Multimodal Interfaces for In-Vehicle Applications

Roman Vilimek, Thomas Hempel, and Birgit Otto

Siemens AG, Corporate Technology, User Interface Design
Otto-Hahn-Ring 6, 81730 Munich, Germany
{roman.vilimek.ext,thomas.hempel,birgit.otto}@siemens.com

Abstract. This paper identifies several factors that were observed as being crucial to the usability of multimodal in-vehicle applications – a multimodal system is not of value in itself. Focusing in particular on the typical combination of manual and voice control, this article describes important boundary conditions and discusses the concept of natural interaction.

Keywords: Multimodal, usability, driving, in-vehicle systems.

1 Motivation

The big-picture goal of interaction design is not accomplished only by enabling the users of a certain product to fulfill a task by using it. Rather, the focus should reside on a substantial facilitation of the interaction process between humans and technical solutions. In the majority of cases the versatile and characteristic abilities of human operators are widely ignored in the design of modern computer-based systems. Buxton [1] depicts this situation quite nicely by describing what a physical anthropologist in the far future might conclude when discovering a computer store of our time (p. 319): "My best guess is that we would be pictured as having a well-developed eye, a long right arm, uniform-length fingers and a 'low-fi' ear. But the dominating characteristic would be the prevalence of our visual system over our poorly developed manual dexterity."

This statement relates to the situation in the late 1980s, but it still has an unexpected topicality. However, with the advent of multimodal technologies and interaction techniques, a considerable amount of new solutions emerge that can reduce the extensive overuse of the human visual system in HCI. By involving the senses of touch and hearing, the heavy visual monitoring load of many tasks can be considerably reduced. Likewise, the activation of an action does no longer need to be carried out exclusively by pressing a button or turning a knob. For instance, gesture recognition systems allow for contact-free manual input and eye gaze tracking can be used as an alternative pointing mechanism. Speech recognition systems are the prevalent foundation of "eyes-free, hands-free" systems.

The trend to approach the challenge of making new and increasingly complex devices usable with multimodal interaction is particularly interesting, as not only universities but also researchers in the industrial environment spend increasing efforts in evaluating the potential of multimodality for product-level solutions. Research has shown that the advantages of multimodal interfaces are manifold, yet their usability

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 216–224, 2007.

does not always meet expectations. Several reasons account for this situation. In most cases, the technical realization is given far more attention than the users and their interaction behavior, their context of use or their preferences. This leads to systems which do not provide modalities really suited for the task. That may be acceptable for a proof-of-concept demo, but is clearly not adequate for an end-user product. Furthermore the user's willingness to accept the product at face value is frequently overestimated. If a new method of input does not work almost perfectly, users will soon get annoyed and do not act multimodally at all. Tests in our usability lab have shown, that high and stable speech recognition rates are necessary (>90%) for novice users of everyday products. And these requirements have to be met in everyday contexts – not only in a sound-optimized environment! Additionally, many multimodal interaction concepts are based on misconceptions about how users construct their multimodal language [2] and what "natural interaction" with a technical system should look like.

Taken together, these circumstances seriously reduce the expected positive effects of multimodality in practice. The goal of this paper is to summarize some relevant aspects of key factors for successful multimodal design of advanced in-vehicle interfaces. The selection is based on our experience in an applied industrial research environment within a user-centered design process and does not claim to be exhaustive.

2 Context of Use: Driving and In-Vehicle Interfaces

ISO9241-11 [3], an ISO norm giving guidance on usability, requires explicitly to consider the context in which a product will be used. The relevant characteristics of users (2.1), tasks and environment (2.2) and the available equipment (2.3) need to be described. Using in-vehicle interfaces while driving is usually embedded in a multiple task situation. Controlling the vehicle safely must be regarded as the primary task of the driver. Thus, the usability of infotainment, navigation or communication systems inside cars refers not only to the quality of the interaction concept itself. These systems have to be built in a way that optimizes time-sharing and draws as few attentional resources as possible off the driving task. The contribution of multimodality needs to be evaluated in respect to these parameters.

2.1 Users

There are only a few limiting factors that allow us to narrow the user group. Drivers must own a license and thus they have shown to be able to drive according to the road traffic regulations. But still the group is very heterogeneous. The age range goes anywhere from 16 or 18 to over 70. A significant part of them are seldom users, changing users and non-professional users, which have to be represented in usability tests. Quite interestingly older drivers seem to benefit more from multimodal displays than younger people [4]. The limited attentional resources of elderly users can be partially compensated by multimodality.

2.2 Tasks and Environment

Even driving a vehicle itself is not just a single task. Well-established models (e.g. [5]) depict it as a hierarchical combination of activities at three levels which differ in

respect to temporal aspects and conscious attentional demands. The topmost strategic level consists of general planning activities as well as navigation (route planning) and includes knowledge-based processes and decision making. On the maneuvering level people follow complex short-term objectives like overtaking, lane changing, monitoring the own car movements and observing the actions other road users. On the bottom level of the hierarchy, the operational level, basic tasks have to be fulfilled including steering, lane keeping, gear-shifting, accelerating or slowing down the car. These levels are not independent; the higher levels provide information for the lower levels. They pose different demands on the driver with a higher amount of mental demands on the higher levels and in increased temporal frequency of the relevant activities on the lower levels [6]. Thus, these levels have to be regarded as elements of a continuum.

This model delivers valuable information for the design of in-vehicle systems which are not directly related to driving. Any additional task must be created in a way that minimizes conflict with any of these levels. To complicate matters further, more and more driver information systems, comfort functions, communication and mobile office functions and the integration of nomad devices turn out to be severe sources of distraction. Multimodal interface design may help to re-allocate relevant resources to the driving task. About 90% of the relevant information is perceived visually [7] and the manual requirements of steering on the lower levels are relatively high, as long as they are not automated. Thus, first of all interfaces for on-board comfort functions have to minimize the amount of required visual attention. Furthermore they must support short manual interaction steps and an ergonomic posture. Finally, the cognitive aspect may not be underestimated. Using in-vehicle applications must not lead to high levels of mental workload or induce cognitive distraction. Research results show, that multimodal interfaces have a high potential to reduce the mental and physical demands in multiple task situations by improving the time-sharing between primary and secondary task (for an overview see [8]).

2.3 Equipment

Though voice-actuation technology has proven to successfully keep the driver's eyes on the road and the hands on the steering wheel, manual controls will not disappear completely. Ashley [9] comes to the conclusion, that there will be fewer controls and that they will morph into a flexible new form. And indeed there is a general trend among leading car manufacturers to rely on a menu-based interaction concept with a central display at the top of the center console and single manual input device between the front seats. The placement of the display allows for a peripheral detection of traffic events while at the same time the driver is able to maintain a relaxed body posture while activating the desired functions. It is important to keep this configuration in mind when assessing the usability of multimodal solutions as the have to fit into this context. Considering the availability of a central display, the speech dialog concept can make use of the "say what you see" strategy [10] to inform novice users about valid commands without time-consuming help dialogs. Haptic or auditory feedback can improve the interaction with the central input device like and reduce visual distraction like for example the force feedback of BMW's iDrive controller [9].

3 Characteristics of Multimodal Interfaces

A huge number of different opinions exist on the properties of a multimodal interface. Different researchers mean different things when talking about multimodality, probably because of the interdisciplinary nature of the field [11]. It is not within the scope of this paper to define all relevant terms. However, considering the given situation in research it seems necessary to clarify at least some basics to narrow down the subject.

The European Telecommunications Standards Institute [12] defines *multimodal* as an "adjective that indicates that at least one of the directions of a two-way communication uses two sensory modalities (vision, touch, hearing, olfaction, speech, gestures, etc.)" In this sense, *multimodality* is a "property of a user interface in which: a) more than one sensory is available for the channel (e.g. output can be visual or auditory); or b) within a channel, a particular piece of information is represented in more than one sensory modality (e.g. the command to open a file can be spoken or typed)." The term sensory is used in wide sense here, meaning human senses as well as sensory capabilities of a technical system.

A key aspect of a multimodal system is to analyze how input or output modalities can be combined. Martin [13, 14] proposes a typology to study and design multimodal systems. He differentiates between the following six "types of cooperation":

- Equivalence: Several modalities can be used to accomplish the same task, i.e. they can be used alternatively.
- Specialization: A certain piece of information can only be conveyed in a specially designated modality. This specialization is not necessarily absolute: Sounds, for example, can be specialized for error messages, but may also be used to signalize some other important events.
- Redundancy: The same piece of information is transmitted by several modalities at the same time (e.g., lip movements and speech in input, redundant combinations of sound and graphics in output). Redundancy helps to improve recognition accuracy.
- Complementarity: The complete information of a communicative act is distributed across several modalities. For instance, gestures and speech in man-machine interaction typically contribute different and complementary semantic information [15].
- Transfer: Information generated in one modality is used by another modality, i.e. the interaction process is transferred to another modality-dependent discourse level. Transfer can also be used to improve the recognition process. Contrary to redundancy, the modalities combined by transfer are not naturally associated.
- Concurrency: Several independent types of information are conveyed by several modalities at the same time, which can speed up the interaction process.

Martin points out that redundancy and complementarity imply a fusion of signals, an integration of information derived from parallel input modes. Multimodal fusion is generally considered to be the supreme discipline of multimodal interaction design. However, it is also the most complex and cost-intensive design option – and may lead to quite error prone systems in real life because the testing effort is drastically increased. Of course so-called mutual disambiguation can lead to a recovery from unimodal recognition errors, but this works only with redundant signals. Thus, great care

has to be taken to identify whether there is a clear benefit of modality fusion within the use scenario of a product or whether a far simpler multimodal system without fusion will suffice.

One further distinction should be reported here because of its implication for cognitive ergonomics as well as for usability. Oviatt [16] differentiates between active and passive input modes. Active modes are deployed intentionally by the user in form of an explicit command (e.g., a voice command). Passive input modes refer to spontaneous automatic and unintentional actions or behavior of the user (e.g., facial expressions or lip movements) which are passively monitored by the system. No explicit command is issued by the user and thus no cognitive effort is necessary. A quite similar idea is brought forward by Nielsen [17] who suggests non-command user interfaces which do no longer rely on an explicit dialog between the user and a computer. Rather the system has to infer the user intentions by interpreting user actions. The integration of passive modalities to increase recognition quality surely improves the overall system quality, but non-command interfaces are a two-edged sword: On the one hand they can lower the consumption of central cognitive resources, on the other the risk of over-adaptation arises. This can lead to substantial irritation of the driver.

4 Designing Multimodal In-Vehicle Applications

The benefits of successful multimodal design are quite obvious and have been demonstrated in various research and application domains. According to Oviatt and colleagues [18], who summarize some of the most important aspects in a review paper, multimodal UIs are far more flexible. A single modality does not permit the user to interact effectively across all tasks and environments while several modalities enable the user to switch to a better suited one if necessary. The first part of this section will try to show how this can be achieved for voice and manual controlled in-vehicle applications. A further frequently used argument is that multimodal systems are easier to learn and more natural, as multimodal interaction concept can mimic man-man-communication. The second part of this section tries to show that natural is not always equivalent to usable and that natural interaction does not necessarily imply human-like communication.

4.1 Combining Manual and Voice Control

Among the available technologies to enhance unimodal manual control by building a multimodal interface, speech input is the most robust and advanced option. Bengler [19] assumes, that any form of multimodality in the in-vehicle context will always imply the integration of speech recognition. Thus, one of the most prominent questions is how to combine voice and manual control so that their individual benefits can take effect. If for instance the hands cannot be taken off the steering wheel on a wet road or while driving at high speed, speech commands ensures the availability of comfort functions. Likewise, manual input may substitute speech control if it is too noisy for successful recognition. To take full advantage of the flexibility offered by multimodal voice and manual input, both interface components have to be completely equivalent. For any given task, both variants must provide usable solutions for task completion.

How can this be done? One solution is to design manual and voice input independently: A powerful speech dialog system (SDS) may enable the user to accomplish a task completely without prior knowledge of the system menus used for manual interaction. However, using the auditory interface poses high demands on the driver's working memory. He has to listen to the available options and keep the state of the dialog in mind while interrupting it for difficult driving maneuvers. The SDS has to be able to deal with long pauses by the user which typically occur in city traffic. Furthermore the user cannot easily transfer acquired knowledge from manual interaction, e.g. concerning menu structures. Designing the speech interface independently also makes it more difficult to meet the usability requirement of consistency and to ensure that really all functions available in the manual interface are incorporated in the speech interface.

Another way is to design according to the "say what you see" principle [10]: Users can say any command that is visible in a menu or dialog step on the central display. Thus, the manual and speech interface can be completely parallel. Given that currently most people still prefer the manual interface to start with the exploration of a new system, they can form a mental representation of the system structure which will also allow them to interact verbally more easily. This learning process can be substantially enhanced if valid speech commands are specially marked on the GUI (e.g., by font or color). As users understand this principle quickly, they start using expert options like talk-ahead even after rather short-time experience with the system [20].

A key factor for the success of multimodal design is user acceptance. Based on our experience, most people still do not feel very comfortable interaction with a system using voice commands, especially when other people are present. But if the interaction is restricted to very brief commands from the user and the whole process can be done without interminable turn-taking dialogs, the users are more willing to operate by voice. Furthermore, users generally prefer to issue terse, goal-directed commands rather than engage in natural language dialogs when using in-car systems [21]. Providing them with a simple vocabulary by designing according to the "say what you see" principle seems to be exactly what they need.

4.2 Natural Interaction

Wouldn't it be much easier if all efforts were undertaken to implement natural language systems in cars? If the users were free to issue commands in their own way, long clarification dialogs would not be necessary either. But the often claimed equivalency between naturalness and ease is not as valid as it seems from a psychological point of view. And from a technological point of view crucial prerequisites will still take a long time to solve. Heisterkamp [22] emphasizes that fully conversational systems would need to have the full human understanding capability, a profound expertise on the functions of an application and an extraordinary understanding of what the user really intends with a certain speech command. He points out that even if these problems could be solved there are inherent problems in people's communication behavior that cannot be solved by technology. A large number of recognition errors will result, with people not exactly saying what they want or not providing the information that is needed by the system. This assumption is supported by findings of Oviatt [23].

She has shown that the utterances of users get increasingly unstructured with growing sentence length. Longer sentences in natural language are furthermore accompanied by a huge number of hesitations, self-corrections, interruptions and repetitions, which are difficult to handle. This holds even for man-man communication. Additionally, the quality of speech production is substantially reduced in dual-task situations [24]. Thus, for usability reasons it makes sense to provide the user with an interface that forces short and clear-cut speech commands. This will help the user to formulate an understandable command and this in turn increases the probability of successful interaction.

Some people argue that naturalness is the basis for intuitive interaction. But there are many cases in everyday life where quite unnatural actions are absolutely intuitive – because there are standards and conventions. Heisterkamp [22] comes up with a very nice example: Activating a switch on the wall to turn on the light at the ceiling is not natural at all. Yet, the first thing someone will do when entering a dark room is to search for the switch beside the door. According to Heisterkamp, the key to success are conventions, which have to be omnipresent and easy to learn. If we succeed in finding conventions for multimodal speech systems, we will be able to create very intuitive interaction mechanisms. The "say what you see" strategy can be part of such a convention for multimodal in-vehicle interfaces. It also provides the users with an easy to learn structure that helps them to find the right words.

5 Conclusion

In this paper we identified several key factors for the usability of multimodal in-vehicle applications. These aspects may seem trivial at first, but they are worth considering as they are neglected far too often in practical research. First, a profound analysis of the context of use helps to identify the goals and potential benefit of multimodal interfaces. Second, a clear understanding of the different types of multimodality is necessary to find an optimal combination of single modalities for a given task. Third, an elaborate understanding of the intended characteristics of a multimodal system is essential: Intuitive and easy-to-use interfaces are not necessarily achieved by making the communication between man and machine as "natural" (i.e. human-like) as possible. Considering speech-based interaction, clear-cut and non-ambiguous conventions are needed most urgently.

To combine speech and manual input for multimodal in-vehicle systems, we recommend designing both input modes in parallel, thus allowing for transfer effects in learning. The easy-to-learn "say what you see" strategy is a technique in speech dialog design that structures the user's input and narrows the vocabulary at the same time and may form the basis of a general convention. This does not mean that command-based interaction is from a usability point of view generally superior to natural language. But considering the outlined technological and user-dependent difficulties, a simple command-and-control concept following universal conventions should form the basis of any speech system as a fallback. Thus, before engaging in more complex natural interaction concepts, we have to establish these conventions first.

References

1. Buxton, W.: There's More to Interaction Than Meets the Eye: Some Issues in Manual Input. In: Norman, D.A., Draper, S.W. (eds.) User Centered System Design: New Perspectives on Human-Computer Interaction, pp. 319–337. Lawrence Erlbaum Associates, Hillsdale, NJ (1986)
2. Oviatt, S.L.: Ten Myths of Multimodal Interaction. Communications of the ACM 42, 74–81 (1999)
3. ISO 9241-11 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability. International Organization for Standardization, Geneva, Switzerland (1998)
4. Liu, Y.C.: Comparative Study of the Effects of Auditory, Visual and Multimodality Displays on Driver's Performance in Advanced Traveller Information Systems. Ergonomics 44, 425–442 (2001)
5. Michon, J.A.: A Critical View on Driver Behavior Models: What Do We Know, What Should We Do? In: Evans, L., Schwing, R. (eds.) Human Behavior and Traffic Safety, pp. 485–520. Plenum Press, New York (1985)
6. Reichart, G., Haller, R.: Mehr aktive Sicherheit durch neue Systeme für Fahrzeug und Straßenverkehr. In: Fastenmeier, W. (ed.): Autofahrer und Verkehrssituation. Neue Wege zur Bewertung von Sicherheit und Zuverlässigkeit moderner Straßenverkehrssysteme. TÜV Rheinland, Köln, pp. 199–215 (1995)
7. Hills, B.L.: Vision, Visibility, and Perception in Driving. Perception 9, 183–216 (1980)
8. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance. Prentice Hall, Upper Saddle River, NJ (2000)
9. Ashley, S.: Simplifying Controls. Automotive Engineering International March 2001, pp. 123-126 (2001)
10. Yankelovich, N.: How Do Users Know What to Say? ACM Interactions 3, 32–43 (1996)
11. Benoît, J., Martin, C., Pelachaud, C., Schomaker, L., Suhm, B.: Audio-Visual and Multimodal Speech-Based Systems. In: Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation, pp. 102–203. Kluwer Academic Publishers, Boston (2000)
12. ETSI EG 202 191: Human Factors (HF); Multimodal Interaction, Communication and Navigation Guidelines. ETSI. Sophia-Antipolis Cedex, France (2003) Retrieved December 10, 2006, from http://docbox.etsi.org/EC_Files/EC_Files/eg_202191v010101p.pdf
13. Martin, J.-C.: Types of Cooperation and Referenceable Objects: Implications on Annotation Schemas for Multimodal Language Resources. In: LREC 2000 pre-conference workshop, Athens, Greece (1998)
14. Martin, J.-C.: Towards Intelligent Cooperation between Modalities: The Example of a System Enabling Multimodal Interaction with a Map. In: IJCAI'97 workshop on intelligent multimodal systems, Nagoya, Japan (1997)
15. Oviatt, S.L., DeAngeli, A., Kuhn, K.: Integration and Synchronization of Input Modes During Human-Computer Interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 415–422. ACM Press, New York (1997)
16. Oviatt, S.L.: Multimodal Interfaces. In: Jacko, J.A., Sears, A. (eds.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 286–304. Lawrence Erlbaum Associates, Mahwah, NJ (2003)
17. Nielsen, J.: Noncommand User Interfaces. Communications of the ACM 36, 83–99 (1993)

18. Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D.: Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. Human-Computer Interaction 15, 263–322 (2000)
19. Bengler, K.: Aspekte der multimodalen Bedienung und Anzeige im Automobil. In: Jürgensohn, T., Timpe, K.P. (eds.) Kraftfahrzeugführung, pp. 195–205. Springer, Berlin (2001)
20. Vilimek, R.: Concatenation of Voice Commands Increases Input Efficiency. In: Proceedings of Human-Computer Interaction International 2005, Lawrence Erlbaum Associates, Mahwah, NJ (2005)
21. Graham, R., Aldridge, L., Carter, C., Lansdown, T.C.: The Design of In-Car Speech Recognition Interfaces for Usability and User Acceptance. In: Harris, D. (ed.) Engineering Psychology and Cognitive Ergonomics: Job Design, Product Design and Human-Computer Interaction, Ashgate, Aldershot, vol. 4, pp. 313–320 (1999)
22. Heisterkamp, P.: Do Not Attempt to Light with Match! Some Thoughts on Progress and Research Goals in Spoken Dialog Systems. In: Eurospeech 2003. ISCA, Switzerland, pp. 2897–2900 (2003)
23. Oviatt, S.L.: Interface Techniques for Minimizing Disfluent Input to Spoken Language Systems. In: Proceedings of the Sigchi Conference on Human Factors in Computing Systems: Celebrating Interdependence (CHI'94), pp. 205–210. ACM Press, New York (1994)
24. Baber, C., Noyes, J.: Automatic Speech Recognition in Adverse Environments. Human Factors 38, 142–155 (1996)