Potentials of Multimodal Interaction in the Vehicle - More Effective, More Natural, More Persuasive

Florian Roider

BMW Group Research and Technology, Munich, Germany
University of Bamberg, Bamberg, Germany
florian.roider@bmw.de

Abstract. Multimodal automotive interaction concepts enable drivers to use a broader bandwidth of interaction possibilities by taking multiple modalities into account. This allows not only to cope the increasing number and complexity of functions in modern vehicles, but has also the potential to increase naturalness and persuasiveness of the interaction. However, multimodal interaction concepts in the vehicle vary greatly in their ways of combining individual modalities. This paper summarizes different approaches based on concepts presented in literature and discusses them regarding their benefits and limitations. Considerations for overcoming these limitations in future interaction concepts are made in order to create a more effective and more natural way of interaction for the driver.

Keywords: Multimodal interaction, natural user interface, automotive user interfaces

1 Introduction

Multimodal interaction is one of the most promising approaches to cope with the growing number of functions of increased complexity in the vehicle. Several concept cars that have been presented in the recent past include multimodal interaction concepts to replace conventional input devices, such as the BMW i8 Spyder[1] and the Porsche Mission E[2]. The underlying idea of such concepts is to enable the driver to use natural modalities such as speech, gesturing and gazes to interact with the vehicle, in order to make interaction easier, more effective and less distracting from the driving task. However, there seems to be little consensus in literature of how multimodal interaction is concretely applied best to achieve these goals.

As the main modality in interhuman communication, interaction based on speech recognition is a common feature in modern vehicles. However, speech control still fights for acceptance among drivers. Problems include technical issues, such as

[1] https://www.youtube.com/watch?v=09yEQfa4Bqw
[2] https://www.youtube.com/watch?v=fBJ0pXOSUug

errors in recognition, waiting times while input is processed and the difficulties of natural language understanding. Furthermore, the driver may experience an increased cognitive load for formulating appropriate commands [8]. This is due to the fact that purely verbal information does not suffice to provide a complete picture of human communication capability [12]. Multimodality allows to overcome such limitations by combining input from various modalities.

2 Interaction in the Vehicle

Many interaction concepts that aim to control vehicle functions such as mirrors, windows, or ventilation distinguish between two phases in the interaction process [7, 3]. In a first phase, the selection of the interaction object takes place, determining the context of the interaction, e.g. identifying a window that shall be opened. The second phase describes the manipulation of this object or the function that is applied to the selected context e.g. opening the selected window. This differentiation is also applicable for interaction with more abstract objects in the vehicle, such as the integrated navigation system or multimedia controls. Although those entities usually do not have visual representatives except the display they are shown on, they can still be selected. Functions can then be applied based on this selected context, such as switching to an alternative route or to a different radio station.
The greatest costs in vehicle interaction are currently caused by the phase of object identification [4]. Multimodal interaction has the potential to overcome these costs by offering a more intuitive way to select the context, for example by just looking [9] and/or pointing [10] at objects the driver wants to interact with. Thereby, the costs for identifying objects and remembering commands to activate them could be reduced and even overcome for certain cases.

3 Combining Modalities

Oviatt defined multimodal systems 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" [5]. Based on this definition, the term has been interpreted in different ways.
The following sections provide a short overview over a differentiation between possibilities how to combine modalities based on [4]. Figure 1 illustrates the differences between three approaches using the example of combining gaze and gesture input.

Temporally Cascaded Modalities: *Temporally cascaded modalities* are combined in a particular temporal order. Input using one modality in an earlier interaction step is constraining the interpretation of input using another modality in a later step [5]. However, in this case each interaction step is connected to a specific modality [4]. Plfeging et al. present a concept that combines speech

Fig. 1. Differences between temporally cascaded, redundant and fused modalities using the example of gaze and gesture input in the object selection phase.

and touch gestures in a temporally cascaded manner [7]. Objects in the vehicle can be selected by speech and in a second step modified by using predefined touch gestures. Rümelin et al. use pointing gestures to identify objects outside the vehicle to allow further interaction in subsequent steps [10]. Poitschke et al. present a concept that allows drivers to control multiple displays with the same steering wheel button based on the drivers gaze [9].

Redundant Modalities According to Müller et al. *redundant modalities* are a special form of temporally cascaded modalities [4]. Instead of being bound to fixed assignments of one modality to an interaction step, drivers have the choice to pick a modality whichever they consider best suited for their needs. This results in a more flexible way of interacting with the vehicle which might help to reduce driver distraction in various traffic situations. Yang et al. [13] present a concept that allows full control of typical automotive domains like radio and air conditioning via voice or predefined finger gestures on the wheel alternatively.

Fused Modalities: In *fused modalities* multiple input modes play a part in a single interaction step. The information from different modalities is fused in order to clarify the intended interaction [4]. This enables users to make use of a broader communication bandwidth and therefore bears the greatest potential to simplify the input of complex information.
Two modalities can transport redundant information by using another input mode. For example, speech recognition can be combined with facial expressions in order to improve recognition accuracy in loud environments, by enhancing the auditory input with a visual one [11].
Modality fusion can also be interpreted by transmitting complementary information with each medium. Ideally, each modality transports only the part of the message that it is best suited for. Bolts "put-that-there" [1] is an example for this type of fusion of gestures and speech. Spacial information is communicated by a pointing gesture, whereas the actual action is uttered using speech. Deictic references "that" and "there" provide a semantic and temporal connection between both modalities.

4 Natural Interaction

When communicating face-to-face with other human beings we make use of a variety of different communication channels at the same time to precisely

describe what we want to express. Natural human-computer interaction fuses those interaction channels that are typically used in inter-human communication, such as speech and hand gestures. Besides those intentionally used modalities, we provide a variety of unintentionally expressed information, such as gaze, facial expressions and body language that contribute to clarifying our intents. In literature this difference is represented by distinguishing active and passive input modes [5]. Active modes are intentionally expressed as a command towards the system, whereas passive input modes describe natural occurring behavior, that can be detected by the system by monitoring the user. Oviatt describes the combination of at least one active mode with at least one passive input mode as blended interaction [5]. The concept of fused multimodal interaction allows to incorporate active and passive input from a multitude of channels. Accordingly, it is frequently put into context with natural interaction as a means to provide a more natural and more intuitive user experience.

5 Multimodality and Persuasion

While multimodality can enhance user interaction with the system, it also allows the system to have a greater influence on the user. Natural interaction that incorporates passive information from speech, gaze and body language allows to better assess the driver's affective state and might therefore help to create a more human-like communication with the vehicle. The ability to sense input from a variety of different channels about the user's state and react accordingly can increase naturalness and persuasiveness of the system [6]. Similarities with interhuman communication also promote the acceptance of proactive suggestions by the system. Allowing to trigger interactions in the right moment is crucial for influencing user behavior [2].
On the other side, persuasion can support the efficient use of multimodality. Despite the flexibility of multimodal systems, sometimes it makes sense to suggest adequate modalities, based on switch costs, suitability of modality combinations and situation depended driver loads. For example, if the driver wants to interact with the vehicle while talking to a co-driver, speech might be disturbing the conversation and gaze and gesture based interaction should be offered instead. Besides the identification of suited modalities, the main challenge here consists in persuading the driver to use the most adequate modality, while still retaining full flexibility and not imposing rules.

6 Discussion

Besides the need for adequate sensors to capture all relevant information provided by the driver, the greatest challenge lies in the interpretation of input from different channels in order to conclude the driver's intent. This is due to the fact that the relations of multiple modalities in human-human communications cannot be generalized. Interplay between gesture and speech for example is highly adaptive to various situations [12]. Information might be transmitted primarily

on either of the channels according to the context. For successful integration of multiple modalities systems have to understand what information is transmitted within each modality. The great variability over possible traffic situations and individual users demand for highly flexible solutions.
The choice of suited modalities is another important point. Speech, gestures and gaze are frequently incorporated in the automotive domain. The majority of concepts in literature make fixed assignments of modalities to single interaction phases. However, this does not provide sufficient flexibility to allow a natural interaction for a wide range of different functions and situations. Drivers should not be bound to certain modalities, since it is difficult to determine a fixed set of modalities for all interactions in the car. There are functions that might be best operated with a combination of gaze and gestures, whereas in other cases a combination of speech and gestures, or a simple hand gesture only might be the most intuitive way to achieve the desired effect. Accordingly, multimodality should not only be supported between interaction steps but also within single interaction phases. Redundant and fused modalities follow this idea by enabling drivers to use alternative modalities and modality combinations. Transmitted information has to be decoupled from the transmitting modalities.

7 Future Research

Future research will focus on the exploration of the interplay of different modality combinations on a more general level. In this course, analysis of driver loads for switching between input modes and for simultaneous application of multiple modalities is planned. Results could be used to determine most adequate modalities and to foreground interaction with those modalities. This is closely connected to the expressiveness of individual modalities for specific in-vehicle use cases, which will be part of future research.
Only few automotive interaction concepts combine speech with co-verbal gestures, such as pointing. Since most of the infotainment functionality in modern cars is bound to a relatively small central information display, the potential for context selection by pointing on content on the screen is limited. This may change with the development of further growing displays, which goes hand in hand with displays moving further away from the driver. Experiments are planned to investigate the suitability of pointing gestures in this context more closely.
Following the idea of multimodal object selection, it makes sense to shift the manipulation phase also towards the object in order to allow a more direct manipulation. Drivers do not have to search for a dedicated input device or an option hidden in a menu hierarchy. Research is needed how to make possible interactions clear for the user and also how to provide adequate feedback during interaction.
Another topic is the influence of multimodal interaction on the interaction with a proactive agent as a communication partner. In this course, more natural interaction with the vehicle and enhanced awareness about the driver's state shall

be utilized to increase acceptance and persuasiveness of the agent.

References

1. Richard a Bolt. Put-that-there: Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques - SIGGRAPH '80*, volume 14, pages 262–270. ACM, 1980.
2. BJ Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology - Persuasive '09*, page 1, New York, New York, USA, apr 2009. ACM.
3. Monika Mitrevska, Mohammad Mehdi Moniri, Robert Nesselrath, Tim Schwartz, Michael Feld, Yannick Korber, Matthieu Deru, and Christian Muller. SiAM - Situation-Adaptive Multimodal Interaction for Innovative Mobility Concepts of the Future. In *Proceedings of the International Conference on Intelligent Environments - IE '15*, pages 180–183. IEEE, 2015.
4. Christian Müller, Garrett Weinberg, and Anthony Vetro. Multimodal input in the car, today and tomorrow. *IEEE Multimedia*, 18(1):98–103, 2011.
5. Sharon Oviatt. Multimodal Interfaces. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, 14:405–430, 2012.
6. Maja Pantic and Leon J M Rothkrantz. Toward an Affect-Sensitive Multimodal Human Computer Interaction. In *Proceedings of the IEEE*, volume 91, pages 1370–1390, 2003.
7. Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. Multimodal interaction in the car - combining speech and gestures on the steering wheel. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12*, number c, pages 155–162. ACM, 2012.
8. C.a. Carl a. Pickering, K.J. Keith J. Burnham, and Michael J. M.J. Richardson. A review of automotive human machine interface technologies and techniques to reduce driver distraction. In *System Safety, 2007 2nd Institution of Engineering and Technology International Conference on*, pages 223–228. IEEE, 2007.
9. Tony Poitschke, Florian Laquai, Stilyan Stamboliev, and Gerhard Rigoll. Gaze-based interaction on multiple displays in an automotive environment. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 543–548. IEEE, 2011.
10. Sonja Rümelin, Chadly Marouane, and Andreas Butz. Free-hand pointing for identification and interaction with distant objects. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, pages 40–47, New York, New York, USA, oct 2013. ACM.
11. Tevfik Metin Sezgin, Ian Davies, and Peter Robinson. Multimodal Inference for Driver-Vehicle Interaction. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI '09*, pages 193–197. ACM, 2009.
12. Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, feb 2014.
13. Seungmin Yang and Younghwan Pan. A Study on Methods of Multimodal Interaction. In *HCI International 2014 - Posters Extended Abstracts*, pages 484–489. Springer International Publishing, 2014.