Multimodal Input in the Car, Today and Tomorrow

Muller, C.; Weinberg, G.

TR2011-002 January 2011

Abstract

With the increased functionality offered by in-vehicle systems, multimodal input is emerging as an effective means of interaction to minimize driver distraction. This article describes the current state of this technology for automotive applications, various ways to combine modalities, and outlooks toward the future.

IEEE MultiMedia

Multimodal Input in the Car, Today and Tomorrow

Christian Müller and Garrett Weinberg

***Index Terms*—Multimodal, input, automotive, speech, voice, gesture, touch, gaze.**

AFTER a surge in horrific automobile accidents in which distracted driving was proven to be a factor, 38 US states have enacted texting-while-driving bans [9]. While nearly everyone can agree that pecking out a love note on a tiny mobile phone keypad while simultaneously trying to operate a vehicle is bad idea, what about the other activities that we perform on a day-to-day basis using the electronic devices either built in or brought in to our cars? Finding a nearby restaurant acceptable to the vegetarian in the back set? Locating and queuing up that new album you downloaded to your iPod?

This article offers a brief overview of multimodal (speech, touch, gaze, etc.) input theory as it pertains to common in-vehicle tasks and devices. After a brief introduction, we walk through a sample multimodal interaction, detailing the steps involved and how information necessary to the interaction can be obtained by combining input modes in various ways. We also discuss how contemporary in-vehicle systems take advantage of multimodality (or fail to do so), and how the capabilities of such systems might be broadened in the future via clever multimodal input mechanisms.

I. The Unique Problems of In-Vehicle Interaction

The reason activities such as finding music or deciding on a restaurant are challenging, and indeed sometimes dangerous, is that humans have a limited capacity for carrying out multiple tasks at once (see [19] for a thorough treatment of this topic). Geiser classifies driving-related activities into the following categories: 1) primary tasks, involved in maneuvering (e.g., turning the steering wheel and operating the pedals); 2) secondary tasks, involved in maintaining safety (e.g., turn signals, windshield wipers); and 3) tertiary tasks, involving all other comfort, information, and entertainment functions [8]. While there has been some progress made in the design and development of workload managers that automatically lock out some or all tertiary functions as the difficulty of the primary task increases [10], [1], there are still numerous technical challenges to overcome. In the meantime, car makers and electronics suppliers have taken an ad-hoc approach toward building in-car interfaces that minimize distraction. Internationally-recognized standards are few and far between; "best practices" dominate instead. The Society of Automotive Engineers recommends, for example, that any tertiary task taking more than 15 seconds to carry out while stationary be disallowed while the vehicle is in motion (the so-called "15-Second Rule") [17].

A. Speech to the Rescue?

Voice-activated controls are explicitly exempted from the 15-Second Rule. But should they be? Some data suggest that certain kinds of voice interfaces impose inappropriately high cognitive loads and can negatively affect driving performance (e.g., [7], [6]). This is due to technical limitations within of the underlying automatic speech recognition (ASR) engines (in particular, the inability to distinguish among acoustically similar words given a large enough vocabulary), as well as usability flaws such as confusing or inconsistent command sets and unnecessarily deep and complex dialog structures.

This situation was in part brought about by car makers' "feature-itis;" in an intensely competitive market, each manufacturer wanted to bring as many products having voice recognition capability onto the market as quickly as possible. Speech was often "bolted on" to existing systems as a separate and independent feature. This led to a situation still common in vehicular interfaces: there is a manual way to accomplish something, and a voice-enabled way to accomplish something, and never the twain shall meet. The remainder of this article discusses how this quandary can be overcome, and how current research into combinations of speech and other forms of input will eventually enable in-car devices to accomplish what might today seem far-fetched.

B. Multimodality

Oviatt defines multimodal systems as "those that process two or more combined user input modes—such as speech, pen, touch, manual gestures, gaze, and head and body movements—in a coordinated manner with multimedia system output" [14]. We will focus on the multimodal input in this article, but there is burgeoning research on multimodal output in the vehicular context as well. For example, visual, audible and haptic alerts can be combined to notify the driver about the proximity of other vehicles during lane changes [16], [2].

To understand both the advantages and limitations of today's multimodal in-vehicle interfaces, and to better understand what the future might hold, we need to "think multimodally." The best way to learn to do this is to deconstruct a sample in-vehicle task.

II. Thinking Multimodally

Figure 1 illustrates a simple multimodal interaction scenario: a driver lowers the front-right window a little bit, and then—before performing another tertiary task—lowers it a little bit more. We will refer to this example in explaining

Christian Müller is a Senior Researcher at the German Research Institute for Artificial Intelligence, Saarbrücken, Germany. http://www.dfki.de/cmueller

Garrett Weinberg is a Member of Research Staff at Mitsubishi Electric Research Labs, Cambridge, Massachusetts, USA. http://www.garrettweinberg.com

Manuscript received July XX, 2010; revised XX.

the “nuts and bolts” of multimodal input for drivers. We depict the interaction as a directed acyclic graph; nodes of the graph correspond to individual interaction subgoals, while edges correspond to the means for accomplishing these goals. A research prototype implementing much of what is discussed in this section was developed and studied by [4]. See [3] for a video.

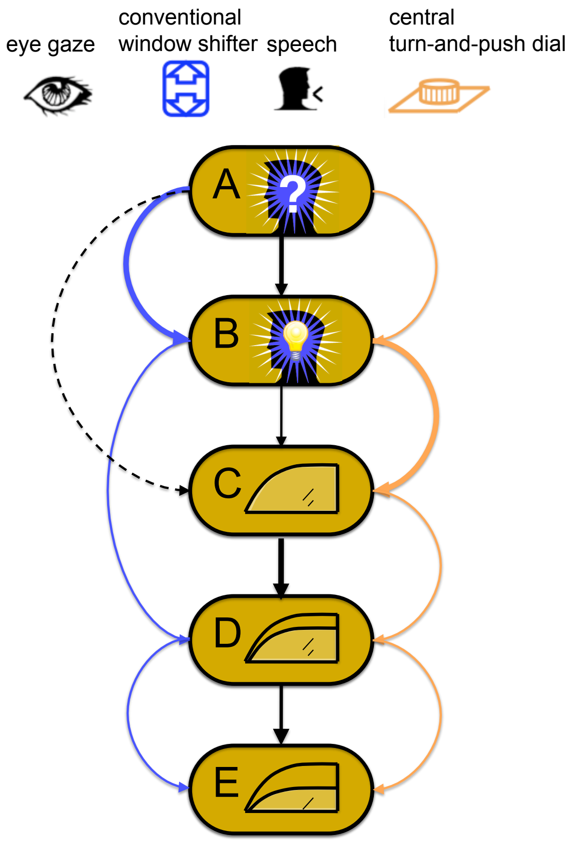

Fig. 1. Directed acyclic graph representing a driver’s intention and action to open the front-right window a little bit. We use this sample interaction to discuss the advantages and disadvantages of individual input modalities, various techniques for combining them, as well as extensions to multimodal input theory such as implicit interaction, inferred interaction, and an interaction cost model. Heavier line weights correspond to high-cost interaction steps.

A. Advantages and Drawbacks of Individual Modalities

Multiple modalities are available at any time during this sample interaction. In this example, there is the conventional electric window shifter button (often mounted on the door), speech command, and a multifunction “turn-and-push” knob (often mounted between the driver and front passenger seats). For each modality, we will focus in particular on the difficult steps in the interaction—those drawn with heavier line weights in the figure—as the other steps are as straightforward to understand as they are to carry out in a real car.

The blue edges in the graph correspond to using the conventional window shifters. Here, the first step consists of knowing where the button is, which we assume to be demanding in cases when the driver is not familiar with the particular car. For example, think how difficult it can be to find the windshield wiper controls on a rental car.

The orange edges correspond to the use of a multifunction turn-and-push knob of the sort found in many current luxury-tier vehicles. Here, the costliest step is the context selection step, i.e. determining how to carry out what you want to do. This is because multifunction devices tend to engender UI designs where the driver must browse through hierarchical menus in order to pick the desired action, and this process may in fact be rather demanding. However, once the correct node is selected within the correct subtree, the manipulation itself (lowering the window a little bit) can be done intuitively and gradually by pulling or twisting the knob ($\overrightarrow{CD}$ and $\overrightarrow{DE}$).

This gradual manipulation step ($\overrightarrow{CD}$) is the Achilles heel of speech-based interaction (the black edges in the graph above). Opening the window *just slightly* is not at all an intuitive operation to perform solely via speech. First of all, gradual manipulation is lost—the window can only be opened in discrete steps from completely closed to fully open. Secondly, it is not always easy to map an in-vehicle concept onto a natural speech command. Think of the last time you were the driver in a car without power mirrors and you had to describe to your passenger how to make such an adjustment. Note, however, that $\overrightarrow{DE}$, i.e. lowering the window another small amount, is once again relatively easy to do using speech, for example by saying “more.”

A final consideration for the voice modality is the user’s need to memorize and formulate a command that is valid for a particular system state. This can be somewhat demanding as well, so we have drawn the relevant edge ($\overrightarrow{AB}$) using in intermediate line weight.

The dashed edge $\overrightarrow{AC}$ represents the use of eye gaze as an *implicit* interaction modality (one that “refer[s] to naturally occurring user behavior or actions [without] requir[ing] any explicit command” [14]). An intelligent system (equipped with an eye tracker and a sophisticated user model) could infer the driver’s intention to open the window from her gaze, and thus set the interaction context accordingly (bypassing node B in the graph). If the system takes action in this manner based on an established belief about a user’s intention, this is termed an *inferred* interaction.

B. Methods of Combining Modalities

The disadvantages of any single modality can often be overcome by combining them intelligently. This can be accomplished in various ways:

Temporally cascaded modalities. According to [14], two or more modalities are temporally cascaded if sequenced in a particular order such that partial information supplied by recognition of the earlier mode is able to constrain the interpretation of the later mode. Suppose you say “front-right window” ($\overrightarrow{ABC}$) and then immediately you push the multifunction knob downwards ($\overrightarrow{CD}$). Then, obviously, the knob manipulation should be interpreted in the context of the preceding utterance. Alternatively, you could lower the window initially by pressing the conventional window shifter ($\overrightarrow{ABD}$). If you then said “more” ($\overrightarrow{DE}$), the speech command would be interpreted using knowledge derived from the preceding manual interaction.

In terms of industry deployments, the current Sync offering by Ford and Microsoft typifies temporally cascading multi-

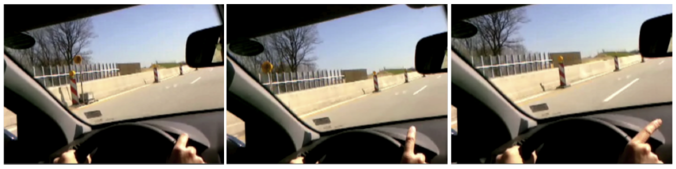

Fig. 2. Left: a sequence of video captures showing a person executing a clockwise “rotary dial” gesture while driving. The combination of this gesture with the utterance of a person’s name (e.g. “John”) could comprise a multimodal interaction for initiating phone calls [5].

modal systems as they are found in today’s vehicles. Pressing the “phone” button on the steering wheel or dashboard activates the dialing and address-book ASR grammar, constraining the interpretation of subsequent voice commands. By the same token, if the user finds herself in the USB media player mode, she could issue the “phone” command by voice, after which a press of the “menu” key on the dashboard brings up a phone-specific menu rather than a USB-specific menu.

Redundant modalities. We define redundant modalities as an special form of temporally cascaded modalities where each mode is available in each interaction step. The user can then freely pick the means by which she feels most comfortable beginning or continuing an interaction. A system offering this form of multimodality would have an interaction graph roughly corresponding to the entirety of Figure 1. If employed consistently, modality redundancy offers two significant advantages for in-car use. It allows users to accomplish interactions using the modality most appropriate to the driving situation—perhaps reserving speech for heavier-traffic situations when hands must be kept on the wheel. It also allows them to transfer longer interactions from one modality to another fluidly and transparently.

Car navigation systems featuring modality redundancy have already begun to appear on the market. The current Acura TL and Mercedes Benz E-Class, for example, feature menu items that can be activated either by means of the turn-and-push knob or by voice (as is standard throughout the industry, a steering wheel-mounted push-to-talk button initiates each voice command). While these UIs’ organization does impose a heavily hierarchical, step-wise interaction scheme, the user is given the freedom carry out each step using either input mode. Contrast this with earlier-generation systems whose voice dialog nodes lacked one-to-one correspondence with the systems’ visual/manual interfaces. Users found such systems disorienting because the available voice commands had little or nothing to do with what was showing on the screen at any given time.

Fused modalities. The most elaborate form of multimodality is modality fusion [14]. Here, multiple modes play a part in a single interaction step. To take up the “calling John” example again, suppose that, in addition to saying “John,” you write the letters “J.O.H.N.” in the air or on a touchpad using your index finger (see Figure 2). In this case the hypotheses stemming from the speech recognizer and those from the gesture recognizer could be be combined in order to improve the overall recognition accuracy. It is apparent that with high levels of background noise and larger vocabulary sizes this might offer a considerable advantage, as ASR engines can stumble in such situations. Generally, the fusion of two probabilistic knowledge sources tends to be most fruitful if the reasons for failures of the individual streams are different. In this example, background noise and cross-talk hurt speech recognition while (optical) gesture recognition is a most compromised by dynamically changing lighting conditions.

Depending whether fusion is carried out on the feature level (fusing acoustic features with optical features) or on the level of final modality-specific hypotheses, this is termed early fusion or late fusion, respectively [14].

Another example of fusion is illustrated in the following scenario. Say you’re driving past the Eiffel Tower in Paris and you wonder what this beautiful structure is called. With a suitably advanced system, you could point at the structure, simultaneously say “what building is this?,” and receive an answer. The referent of the deictic expression “this” would be disambiguated via the pointing gesture.[1]

C. Design Considerations

As discussed in section I, in-car UI’s should be designed with a focus on highly efficient interactions and a minimum of driver distraction. Therefore it is important to accompany each stage of system design—from early prototypes to mature products—with comprehensive and well-designed user studies. A number of methods can be applied to evaluate design choices and implementation parameters, starting with questionnaires (for example, the Driver Activity Load Index [15]) and proceeding into various forms of driving simulation and instrumented-vehicle experiments, if possible incorporating physiological measures of driver state such as heart rate and skin conductance (see e.g., [13]).

A given speech or multimodal interface technique might incubate at a university or corporate research lab, where usability evaluations can be carried out quickly and inexpensively using a low- or mid-fidelity simulator that supports, for example, the Lane-Change Task [11] as a measure of distraction. Later in the iterative development cycle, evaluations could be carried out on custom, high-fidelity, full-motion simulators such as those owned by the major carmakers, and eventually in real vehicles operated on closed test tracks.

III. Outlook for the Future

A. Input Cost Model

In lieu of these sometimes resource- and time-prohibitive studies, researchers and practitioners seeking to bootstrap novel multimodal interactions could benefit from a generic input cost model. In such a model, each step in a given in-vehicle interaction would be assigned a cost function that took into account such variables as vehicle speed, traffic density, and the amount of physical and cognitive energy required to perform the step. Steps would be abstracted into interaction “atoms” such as *select one item from a list of* n *items*, or

[1] The German Research Center for Artificial Intelligence is investigating this kind of interaction in the research project *Car Oriented Multimodal Interface Architectures* (CARMINA). At the time of this writing, pertinent publications were still under review. See http://automotive.dfki.de for updates.

gradually increase/decrease a scalar quantity or variable. Such a cost model would give designers some sense of how a given input technique or UI might fare in a simulator or test vehicle without necessarily taking the time to prototype it.

B. Standards

As mentioned above, there is a dearth of standards that specifically relate to the operation of multimodal in-vehicle interfaces. However there has been significant progress towards the standardization of multimodal interfaces in general. The Multimodal Interaction Working Group [21] focuses on markup languages and architectures that support the creation and consumption of multimodal (often voice+pen) websites. The HTML Speech Incubator group [20] is working on extensions to HTML5 that will make speech recognition available as a first-class input mechanism for web forms and fields. Considering that many operating systems for in-vehicle platforms include browsers that already support or will soon support HTML5, members of the automotive user interface community should pay close attention to the output of these two standards bodies.

C. Streamline for Safety

Few automotive UI designers would debate the advantages of modality redundancy (see section II-B) in reducing cognitive load; it's an obvious advantage if the driver can avoid having to think about whether she must proceed through an interaction using tactile or voice input, and can instead always use either modality. An interesting question for the design of future systems is whether the judicious use of multimodality can actually streamline tasks temporally in addition to cognitively. Early research results are promising. [18], for example, discusses a design in which the mode-switching buttons that are often clustered around navigation systems' screens are dual-purposed as domain-specific push-to-talk buttons. A single tap on one of these buttons changes to a given mode (as is normal for such buttons), but a double-tap immediately opens the microphone for voice search within that mode. While from a theoretical point of view this is simply another form of temporal modality cascading, the design combines domain selection (pressing the "phone" key in the Ford Sync example above) and push-to-talk into a single step, reducing overall interaction time by approximately 40% versus a traditional design [18]. And this encourages today's multitasking, hyper-connected driver to get back to doing what he seems increasingly loath to do: actually driving.

ACKNOWLEDGMENT

This work was partially funded by the German Ministry of Education and Research (project Car-Oriented Multimodal Interface Architectures, grant number 01IW08004).

REFERENCES

[1] R. Broström, J. Engström, A. Agnvall, and G. Markkula. Towards the next generation intelligent driver information system (IDIS): The Volvo cars interaction manager concept. In *Proceedings of the 2006 ITS World Congress*, London, Oct. 2006.

[2] Y. Cao, A. Mahr, S. Castronovo, M. Theune, C. Stahl, and C. Müller. Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2010)*, pages 239–248, Hong Kong, China, February 2010. ACM.

[3] S. Castronovo. Combining speech and turn-and-push dial to control comfort functions, 2010. http://www.youtube.com/watch?v=EhfFbmyzdR0.

[4] S. Castronovo, A. Mahr, and C. Müller. Multimodal dialog in the car: Combining speech and turn-and-push dial to control comfort functions. In *Proceedings of Interspeech (2010)*, pages 510–513. ISCA, Makuhari, Japan, 26–30 September 2010.

[5] C. Endres, T. Schwartz, and C. Müller. "geremin": 2d microgestures for drivers based on electric field sensing. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2011)*, 2011.

[6] L. Garay-Vega, A. Pradhan, G. Weinberg, B. Schmidt-Nielsen, B. Harsham, Y. Shen, G. Divekar, M. Romoser, M. Knodler, and D. Fisher. Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. *Accident Analysis & Prevention*, 42(3):913–920, May 2010.

[7] U. Gärtner, W. König, and T. Wittig. Evaluation of manual vs. speech input when using a driver information system in real traffic. In *International driving symposium on human factors in driver assessment, training and vehicle design*, 2001.

[8] G. Geiser. Man machine interaction in vehicles. *ATZ*, 87:74–77, 1985.

[9] State cell phone use and texting while driving laws. http://www.ghsa.org/html/stateinfo/laws/cellphone_laws.html.

[10] P. A. Green. Driver distraction, telematics design, and workload managers: Safety issues and solutions. Technical report, 2004.

[11] ISO 26022:2010 - road vehicles – ergonomic aspects of transport information and control systems – simulated lane change test to assess in-vehicle secondary task demand. Technical report.

[12] D. Lavrinc. Audi bringing new nav display, optional touch pad to 2012 a6 and a7/s7, 2010. http://www.autoblog.com/2010/05/19/audi-bringing-new-nav-display-optional-touch-pad-to-2012-a6-and/.

[13] B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(-1):6–12, 2009.

[14] S. Oviatt. Multimodal interfaces. In A. Sears and J. A. Jacko, editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Second Edition*, pages 413–432. CRC Press, 2 edition, Sept. 2007.

[15] A. Pauzié and G. Pachiaudi. Subjective evaluation of the mental workload in the driving context. *Traffic and Transport Psychology*, pages 173–182, 1997.

[16] M. J. Pitts, M. A. Williams, T. Wellings, and A. Attridge. Assessing subjective response to haptic feedback in automotive touchscreens. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '09, pages 11–18, New York, NY, USA, 2009. ACM.

[17] Society of Automotive Engineers (SAE). SAE recommended practice: Navigation and route guidance function accessibility while driving (SAE 2364). Technical Report SAE 2364, Jan. 2000.

[18] G. Weinberg, B. Harsham, C. Forlines, and Z. Medenica. Contextual push-to-talk: shortening voice dialogs to improve driving performance. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pages 113–122. ACM, 2010.

[19] C. D. Wickens. Multiple resources and mental workload. *Human Factors*, 50(3):449, 2008.

[20] World Wide Web Consortium. Html speech incubator group. http://www.w3.org/2005/Incubator/htmlspeech/charter.

[21] World Wide Web Consortium. W3c multimodal interaction working group. http://www.w3.org/2002/mmi/.