Anthony Vetro
Mitsubishi Electric Research Labs

Multimodal Input in the Car, Today and Tomorrow

Christian Müller
German Research Institute for Artificial Intelligence

Garrett Weinberg
Mitsubishi Electric Research Labs

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 (see http://www.ghsa.org/html/stateinfo/laws/cellphone laws.html). 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 group? Locating and queuing up that new album you downloaded to your iPod?

This article offers a brief overview of multimodal (speech, touch, gaze, and so on) 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.

The unique problems of in-vehicle interaction

The reason in-vehicle 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. Geiser classifies driving-related activities into the following categories:

> primary tasks involved in maneuvering (for example, turning the steering wheel and operating the pedals);
>
> secondary tasks involved in maintaining safety (for example, turn signals, windshield wipers); and
>
> tertiary tasks involving all other comfort, information, and entertainment functions.

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, 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 designed to 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. This is the so-called 15-second rule.[1]

Voice-activated controls are explicitly exempted from the 15-second rule. But should they be? Some data suggests that certain kinds of voice interfaces impose inappropriately high cognitive loads and can negatively affect driving performance.[2,3] This negative affect is due to the technical limitations of the underlying automatic-speech-recognition (ASR) engines (in particular, the inability to distinguish

Editor's Note

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.

—Anthony Vetro

1070-986X/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society

among acoustically similar words given a large enough vocabulary), as well as to 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.

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."[4] 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.[5]

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 examine a sample in-vehicle task closely.

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 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 in other work.[6] See http://www.youtube.com/watch?v=EhfFbmyzdR0 for a video.

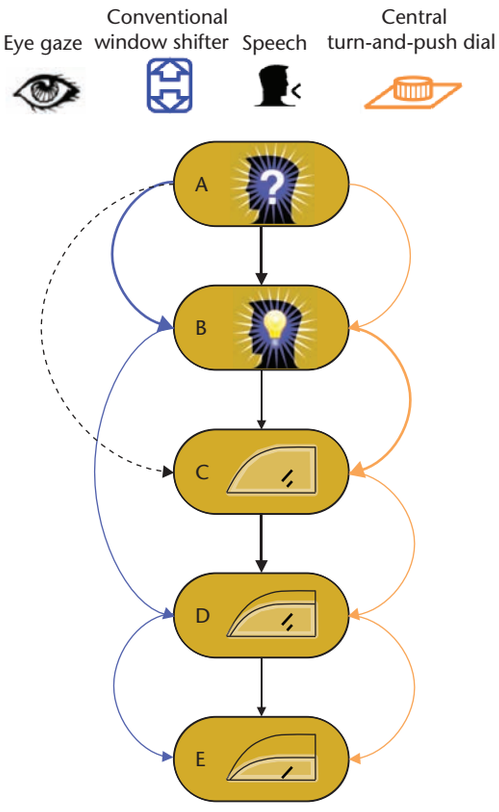

Figure 1. Directed acyclic graph representing a driver's intention and action to open the front-right window a little bit. We use this sample interaction to discuss the advantages and disadvantages of individual input modalities, various techniques for combining them, as well as extensions to multimodal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

Advantages and drawbacks

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, that is, 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 to pick the desired action. This process, might, 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 (C→D and D→E).

This gradual manipulation step (C→D) is the Achilles heel of speech-based interaction (the black edges in the graph in Figure 1). 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's 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 D→E, that is, 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 valid for a particular system state. This can be somewhat demanding as well, so we have drawn the relevant edge (A→B) using an intermediate line weight.

The dashed edge A→C represents the use of eye gaze as an implicit interaction modality (one that refers to naturally occurring user behavior or actions without requiring any explicit command[4]). An intelligent system (equipped with an eye tracker and a sophisticated user model) could infer the driver's intention to open the window from his or her gaze, and thus set the interaction context accordingly (bypassing node B in the graph). If the system takes action in this manner on the basis of an established belief about a user's intention, this is termed an inferred interaction.

Combining modalities

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

According to Oviatt, two or more modalities are temporally cascaded if sequenced in a particular order so partial information supplied by recognition of the earlier mode is able to constrain the interpretation of the later mode.[4] Suppose you say ''front-right window'' (A→B→C) and then immediately push the multifunction knob downward (C→D). 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 (A→B→D). If you then said ''more'' (D→E), 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 the temporally cascading modal systems 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 him- or herself in the USB media-player mode, the ''phone'' command could be issued by voice, after which a press of the menu key on the dashboard would bring up a phone-specific menu rather than a USB-specific menu.

We define redundant modalities as a special form of temporally cascaded modalities where each mode is available in each interaction step. The user can then freely pick the means that feels most comfortable to begin or continue 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

Figure 2. 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 (for example, "John") could comprise a multimodal interaction for initiating phone calls.[7]

throughout the industry, a steering wheel-mounted push-to-talk button initiates each voice command). While the organization of these UIs does impose a heavily hierarchical, step-wise interaction scheme, the user is given the freedom to 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 and 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.

The most elaborate form of multimodality is modality fusion.[4] Here, multiple modes play a part in a single interaction step. 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 combined to improve the overall recognition accuracy. It's apparent that with high levels of background noise and larger vocabulary sizes, this method 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 most compromised by dynamically changing lighting conditions. Depending on 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 called early fusion or late fusion, respectively.[4]

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

Design considerations

As discussed earlier, in-car UIs should be designed with a focus on highly efficient interactions and a minimum of driver distraction. Therefore it's important to accompany each stage of system design—from early prototypes to mature products—with comprehensive and well-designed user studies. Several methods can be applied to evaluate design choices and implementation parameters, starting with questionnaires (for example, the Driver Activity Load Index[8]) 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.[9]

Given speech- or multimodal-interface techniques could incubate at university or corporate research labs, where usability evaluations could be carried out quickly and inexpensively using low- or mid-fidelity simulators that support, for example, the lane-change task[10] 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.

Outlook for the future

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 takes 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 selecting one item from a list of n items, or gradually increasing or decreasing 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.

As mentioned previously, there is a dearth of standards that specifically relate to the operation of multimodal in-vehicle interfaces. However, there has been significant progress toward the standardization of multimodal interfaces in general. The Multimodal Interaction Working Group (see http://www.w3.org/2002/mmi/) focuses on markup languages and architectures that support the creation and consumption of multimodal (often voice and pen) websites. The HTML Speech Incubator group (see http://www.w3.org/2005/Incubator/htmlspeech/charter) 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 UI community should pay close attention to the output of these two standards bodies.

Few automotive UI designers would debate the advantages of modality redundancy in reducing cognitive load; it's an obvious advantage if the driver can avoid having to think about whether to proceed through an interaction using tactile or voice input, and can instead simply 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. One 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.[11] 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 previous Ford Sync example) and push-to-talk into a single step, reducing overall interaction time by approximately 40 percent versus a traditional design[11] and encouraging today's multitasking, hyper-connected driver to get back to actually driving. **MM**

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. Society of Automotive Engineers, SAE Recommended Practice: Navigation and Route Guidance Function Accessibility While Driving (SAE 2364), tech. report SAE 2364, Jan. 2000.
2. L. Garay-Vega et al., ''Evaluation of Different Speech and Touch Interfaces to In-Vehicle Music Retrieval Systems,'' *Accident Analysis & Prevention,* vol. 42, no. 3, 2010, pp. 913-920.
3. U. Gärtner, W. König, and T. Wittig, ''Evaluation of Manual vs. Speech Input When Using a Driver Information System in Real Traffic,'' *Proc. Int'l Driving Symp. Human Factors in Driver Assessment, Training, and Vehicle Design,* Transportation Research Board, 2001.
4. S. Oviatt, ''Multimodal Interfaces,'' T *he Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications,* 2nd ed., A. Sears, and J.A. Jacko, eds., CRC Press, 2007, pages 413-432.
5. M.J. Pitts et al., ''Assessing Subjective Response to Haptic Feedback in Automotive Touchscreens,'' *Proc. 1st Int'l Conf. Automotive User Interfaces and Interactive Vehicular Applications,* ACM Press, 2009, pp. 11-18.
6. S. Castronovo, A. Mahr, and C. Müller, ''Multimodal Dialog in the Car: Combining Speech and Turn-and-Push Dial to Control Comfort Functions,'' *Proc. Interspeech,* ISCA Press 2010, pp. 510-513.
7. C. Endres, T. Schwartz, and C. Müller, '''Geremin': 2D Microgestures For Drivers Based On Electric Field Sensing,'' *Proc. Int'l Conf. Intelligent User Interfaces,* ACM Press, 2011.

8. A. Pauzié, and G. Pachiaudi, ''Subjective Evaluation of the Mental Workload in the Driving Context,'' *Traffic and Transport Psychology: Theory and Application,* T. Rothengatter, and E. Carbonell Vaya, eds., Pergamon, 1997, pp. 173-82.
9. B. Mehler et al., ''Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers,'' *Transportation Research Record: J. Transportation Research Board,* vol. 2138, no. 1, 2009, pp. 6-12.
10. *ISO 26022:2010—Road Vehicles—Ergonomic Aspects of Transport Information and Control Systems—Simulated Lane Change Test to Assess In-Vehicle Secondary Task Demand,* ISO, 2010.
11. G. Weinberg et al., ''Contextual Push-to-Talk: Shortening Voice Dialogs to Improve Driving Performance,'' *Proc. 12th Int'l Conf. Human–Computer Interaction with Mobile Devices and Services,* ACM Press, 2010 pp. 113-122.

Contact author Christian Müller at christian.mueller@dfki.de.

Contact editor Anthony Vetro at avetro@merl.com.