

Society of Automotive Engineers, 2004. SAE Recommended Practice Navigation and Route Guidance Function Accessibility While Driving (SAE 2364). Society of Automotive Engineers, Warrendale, Pennsylvania, U.S.A.

Wolf, P. et al., 2004. SpokenQuery: an alternate approach to chosing items with speech. In *Proceedings of the 8th International Conference on Spoken Language Processing*. Jeju Island, Korea, pp. 221-224