Natural Multimodal Interaction in the Car

Generating Design Support for Speech, Gesture, and Gaze Interaction while Driving

Inaugural-Dissertation

an der Fakultät für Wirtschaftsinformatik und Angewandte Informatik

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Florian Roider

aus München

Bamberg, den 09.10.2020

URN: urn:nbn:de:bvb:473-irb-518266
DOI: https://doi.org/10.20378/irb-51826

Tag der mündlichen Prüfung: 10.05.2021

Dekan/Dekanin: Universitätsprofessor/-in Dr. Tom Gross

Betreuer/-in: Universitätsprofessor/-in Dr. Tom Gross

Weitere/r Gutachter/-in: Universitätsprofessor/-in Dr. Diedrich Wolter

Abstract

Driving a modern car is more than just maneuvering the vehicle on the road. At the same time, drivers want to listen to music, operate the navigation system, compose, and read messages and more. Future cars are turning from simple means for transportation into smart devices on wheels. This trend will continue in the next years together with the advent of automated vehicles. However, technical challenges, legal regulations, and high costs slow down the penetration of automated vehicles. For this reason, a great majority of people will still be driving manually at least for the next decade. Consequently, it must be ensured that all the features of novel infotainment systems can be used easily, efficiently without distracting the driver from the task of driving and still provide a high user experience.

A promising approach to cope with this challenge is multimodal in-car interaction. Multimodal interaction basically describes the combination of different input and output modalities for driver-vehicle interaction. Research has pointed out the potential to create a more flexible, efficient, and robust interaction. In addition to that, the integration of natural interaction modalities such as speech, gestures and gaze, the communication with the car could increase the naturalness of the interaction.

Based on these advantages, the researcher community in the field of automotive user interfaces has produced several interesting concepts for multimodal interaction in vehicles. The problem is that the resulting insights and recommendations are often easily applicable in the design process of other concepts because they too concrete or very abstract. At the same time, concepts focus on different aspects. Some aim to reduce distraction while others want to increase efficiency or provide a better user experience. This makes it difficult to give overarching recommendations on how to combine natural input modalities while driving. As a consequence, interaction designers of in-vehicle systems are lacking adequate design support that enables them to transfer existing knowledge about the design of multimodal in-vehicle applications to their own concepts.

This thesis addresses this gap by providing empirically validated design support for multimodal in-vehicle applications. It starts with a review of existing design support for automotive and multimodal applications. Based on that we report a series of user experiments that investigate various aspects of multimodal in-vehicle interaction with more than 200 participants in lab setups and driving simulators. During these experiments, we assessed the potentials of multimodality while driving, explored how user interfaces can support speech and gestures, and evaluated novel interaction techniques. The insights from these experiments extend existing knowledge from literature in order to create the first pattern collection for multimodal natural in-vehicle interaction. The collection contains 15 patterns that describe solutions for reoccurring problems when combining natural input with speech, gestures, or gaze in the car in a structured way. Finally, we present a prototype of an in-vehicle information system, which demonstrates the application of the proposed patterns and evaluate it in a driving-simulator experiment.

This work contributes to field of automotive user interfaces in three ways. First, it presents the first pattern collection for multimodal natural in-vehicle interaction. Second, it illustrates and evaluates interaction techniques that combine speech and gestures with gaze input. Third, it provides empirical results of a series of user experiments that show the effects of multimodal natural interaction on different factors such as driving performance, glance behavior, interaction efficiency, and user experience.

Zusammenfassung

Ein modernes Auto zu fahren ist mehr als nur die reine Steuerung des Fahrzeugs auf der Straße. Gleichzeitig möchten die Fahrer Musik hören, das Navigationssystem bedienen, Nachrichten verfassen und lesen und vieles mehr. Die Autos der Zukunft entwickeln sich von einfachen Transportmitteln zu intelligenten Geräten auf Rädern. Dieser Trend wird sich auch in den nächsten Jahren fortsetzen, zusammen mit der Entwicklung automatisierter Fahrzeuge. Allerdings verlangsamen technische Herausforderungen, gesetzliche Vorschriften und hohe Kosten die Verbreitung automatisierter Fahrzeuge. Es ist daher anzunehmen, dass eine große Mehrheit der Menschen auch im nächsten Jahrzehnt noch manuell fahren wird. Folglich muss sichergestellt werden, dass sämtliche Funktionen neuartiger Infotainment-Systeme einfach und effizient genutzt werden können, ohne den Fahrer vom Fahren abzulenken, und dennoch eine hohe Benutzerfreundlichkeit bieten.

Ein vielversprechender Ansatz zur Bewältigung dieser Herausforderung ist die multimodale Interaktion. Im Automotive Kontext beschreibt multimodale Interaktion die Kombination verschiedener Ein- und Ausgabemodalitäten zur Interaktion zwischen Fahrer und Fahrzeug. Forschungsarbeiten auf dem Gebiet der Human-Computer Interaction haben auf das Potenzial zur Schaffung einer flexibleren, effizienteren und robusteren Interaktion aufgezeigt. Darüber hinaus könnte die Integration natürlicher Interaktionsmodalitäten wie Sprache, Gestik und Blick, die Einfachheit und Natürlichkeit der Kommunikation mit dem Auto erhöhen.

Basierend auf diesen Vorteilen hat die Forschung auf dem Gebiet der automobilen Benutzerschnittstellen eine Reihe von interessanten Konzepten für die Anwendung multimodaler Interaktion im Fahrzeug hervorgebracht. Die daraus resultierenden Erkenntnisse und Empfehlungen sind allerdings oft nicht leicht in den Designprozess anderer Konzepte zu integrieren, weil sie zu konkret oder sehr abstrakt sind. Gleichzeitig konzentrieren sich die Konzepte auf verschiedene Aspekte. Einige zielen darauf ab, die Ablenkung zu verringern, während andere die Effizienz erhöhen oder eine bessere Benutzererfahrung bieten wollen. Dies macht es schwierig, übergreifende Empfehlungen zu geben, wie natürliche Eingabemodalitäten während der Fahrt kombiniert werden können. Es fehlt eine angemessene Form der Designunterstützung für Interaktionsdesigner, welche es diesen ermöglicht, das Wissen über multimodale Anwendungen auf ihre eigenen Konzepte im Fahrzeug zu übertragen.

Die vorliegende Arbeit schließt diese Lücke, indem sie fundierte Design-Unterstützung für multimodale Anwendungen im Fahrzeug basierend bereitstellt. Sie beginnt mit einem Überblick über die bestehende Entwurfsunterstützung für automobile und multimodale Anwendungen. Darauf aufbauend wird eine Reihe von Nutzerexperimenten berichtet, die verschiedene Aspekte der multimodalen Interaktion im Fahrzeug mit mehr als 200 Probanden in Laboraufbauten und Fahrsimulatoren untersucht. Diese Experimente bewerten die Potenziale der Multimodalität während der Fahrt, erforschen wie Benutzerschnittstellen Sprache und Gesten unterstützen können, und evaluieren neuartige Interaktionstechniken. Die Erkenntnisse aus den Experimenten erweitern das vorhandene Wissen aus der Literatur, um daraus eine erste Sammlung von Design Patterns für multimodale natürliche Interaktion im Fahrzeug zu erstellen. Die Sammlung enthält 15 Patterns, welche Lösungen für immer wieder auftretende Probleme beschreiben, wenn natürliche Eingaben mit Sprache, Gesten oder Blicken im auf strukturierte Weise kombiniert werden. Schließlich stellen wir einen Prototyp eines Infotainmentsystems vor, der die Anwendung der vorgeschlagenen Design Patterns demonstriert und in einem Fahrsimulator-Experiment evaluiert.

Diese Arbeit trägt in dreierlei Hinsicht zum Bereich der Benutzerschnittstellen im Auto bei. Erstens stellt sie eine erste Mustersammlung für multimodale natürliche Fahrzeuginteraktion vor. Zweitens illustriert und evaluiert sie Interaktionstechniken, die Sprache und Gesten mit Blickeingabe kombinieren. Drittens liefert sie empirische Ergebnisse einer Reihe von Benutzerexperimenten, die die Auswirkungen der multimodalen natürlichen Interaktion auf verschiedene Faktoren wie Fahrleistung, Blickverhalten, Interaktionseffizienz und Benutzererfahrung zeigen.

Preface

This thesis is the result of the time I spent in the team for Human-Machine-Interaction at BMW Group from 2015-2020. During this period, I did not work in isolation, but all my decisions were strongly influenced by innumerable conversations and discussions with my team colleagues. In addition to that, the exchange with researchers, in particular from the Human-Computer Interaction and Media Informatics groups of the LMU, but also at occasions such as the conferences I visited, was very valuable and inspiring.

To emphasize that all those cooperations significantly contributed to the making of this thesis, I decided to use the scientific plural. Additionally, the publications which the sections describing the user experiments and prototypes are based on, are highlighted at the beginning of the respective sections. To enable a consistent presentation in this thesis, the content from these publications has been restructured and, where necessary, supplemented with additional information.

Acknowledgements

First, I want to thank my supervisor **Prof. Tom Gross** at the University of Bamberg for supervising this thesis. Thank you for always providing me help and advice and for having the patience to finally see this dissertation come to an end. My thanks also go to the further committee members **Prof. Diedrich Wolter** and **Prof. Thorsten Staake**. Thank you for your detailed feedback and valuable comments on this thesis.

My biggest thanks go to my colleagues at my team at the BMW research department. In particular, I want to thank **Sonja Rümelin**, who supervised my work on the BMW side. Thank you so much for all the support and advice during the past years. Special thanks to **Ronee Chadowitz** for bringing me into this great team and for the best team events ever (Hoooooome!). Big thanks to my BMW PhD colleagues **Oliver Jarosch** and **Michael Braun** for the good times in the office, for the great trips we had together in the US and Canada, and, of course, for the best MMI barbecue parties! Thanks to all my current and former team members, **Svenja Paradies, Florian Weber, Alex Peters, Elisabeth Schmidt, Lenja Sorokin, Julian Eichhorn, Nora Broy, Sigrid van Veen** for being great colleagues and for your acknowledgement and valuable feedback. Thanks to **Peter Schneider** for poking me to finally finish this dissertation and for giving me the time to do so.

I want to thank my students **Ahmet Firentepe**, **Carina Rothe**, **Isabel Schönewald**, **Valerie Hentschel**, **Mario Burrafato**, **Konstantin Raab,** and **Lars Reisig**, who did a great job in building prototypes, conducting user studies and many more things. Without all of you this dissertation would not have been possible. Special thanks go to our long-term working student **Jesper Bellenbaum** for refreshing my statistics skills and for all the after-lunch dart sessions.

I want to thank **Bastian Pfleging** for all the support, especially with my first publications. A big thanks also goes to **Felix Schwarz,** who brought me into this topic and helped me to start this dissertation. I want to thank **Teo Babic** for all the ideas we exchanged and discussed on Friday afternoons.

Finally, I want to thank my **family** for their help and advice during all the years of school and studying that finally led my here. Thank you for your support and trust.

Table of Contents

List of Abbreviations

AAM	Alliance of Automotive Manufacturers
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
AF	Acoustic Feedback
CID	Central Information Display
CTT	Critical Tracking Task
DALI	Driver Activity Load Index
DOF	Degrees of Freedom
EFRC	Eye-finger Ray-casting
ESOP	European Statement of Principles
ETSI	European Telecommunication Standards Institute
FPK	Freely Programmable Kombi Display (Instrument Cluster)
GUI	Graphical User Interface
HCI	Human-computer Interaction
HUD	Head-up display
ISO	International Organization for Standardization
IVIS	In-Vehicle Information System
JAMA	Japan Automobile Manufacturers Association
LCT	Lane Change Task
LTT	Look-to-talk
MFL	Multifunctional Steering Wheel
MRT	Multiple Resource Theory
NASA-TLX	NASA Task Load Index
NHTSA	US National Highway Traffic Safety Administration
NLU	Natural Language Understanding
NVUI	Natural Voice User Interface
PDA	Personal Digital Assistant
PLF	Peripheral Light Feedback
PTT	Push-to-talk
SAE	Society of Automotive Engineers
SDD	Standard Deviation of Distance
SDLP	Standard Deviation of Lateral Position
STT	Speech-to-text
TCT	Task Completion Time

TGT	Total Glance Time
TTC	Time to Collision
UEQ	User Experience Questionnaire
UC	Use Case
UI	User Interface
WIMP	Windows-icons-menus-pointers
WOZ	Wizard-of-Oz

1. Introduction

This first chapter provides an introduction and brief overview over this dissertation. First, we motivate the need for research in field of multimodal human computer interaction in the car that respects the specific requirements of the automotive domain. This is followed by a definition of the scope of this work, the main research questions that will be addressed, as well as a brief outlook of the main contributions. Finally, the dissertation roadmap points out the structure of this dissertation.

1.1. Motivation

User input is a key aspect of human-computer interaction (HCI). It determines how humans interact with machines and influences the user interface (UI). In the early days of computers, users controlled computers mainly using keyboards to write commands into a command line interface. In the 1970 the computer mouse was invented in Xerox Park. In combination with a graphical user interface (GUI), it was seen as a revolution in HCI as it greatly facilitated the interaction with computers. Especially for novice users the benefit was huge, as they were no longer required to learn abstract commands, but a simple click on an icon was enough to issue a command to the computer. However, the mouse has not replaced the keyboard. Instead, both input devices have evolved into a partnership, in which the mouse compensates for the limitations of the keyboard and vice versa. The availability of both devices empowers users to perform a greater variety of task on their computers. They write emails on the keyboard or use the mouse to draw digital paintings. Moreover, both input devices can not only be operated individually, but they can be combined to enable a more powerful form of input, which is necessary to cope with the requirements of modern applications, such as three-dimensional games, or complex graphics editors and modelling tools.

The suitability of an input mode is always closely connected to the usage context. Most UIs for desktop and laptop computers are optimized for mouse and keyboard input. This works great for many tasks in a desktop setting. However, with computers and HCI becoming ubiquitous, the use of mouse and keyboard was no longer feasible due to a variety of contexts of use. With the introduction of the first smartphones in 2007, touch interaction on an interactive screen has emerged as the primary input modality in the mobile context during the past decade. It allows easy and convenient operation of all kinds of electronic devices. Besides touch, other input modalities have evolved mainly in more specific application scenarios in HCI. Speech input, for example, can be very useful when the user's hands are occupied. Accordingly, the Amazon Echo smart speaker is often placed in the kitchen[1]. Users can interact with the device to set timers and control the music while preparing the meal or doing the dishes. Another example for a specific context of use is the medical domain. During an operation, surgeons may have to inspect and control images on a screen without touching anything to maintain sterilization. In this context, gesture recognition sensors such as the Microsoft Kinect have been used to support touchless remote manipulation of 2D and 3D images (Cheng Chang & Gao, 2016). Just like the

[1] https://www.recode.net/2016/9/21/12997080/amazon-echo-survey-kitchen, accessed on April 2nd 2019

mouse and the keyboard, these new input modalities do not only work independently, but they can be combined in order to provide new interaction possibilities. A famous example is Richard Bolt's "Put-that-there", which demonstrates how speech and mid-air pointing gesture input can be combined to interact with large screen display surfaces in a natural way (Bolt, 1980).

In this thesis, the context of use in focus is interaction in the car. The task of driving a vehicle involves much more than keeping the vehicle safely on the street. Drivers also have to control safety relevant functions such as headlight and windshield wipers. Moreover, they want to control comfort and entertainment functions that are provided by the in-vehicle information system (IVIS). At the same time, the driving context puts requirements to the interaction with such non-driving related tasks that limit the use of many input devices, such as mouse and keyboard. In earlier IVISs, functions were mainly controlled using simple buttons and knobs. However, with the growing number of features, hardware elements could be no longer directly mapped to vehicle functions. As a consequence, several hierarchical interfaces evolved that are controlled using one central input device, such as BMW iDrive, Daimler COMAND or Audi MMI. The success of touch input in many other domains finally led to the integration of touchscreens and is now replacing haptic input devices in many modern vehicles. One reason for the reluctant integration of in-car touchscreens are drawbacks in distraction and usability of using touch input while driving in comparison to haptic input elements (Lisseman, Diwischek, Essers, & Andrews, 2014). Accordingly, almost all car manufacturers introduced speech input in their vehicles, which is well suited for text entry e.g., to enter a destination for route guidance, but less suited for other tasks. Some manufacturers also included mid-air gestures for a small set of specific functions, as well as driver cameras that allow the system to monitor head and eye movements of the driver. These natural modalities have the potential for a more efficient and less distracting interaction, and could also fulfill the peoples' desire to communicate with intelligent systems instead of operating them (Koons, Sparrell, & Thorisson, 1998).

In summary, there are multiple input modalities available in modern vehicles, such as touch, speech, gestures, and gaze. However, there is only little interconnection between those input modes. Other domains have shown that the combination of input modalities can enable a more powerful interaction between human and computer. In this regard, current in-car interfaces do not exploit the full potential of the available input modalities. This is especially due to the lack of adequate design support that specifically targets the requirements of the automotive domain. For this reason, this dissertation aims to generate design support for the development of multimodal in-vehicle information systems by assessing the potential of multimodal interaction in a driving context and exploring new interaction techniques that combine

1.2. Scope

The scope of this thesis is the investigation and exploration of multimodal interaction techniques in order to generate design-support for multimodal in-vehicle interaction. The focus lies on the interaction with natural modalities, such as speech, touch, gestures and gaze, and their potential in the automotive domain, independent of a concrete in-vehicle information system. We are targeting secondary interaction with non-driving related tasks, while users are driving manually. In this regard, we do not aim to develop new technical solutions for the recognition of user input, but rather to maximize the potential of existing technology from a user centric perspective.

1.3. Research Questions

The thesis aims to provide design support for the development of multimodal IVISs. Therefore, we investigate different aspects of multimodal interaction in the car in several steps. First, we want to understand how drivers can benefit from multimodal interaction and which factors might limit is potential. Based on this, we explore how to use multimodality to support its potentials for in-vehicle interaction. Finally, we summarize the generated design knowledge in a condensed and reusable form to support designers of future multimodal systems in the car. Table 1.1 gives an overview over the concrete research questions and the sections in which they are addressed.

No.	Research Question	Section
R1	**Investigating potentials of multimodality while driving**	4
R1.1	How do modality switches affect the interaction while driving?	4.1
R1.2	How do situational effects change the suitability of input modalities?	4.2
R2	**Supporting the flexible use of alternative input modes**	5
R2.1	How can visual cues promote the use of speech input?	5.1
R2.2	How can the execution of gestures be effectively supported?	5.2
R3	**Enhancing interaction by combining input modalities**	6
R3.1	How can multimodal input be used to confirm gaze input?	6.1
R3.2	How can multimodal input enhance speech input?	6.2
R.3.3	How can multimodal input support gesture input?	6.3
R4	**Providing design support for interaction designers**	7
R4.1	How can design knowledge be made available for designers in a reusable way?	

Table 1.1 *Overview of Research Questions*

Multimodal interaction has been extensively studied in many domains in HCI. However, it is not clear to which extend the potentials can be applied in for in-car interaction. One of the main benefits of multimodality is increased flexibility, which is especially relevant in mobile contexts with changing environmental demands such as the automotive domain. In a first step we investigate how drivers actually benefit from multiple available input modalities (R1). A flexible use of multiple input modes also involves the need to switch between modes. Therefore, it is necessary to investigate the effects of modality switches and their influence on the overall benefit of the interaction (R1.1). Moreover, a key argument for multiple input modes is the possibility to adapt to changing environmental situations. However, it should be assessed to which extend and in which ways different modalities are affected by situational demands (R1.2).

In a second step, we explore possibilities of supporting a more flexible use of alternative input modalities (R2). Speech and gesture input can be used for different in-vehicle tasks, however the availability and execution of both modalities is often unclear in many interfaces. We explore how user interfaces can effectively support the use of speech input (R2.1) and gesture input (R2.2) by providing adequate cues and visualizations.

Besides a more flexible use modalities, multimodal input allows to combine modalities to overcome individual weaknesses of gaze, speech, and gesture input (R3). In particular, gaze input can be a fast pointing modality, however typical ways of confirming gaze input known from other domains (e.g., dwell times) are not well suited while driving. Therefore, we evaluate the suitability of confirming gaze input with gestures and speech (R3.1). Speech input generally has many advantages while driving, but it requires an explicit activation of system, which limits the efficiency and user experience of speech inputs. Multimodal approaches could provide a more natural and more efficient use of speech input (R3.2). Gesture input can be a valuable complement to speech inputs by providing a fast and direct interaction, but the accuracy of gesture input is often limited. This is not only based on the technical quality of the recognition hardware and software, but also due to inaccurate pointing by the driver. It remains to be shown how the integration of redundant input modalities can be used to increase recognition accuracy for pointing gestures while driving (R3.3).

Finally, we aim to provide design support for the development of multimodal in-vehicle systems (R4). In HCI, this is often achieved in the form of design guidelines. However, such guidelines are often difficult to apply for future developers. The knowledge derived from the experiments from the previous parts must be summarized in a structured way. Design patterns are a widely recognized method in HCI to describe proven solutions for reoccurring problems in a formalized way (Granlund, Lafrenière, & Carr, 2001). New design patterns for multimodal in-vehicle interaction must be generated based on the presented user experiments and existing patterns from multimodal HCI must be adopted to the automotive domain (R4.1).

1.4. Contribution

This thesis contributes to the field of HCI in three different ways: design patterns for multimodal in-vehicle systems, interaction techniques for natural in-vehicle interaction, and the insights from empirical evaluations in user experiments. Each contribution is briefly described in the following paragraphs. Please refer to Section 9.1 for a more detailed discussion of these contributions.

1.4.1. Design Patterns for Multimodal In-Vehicle Systems

We present the first collection of interaction design patterns for the integration of speech, gestures, and gaze in multimodal in-vehicle systems. The collection summarizes and organizes the insights from empirical evaluations in a formalized way and puts them into relation. Furthermore, the pattern collection is embedded in a larger context by incorporating existing patterns from literature.

1.4.2. Interaction Techniques for Natural In-Vehicle Interaction.

The thesis presents various prototypes for novel interaction techniques with natural input modalities while driving. In particular, we present techniques to support gesture interaction with objects inside and outside of the vehicle. We prototyped gaze-based interaction techniques in combination with speech and gesture input. Finally, we present a prototypical IVIS that demonstrates how to incorporate the presented design patterns.

1.4.3. Empirical Evaluations of Multimodal Interaction while Driving.

We report empirical results from multiple user studies that investigate the potentials of multimodal input in an automotive setting. The results include objective measures about driving performance and visual distraction, efficiency of interaction, as well as subjective assessments of cognitive demand and user experience.

1.5. Dissertation Roadmap

The remainder of this dissertation is organized as follows:

Chapter 2 gives background information about the two major fields of this thesis. The first one, multimodal human computer interaction, is an active field of research for more than 50 years. We clarify important terms and give a definition of multimodal interaction for the scope of this thesis. There are brief introductions in some theories from cognitive psychology, as they serve as a foundation for many potential benefits of multimodal interfaces. We present design spaces of multimodal systems, as well as important insights and guidelines for the design of multimodal systems. The second field, automotive user interfaces, is determined by the fact that the driver is in a dual task situation, in which the primary task is driving, and all other interactions are secondary tasks. We describe the different tasks the driver has to manage and the resulting forms of driver distraction. This leads to an introduction of state-of-the-art input modalities in modern vehicles. This is followed by a description of multimodal interaction concepts that have been presented in literature. We present goals and evaluation methods for multimodal in-vehicle interaction.

Chapter 3 deals with the problem to provide adequate design support for natural multimodal in-vehicle interaction. There are official standards and guidelines that support and regulate the development of IVISs, but they hardly cover the interaction with natural input modalities. We discuss the advantages of design patterns regarding their suitability to provide adequate design support. We describe methodologies to generate, structure and organize design patterns and present relevant pattern collections. Based on that, we point out the research directions for this thesis.

Chapter 4 investigates the benefits of the availability of multiple input modalities. The usage of multiple input modalities requires users to switch between input modalities. Therefore, in a first user experiment, we assess the potential costs that arise due to the process of switching the input modality. The second experiment in this chapter investigates the effects of situational demands on different input modalities.

Chapter 5 deals with the support of alternative input modalities, such as speech and gestures. We present two user experiments that show how to create a better awareness for both input modes in the cockpit. In particular, we investigate the design and efficacy of visual prompts to promote the use of speech input. For gesture interaction, we present an interactive prototype that uses peripheral light to provide feedback for in-car gestures.

Chapter 6 focuses on the combination of two input modalities in order to overcome individual limitations. The chapter describes three different interaction techniques. A first prototype combines gaze input in combination with speech or gestures input to confirm selections, to overcome the limitations of the dwell time approach. Second, we use gaze input in a passive way to activate the speech system in the vehicle and thereby support the efficiency and naturalness of speech input. The third prototype demonstrates how gaze can be used to increase the accuracy of gestures input without creating additional demand on the user. For each prototype we present an evaluation in a user experiment.

Chapter 7 presents a collection of design patterns for multimodal in-vehicle interaction. These patterns are derived from the individual experiments and prototypes from the previous chapters. The pattern collection provides an overarching organization of the findings, puts them into relation with each other, and links them to literature. Moreover, each pattern is described in a formal way according to design patterns in HCI.

Chapter 8 presents a multimodal prototype that integrates the design patterns presented in the previous chapter. It is based on a combination of speech, gestures, and gaze input. We give a detailed description of the prototype and the design patterns used. A driving simulator experiment is conducted to evaluate the effects of the multimodal approach in comparison to a speech-only interface. The results include insights on driving performance, gaze behavior, and interaction efficiency, as well as subjective ratings of the participants and their usage of the different input modalities.

Chapter 9 concludes this thesis. We summarize the contributions of this work and point out how they answer the research questions presented above. At last, we discuss the future areas of research areas for in-vehicle interaction.

2. Background

This chapter provides fundamental information about the two main research fields of this dissertation. The first section gives an overview over the field of multimodal human computer interaction. The second section describes the context of the automotive domain with a focus on interaction while driving.

2.1. Multimodal Human Computer Interaction

This section describes relevant aspects of multimodal human computer interaction. First, we clarify a few frequently used terms that are used in literature and define multimodal interaction. We give a brief overview over the cognitive foundations that influence the development of multimodal systems. The subsequent sections summarize potential benefits and describe the design space for multimodal systems.

2.1.1. Clarification of Terms

According to the broad range of multimodal systems, there is a number terms relevant to multimodal interaction that have been used in various contexts and with subtly or even significantly different meanings. Especially the terms modality, mode and channel are frequently used and will be explained in the following section.

Modalities and Channels

In human perception, a modality or mode refers the particular human sense that is used for receiving a specific stimulus (Turk, 2014). The five human senses are vision, hearing, touch, smell, and taste. In HCI, most of the input and output modalities can be classified as visual, acoustic, or tactile modalities, while the usage of smell and taste is less common as a means for communication in human-computer interaction. The sensory modalities of a system describe the mode of perception according to these human senses. For example, cameras provide vision to computers, microphones allow them to hear and process spoken language, and haptic sensors enable computers to react touch input. Moreover, there are other input devices, which cannot be directly mapped to human senses. Some examples are keyboard- and mouse input, rotary controllers in cars, or neural input from EEG devices (Jaimes & Sebe, 2007).

Besides specifying the sensory modality of the system that is used to receive information, the term modality is also used to cover the human output modality, which describes the way an intention is expressed, or the manner an action is performed (Nigay & Coutaz, 1993). In a HCI system, these different human output modalities are used to generate user input using various input devices. Users can tap, move, and click, speak voice commands, or make gestures with their heads and hands. Using input devices such as touchscreen, mouse, microphone, and camera the system can interpret these actions as user input. Different user inputs do not necessarily map to accordingly different sensory modalities of the system. For example, head movements (e.g., nodding for confirmation) and hand gestures (e.g., thumbs up for confirmation) typically use the same sensory modality for recognition, namely camera-based vision. However, both forms of user input differ in the way how they convey information and are therefore considered different modalities. Conversely, the same human output modality can be captured using different sensory systems for recognition. Mid-air hand gestures can be detected with a camera-based solution or with an EMG (electromyography) armband, but from

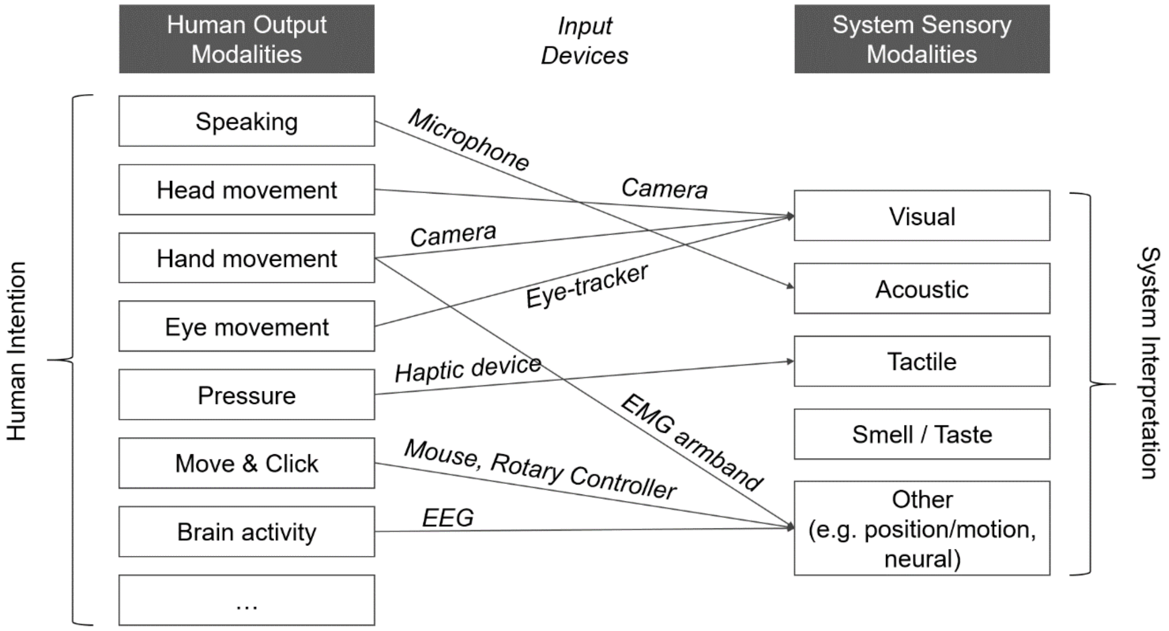


Figure 2.1: Mapping of human output modalities and system sensory modalities using different input devices based on (Sharma et al., 2002).

the user's perspective it is still the same modality. This connection of human output modalities and sensory modalities of the system is illustrated in Figure 2.1.

The specific combination of human output and the sensory modality of the system is defined as a communication *channel*. A communication channel describes one *"particular pathway through which information is transmitted"* (Dumas, Lalanne, & Oviatt, 2009). It is typically represented by one particular combination of user ability and input device capability. In the previous example, the user can nod with the head, or to make a thumbs up gestures with the hand. The system has the capability to detect these actions (e.g., via a camera). Thereby it creates a communication channel through which information can be transmitted. Each combination represents one particular communication channel. Here are some more examples for different channels based on Figure 2.1**:**

- Head movement (e.g., nodding) – Camera
- Hand movement (e.g., mid-air gesture) – Camera
- Hand movement (e.g., mid-air gesture) – EMG armband

In this work, we use the terms *(user) input modality*, or *(user) input mode* from a user-centered perspective if not denoted otherwise. They describe a particular way in which the user transmits information to the system, based on the human output modality and the input device used. Consequently, different user input modalities may share the same sensory capabilities. For example, head nodding, and mid-air hand gestures are different input modalities even though they may both use a camera-based system for recognition. An interaction technique is typically characterized by one or multiple communication channels that are used in a specific way.

Definitions of Multimodality

The term *multimodal* is used in many different contexts and disciplines, also beyond the field of HCI. Therefore, we give a definition of multimodal HCI systems. Literally*, multi* refers to *more than one,* while *modal* may refer to *modality* as well as to *mode* (Nigay & Coutaz, 1993). Accordingly, in the context of user interfaces, the European Telecommunication Standards Institute (ETSI) defines multimodality as a property of a user interface in which more than one sensory modality is available for input or output (e.g. output can be visual or auditory), or in which a particular piece of information is represented in more than one modality (e.g. the command to open a file can be spoken or typed) (European Telecommunications Standards Institute (ETSI), 2003). This definition describes multimodality mainly as the existence of more than one mode or modality for system output or user input. A multimodal HCI system could therefore be described as one that simply responds to inputs in more than one modality or communication channel (e.g. speech, gesture, writing, and others) (Jaimes & Sebe, 2007). In this work, we rather want to refer to one of the most frequently referenced definitions of multimodal HCI systems given by Oviatt (Sharon Oviatt, 2012):

"Multimodal systems 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" (S. Oviatt & Jacko, 2012).

This definition points out some more relevant aspects, which characterize a multimodal system. First, multimodal system processes at least two combined input modes, but they may vary along the number and type of input modalities and communication channels used (Turk, 2014). Second, the definition points out that not only multiple input modes are available, but also that they have some sort of interconnection. Again, there is great variance on how to combine individual modalities. We will give an introduction into the design space of multimodal systems in Section 2.1.4. Third, user input is linked to multimedia system output. This means that the system makes use of multiple output modalities, such as visual, acoustic, or tactile output. It is important to recognize that both, multimodal input and output are part of a multimodal system and must be fully supported in multimodal interaction systems (Turk, 2014).

In practice, most interactive systems make use of multiple input and output modalities. Consequently, the given definition of multimodality fits to a very wide range and variety of interactive systems. For example, a regular desktop workstation is a computer interface that offers multiple input modalities. Users usually use a combination of mouse and keyboard to issue commands to the computer, e.g., to browse the internet. Clicking on a web link produces acoustic (click sound) and visual (the browser loads a new webpage) feedback. In this sense, the workstation with mouse and keyboard as well as many other interactive electronic devices that offer multiple input and output modalities could therefore be classified as multimodal systems.

Towards Natural Interaction

However, the class of multimodal systems represents more than the mere availability of multiple ways of communication between user and system. It does not only refer to the amount of input and output modalities used, but also to the type of modalities used and their impact on user interfaces. *"This new class of interfaces aims to recognize naturally occurring forms of human language and behavior, which incorporate at least one recognition-based technologies* *(e.g. speech, pen, vision)"* (S. Oviatt & Jacko, 2012). Thereby, it also represents a paradigm shift away from conventional user interfaces based on windows, icons, and pointer interaction (Dumas et al., 2009; S. Oviatt & Jacko, 2012). Accordingly, there is a shift away from conventional input devices, such as computer mouse, keyboard, or any other device that is dedicated to the control certain computer functions, towards the recognition of natural human behavior such as human speech, hand gestures, head movements, facial expressions and eye-gaze that enable a more human-like style of communication with computer systems (Dumas et al., 2009).

In summary, multimodal systems enable a communication between humans and computers that integrate multiple input modalities with multimedia output. This can refer to the user's perspective, as well as the system's perspective or both. Moreover, there needs to be some form of coordinated interconnection between the modalities that are used. Finally, multimodal interfaces integrate two or more natural input modalities and combine it with multimedia output in order to enable a more natural form of communication with the system. Natural modalities are those human output modalities that are used in interpersonal communication, such as speech, gestures, gaze, facial expressions, and body language.

2.1.2. Cognitive Foundations

A few decades ago, computers were highly specialized devices for experts. The development of these systems was primarily technology-driven. Users had to adapt to the system in order to use it. Nowadays, computers are ubiquitous tools for all kinds of tasks and a wide range of users in professional, but also in private contexts. This development raised the need for an intuitive communication between humans and computers, which resulted in the establishment of the independent, interdisciplinary research area of HCI. One key factor in HCI is to understand and model users' natural behavior to build user interfaces that are more intuitive, easier to learn and more effective, by respecting human abilities to perceive and process information. This understanding is mostly generated by human-computer interaction studies, but also from theories in cognitive psychology. This section presents two models from cognitive psychology that provide a foundation for the usage of multimodality in HCI.

Theory of Working Memory

The term working memory refers to *"a brain system that provides temporary storage and manipulation of the information necessary for such complex cognitive tasks as language*

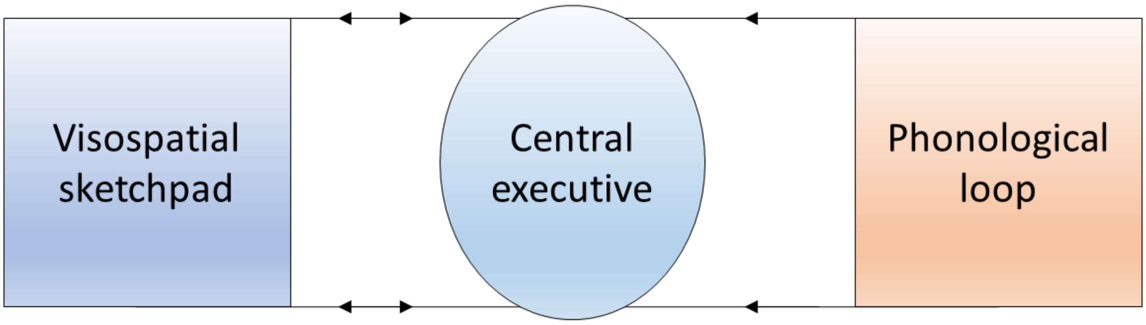


Figure 2.2: The three-component model of working theory is composed of a central executive and two storage systems, the visuospatial sketchpad and the phonological loop (based on (Baddeley, 2003).

comprehension, learning, and reasoning." (Baddeley, 1992). The theory of working memory basically proposes that the human short-term or working memory consists of two processors (see Figure 2.2). Visual images and materials, such as pictures and diagrams are stored in an area of the working area, which is described as the visual-spatial *sketchpad*. Auditory-verbal information is stored in a separate *phonological loop*. Both processors are subsystems that are coordinated by a *central executive,* which acts as attentional-controlling system (Baddeley, 1992). Nevertheless, the visual-spatial sketch pad and the phonological loop are viewed as functioning largely independently. This enables the effective size of working memory to expand, when people perform tasks using multiple modalities for interaction (Sharon Oviatt, Coulston, & Lunsford, 2004).

Multiple Resource Theory

The multiple resource theory (MRT) is a theory of multitasking performance. It aims to predict the level of interference between two concurrently performed tasks (Wickens, 2002). Earlier theories explained the variance in dual task performance mainly because of the difficulty (quantitative resource demand) of the performed tasks. In contrast, the MRT builds on the concept of *resources*, which proposes that different perceptual modalities are connected to associated resources and that the task performance of a user also depends on the use of these resources (qualitative resource demands). Two tasks that claim the same resources will interfere and consequently impair the performance of both tasks. A typical example is human perception with the eyes (visual processing) and the ears (auditory processing), which relate to different perceptual modalities. Multitasking of two visual tasks will be less effective compared to one visual and one auditory task, since they claim different structural resources. Therefore, the

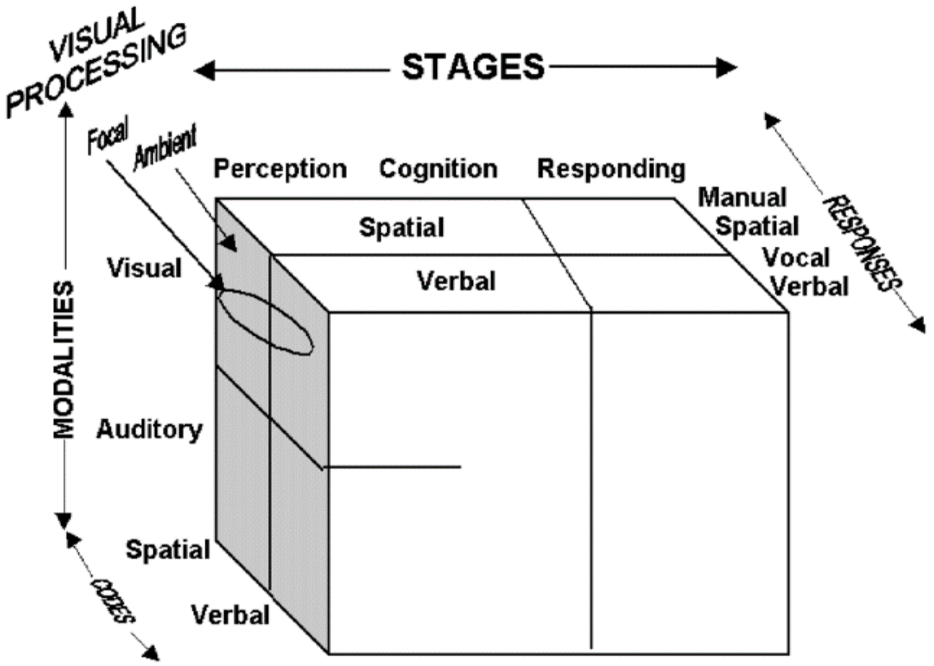


Figure 2.3: A three-dimensional representation of the structure of multiple resources. The fourth dimension of visual processing is illustrated in within the visual modality resources (from (Wickens, 2008))

essential message of the theory is that avoiding competition of modalities by distributing tasks across modalities will result in better multitasking performance (Sharon Oviatt et al., 2004).

Moreover, Wickens' multiple resource model proposes additional structural dimension in addition to the used perceptual modalities. In total, there are four dichotomous structural dimensions of human information processing that can account for variance in multitasking performance. The four dimensions are *stages* (perception/cognition vs. response), perceptual *modalities* (visual vs. auditory), *visual channels* (focal vs. ambient), and processing *codes* (spatial vs. verbal) (Wickens, 2002). The basic idea is that *"two tasks that both demand one level of a given dimension [...] will interfere with each other more than two tasks that demand separate levels on the dimension"* (Wickens, 2002). The multiple resource model illustrates this relationship in a three-dimensional representation of the structure of multiple resources. It is illustrated in Figure 2.3.

Although the multiple resource model is able to account for the majority of variance multitasking performance, there is a number of issues that challenge both the theory and the model. For example, one issue is the fact that the dimension of modalities differentiates only visual and auditory modalities. This could be expanded by adding a third level to the modality dimension related to tactile input (Boles, Bursk, Phillips, & Perdelwitz, 2007; Wickens, 2002) in order to make the model applicable for a greater variety of tasks.

In summary, these two models from cognitive psychology provide basic assumptions that influence the development of multimodal interfaces (Dumas et al., 2009; Sharon Oviatt, 2006). First, humans are able to process modalities partially independently and, thus, presenting information with multiple modalities can increase the effective size of human working memory. Second, human performance is improved when interacting multimodally, due to the way human perception, communication, and memory function works.

2.1.3. Benefits

The previous section described two models from cognitive psychology that are relevant for multimodal interfaces. Accordingly, literature describes several benefits of multimodal interfaces. They are described as a more transparent, flexible, efficient, and powerfully expressive means of human computer interaction. They are easier to learn and use, and thus preferred by a broader range of users for many applications (Sharon Oviatt, 2012). They allow to recognize naturally occurring forms of human language and behavior through the use of recognition based technologies and thereby deliver natural and efficient interaction (Turk, 2014). Based on this, we classify the benefits of multimodal interfaces in four categories: flexibility, efficiency, robustness, and naturalness. The following paragraphs provide a brief summary for each category.

Flexibility

Alternation between input modalities and the integrated use of multiple input and output modalities offer flexibility to users of multimodal interfaces. By providing alternative input modalities, a multimodal system is able to accommodate a wide range of individual user preferences, task requirements, or environmental conditions (Sharon Oviatt, 2012; Reeves et al., 2004; Turk, 2014). Moreover, for users with disabilities, flexible multimodal input is a great way to provide accessibility (Jaimes & Sebe, 2007). Temporary or permanent handicaps can be better supported with a flexible multimodal system that enables alternative input modalities like speech, or head movements (Karray, Karray, Alemzadeh, Saleh, & Arab, 2008). For example, thanks to the integration of speech-based assistants in smartphones, users have the flexibility to do a web search using a voice request and getting a spoken answer, instead of typing a query and reading the response. For users with seeing impairments, this may be the only way to use this function, while non-handicapped users can choose their preferred way of interaction between speech and touch input on the screen. This also applies to users of different skill levels, ages, or cognitive styles. Thereby, multimodal interfaces can accommodate a broader range of users than unimodal systems (Sharon Oviatt, 2012).

A multimodal interface also permits users to better adapt to changing conditions of use. Environmental conditions can have a significant influence on the suitability of individual modalities. For example, touch input is well suited for typing a message on a smartphone in a static environment, like sitting or standing still, but it is less suited in other environments like walking or sitting in a car that drives over a bumpy road. This aspect is especially relevant for mobile applications, which often face continuously changing environmental conditions (Sharon Oviatt, 2012).

Furthermore, flexibility of input modalities allows to prevent overuse of any individual modality (Sharon Oviatt, 2012). This is especially valuable in human computer interaction scenarios that require exhaustive or repetitive use of a single modality. Consequences can range from boredom or annoyance (e.g., caused by the repetitive use a voice command) to physical damage. In particular, the long-term overuse of a computer mouse can lead to injuries of shoulder, forearm and hand, which are often characterized by tendinitis (Pe ina & Bojani , 2003).

The flexibility of multimodal interfaces does not only relate to the modalities used, but also enables flexibility in terms of the temporal relation of multimodal integration. This is necessary since users differ in how they integrate multiple modalities. They can be classified as either sequential or simultaneous integrators (Xiao, Girand, & Oviatt, 2002; Xiao, Lunsford, Coulston, Wesson, & Oviatt, 2003). Multimodal systems should support users in whichever integration pattern they prefer.

Efficiency

Multimodal interfaces support greater efficiency by supporting multiple input and output modalities. Thereby, they can provide efficient input forms for a greater variety of tasks. For example, providing pen input for manipulating graphical information in addition to speech input for text input increases the efficiency of the interaction compared to unimodal input with only pen or speech (Sharon Oviatt, 1997). In this sense, efficiency is a consequence of increased flexibility since alternative input modes allow users to choose the most efficient form of input.

Moreover, there are also positive effects on efficiency based on a more integrated use of different modalities. Simultaneous input via two or more different channels increases the bandwidth for transmitting data. Theoretically, speech and pen input can temporarily overlap or even occur simultaneously, which reduces the duration of the overall interaction. Closely connected input modalities can also have a mutual influence. For example, speech commands that are uttered in a multimodal context are usually shorter and more command-like than in a speech-only application (Sharon Oviatt, 1999).

In terms of perception, humans are able to process information faster when it is presented in different output modalities. It has been shown that the perception of speech is improved when the facial movements of the speaker are visible (van Wassenhove, Grant, & Poeppel, 2005). These examples demonstrate that multimodal interfaces can increase task efficiency. However, Oviatt also points out that efficiency should not be considered the main advantage of multimodal interfaces, since the actual reduction of task completion time is limited (Sharon Oviatt, 1997).

Robustness

Increased robustness can be derived from both, flexibility, and efficiency. Flexibility allows users to choose less error prone input modes and thereby achieve greater precision for many tasks. At the same time, users can avoid the risk for errors that arise from using an unsuitable input modality. In case of an error, alternative input modes provide a suitable way for error recovery and correction, as users profit from the ability to switch to other modalities in case of a recognition error. For example, recognition errors of speech recognition systems can be efficiently corrected with spelling or writing, which results in higher accuracy than unimodal approaches (Suhm, Myers, & Waibel, 2001).

Furthermore, combining different partial information sources can be used for mutual disambiguation. Probabilistic results from each input modality can be fused in a semantically meaningful way, so that the overall output can reach a higher positive recognition rate than interpretation of isolated input modes (Sharon Oviatt, 2012). For example, the accuracy of speech input can be increases by integrating facial expressions or additional touch input (J. Lee, Lee, & Kim, 2017; Sezgin, Davies, & Robinson, 2009).

Naturalness

Multimodal interaction allows users to integrate communication patterns like interpersonal interaction by observing and interpreting natural human behavior. There is a shift away from conventional input devices, such as mouse and keyboard, towards the use of human means of communications, such as speech, gestures, and facial expressions. This has the potential to create interfaces that can understand natural human behavior and can therefore provide a form of interaction, which is more intuitive and easier to learn (Sharon Oviatt, 2012). Another advantage that comes with natural modalities is the support for multi-user and mobile interaction (Dumas et al., 2009). Conventional input devices are typically designed to be used by a single user in a static environment. A computer with a mouse and a keyboard can be controlled only be one user at a time. Speech, gestures, and facial expressions, on the other hand, do not need dedicated input devices to interact with the system, but only the user's ability to speak and move. Multiple users can speak to a system or perform gestures at the same time, given that the system can handle simultaneous input. Furthermore, the lack of dedicated input devices enables mobile interaction concepts that are increasingly independent of the user's physical position. While a mouse or a touchscreen requires the user to be in front of the screen to see the cursor movement and/or to target the desired UI elements, natural input modalities are less locally constrained. Thereby, natural interaction has a great potential for the application in mobile interactive environments.

However, it is important to acknowledge that a system that is based on speech and gesture interaction is not inherently more natural than a conventional system controlled by mouse and keyboard (Norman, 2010). The fact that some interaction modalities occur naturally in communication with other human beings does not necessarily imply that the same form of interaction feels natural in human-computer interaction. Instead, user interfaces must move away from standard WIMP (Windows-Icons-Menus-Pointers) interfaces and emphasize the use of natural modalities. Users have to learn how to formulate speech utterances and how to perform gestures (Norman, 2010). Standards and conventions have to be developed, as they are important factors for the naturalness of interaction with a computer system.

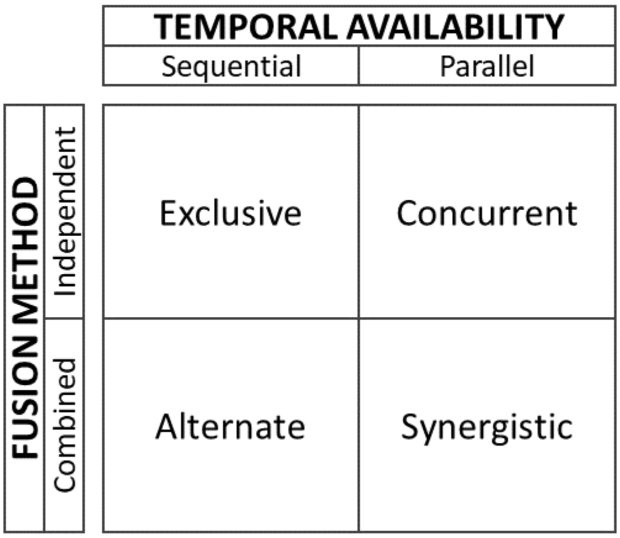
2.1.4. Design Space

The previous section described how the flexibility, efficiency, robustness, and naturalness of human-computer interaction can benefit from multimodality. In the end, the benefits in the real world depend on the actual design of each individual system. Multimodal systems may vary along a number of different aspects, such as the number and type of used input and output modalities, and the form of interconnection between these modalities. The variety of resulting systems can be categorized using design spaces for multimodal systems.

CASE properties

Nigay and Coutaz suggest a design space of multimodal systems based on the temporal availability of modalities (sequential or parallel), and the fusion method (combined or independent) (Nigay & Coutaz, 1993). The CASE properties describe the combinations of these two dimensions, which result in four different classes of multimodal systems: *Concurrent*, *Alternate*, *Synergistic*, and *Exclusive*. They are illustrated in Figure 2.4.

In a sequential multimodal system, each interaction step is completed using only one input modality. There is no temporal overlapping of input from two different modalities. Furthermore, there can be multiple input modalities available, however the user choses one for each interaction step. The counterpart to sequential multimodality is the parallel application of different input channels. Parallel or simultaneous multimodal input means that the input with two different input modalities does at least a partially overlap (Sharon Oviatt et al., 2004).



		TEMPORAL AVAILABILITY	
		Sequential	Parallel
FUSION METHOD	Independent	Exclusive	Concurrent
	Combined	Alternate	Synergistic

Figure 2.4: The design space of multimodal interaction techniques based on the use of modalities and the fusion method (based on (Nigay & Coutaz, 1993)).

Exclusive systems

Sequential systems differ regarding the level of fusion between inputs. In an e*xclusive* system, inputs are completely independent. This means that the interpretation of input with one modality is not depending on previous input with a different modality. For example, a multimodal video library might enable users to page through available videos by either uttering a spoken command ("turn to the next page") or by performing a swipe gesture. The user decides for the form of input that suits him best.

Alternate systems

In an *alternate* system, sequential inputs have a strong interdependency. The system processes different sequential inputs to build up a command. The interpretation of user input is made based on input from a previous step. In literature, this form of combination if also described as *temporally cascaded modalities* (Müller, Weinberg, & Vetro, 2011; Sharon Oviatt, 2012). In the example of a multimodal video library, a spoken deictic reference (e.g., "play this video") only makes sense, if a video has been selected before (e.g., by pointing or looking at it). The information from both sequential sources must be fused to generate the entire command.

Concurrent systems

Concurrent systems support the parallel input from two independent modalities that transmit information. Independent input does not mean that there is no interconnection between both modalities at all, but that the modalities transmit complementary content. In theory, this enables users to profit from a broader bandwidth of communication. In the example of the video library, users could issue a speech command to play a video, and at the same time adjust the turn of the system volume with a mid-air gesture (e.g., by rotating the index finger in a clockwise motion). The theoretical benefit of this class is a broader communication bandwidth. Users can transmit more information in the same time, resulting in shorter interaction times. This can be particularly important for many functions that require efficient interaction. Parallel input of different modalities further allows users to exploit individual strengths of each modality. This could not only promote efficiency, but also increases naturalness of interaction, since it has been shown that people tend to communicate multimodally when their own cognitive load increases (Sharon Oviatt et al., 2004).

Synergistic systems

Finally, *synergistic* systems enable a parallel use of modalities and fuse the information from the different sources. For example, the performance of speech recognition systems can be improved by incorporating the speaker's facial expression (Gibbon, Mertins, & Moore, 2000; Sezgin et al., 2009). From a user's perspective, the user might not interact multimodally, since speech is the only active input modality used. However, from the system's perspective, multiple different information channels are used, which can be fused with active speech input. The combination of one or more active modes with at least one passive mode is also called *blended interaction* (S. Oviatt & Jacko, 2012).

CARE Properties

The CASE properties describe multimodal systems from a system perspective. A way to characterize and assess the usability of multimodal interaction techniques from a user's perspective are the CARE properties (Coutaz et al., 1995). CARE stands for *Complementarity*, *Assignment*, *Redundancy*, and *Equivalence*. They provide a way to describe the relationships of available interaction techniques in a multimodal user interface. As an example, the authors use the CARE properties to describe user input for a multimodal airline travel information system, which allows users to schedule flights using speech, direct manipulation, keyboard, and mouse input.

Complementarity

Complementarity describes multiple modalities that are used in a complementary way within a temporal window, to provide a piece of information. Complementarity may occur sequentially or in parallel within the temporal window.

Example: The user says, "flights to this city", and uses the mouse to select a city on the screen.

Assignment

Assignment describes a fixed allocation of one modality to provide a specific piece of information. In contrast to equivalence, it expresses the absence of choice.

Example: *Window manipulations (e.g., moving, resizing) can only performed with the mouse.*

Redundancy

Redundancy describes multiple modalities that provide the same piece of information. Modalities may provide this information sequentially or parallel, but always within a defined temporal window.

Example: The user says, "flights to Pittsburgh", and uses the mouse to select Pittsburgh.

Equivalence

Equivalence describes the availability of choice between at least two modalities to provide the same piece of information. There are no temporal constraints on equivalent modalities.

Example: The user has the choice of speaking or typing the sentence "flights to Pittsburgh".

Fusion Methods

Multimodal systems can further differ regarding the fusion method that is applied to combine information from different channels. There are several existing classifications of multimodal fusion techniques. Sharma et al. differentiate between data-level fusion, feature-level fusion and decision-level fusion (Sharma, Pavlovic, & Huang, 2002). Accordingly, Sanderson and Paliwal proposed a similar differentiation based on pre-mapping, midst-mapping, and post-mapping fusion (Sanderson & Paliwal, 2002). The main difference is the time at which level of processing the fusion of different sources takes place (Dumas et al., 2009). The following paragraphs give a brief summary based on (Dumas et al., 2009; Schnelle-Walka, Duarte, & Radomski, 2016):

- **Data-level fusion** works on raw data from different recognizers. Usually, the signals from two similar information sources, e.g., from two webcams. As there is no preprocessing of the data, there is no loss of information. On the other hand, the data is highly susceptible to noise and failure.
- **Feature-level fusion** is used when tightly coupled or time synchronized modalities are used. By extracting features from raw data, some information is lost, but noise can be

better handled compared to data-level fusion. The combination of speech with lip movements is an example for feature-level fusion.

- **Decision-level fusion** is using preprocessing information that has semantic information. Due to the relatively great amount of preprocessing, it is less sensitive to noise and failure than feature level fusion, but the quality of the results strongly depends on the quality of processing and the semantic interpretation. Decision-level fusion allows to combine also loosely coupled modalities of differing data structure, such as speech and pen input. Therefore, it is one the most popular form of fusion in multimodal applications. Typical implementations of decision-level fusion are:
  - Frame-based fusion
  - Unification-based fusion
  - Symbolic/statistical fusion

2.1.5. Summary

This section gave an overview over used terms, cognitive foundations, benefits, and design spaces for multimodal human-computer interaction. The term multimodality is used in various contexts in HCI. In this work, multimodal interaction describes a form of interaction that incorporates multiple input modalities in a coordinated manner with system output. One goal of multimodal interaction is a more natural interaction with computer systems. Thus, multimodal research frequently focuses on the application of natural input modalities that are used in human communication, such as speech, gestures, and gaze. Further benefits are increased flexibility, efficiency, and robustness of the interaction. This can be derived from models from cognitive psychology, underpinned by various experiments presented in literature. Multimodal systems can be classified based on the CASE properties, which describe the temporal availability of different input modalities and the level of fusion between these modalities from a system's perspective. The CARE properties, on the other hand, provide a user-centric way to describe the different relationship between multiple input modalities.

2.2. The Automotive Domain

This section describes the research area of automotive user interfaces. First, we present a classification of the driving task as a basis to define driver distraction and explain the concept of different driver resources. Then we give an overview over natural input modalities that are used for in-vehicle interaction and point out the specific benefits of multimodal input in the context of driving. We reference relevant work in literature. The last part of this section describes the methodological evaluation of IVIS including test environments and used measures.

2.2.1. Classification of the Driving Task

Drivers are not just driving while they are driving. Controlling gas and brake pedals, steering wheel and gearshift is basically everything a driver needs for controlling the movement of the vehicle. Besides that, drivers control many other functions that are not directly related to the task of driving, such as windshield wipers, indicators, and light, but also navigation, entertainment, and communication functions. In order to cover this complexity, the task of driving a car can be split into primary, secondary and tertiary tasks (Geiser G, 1985).

Primary Driving Tasks

Primary driving tasks refer to all tasks that the driver must perform to maneuver the car safely from A to B. Basically, this involves the lateral and longitudinal control of the vehicle. Lateral control describes the use of the steering wheel to keep the vehicle centered on the lane or to make a turn at a crossing. Longitudinal control refers to the use of the brake and gas pedals to control the velocity of the vehicle and adapt it adequately to the current traffic situation. Besides the lateral and longitudinal control of the vehicle, the primary driving task also includes the surveillance and assessment of the environment around the car, e.g., checking the distance to nearby vehicles, pedestrians, or objects by checking rear view mirrors and or instruments. Donges presented a description of the driving tasks in a *3-Level-Model* that differentiates three subtasks: Navigation, Maneuvering, and Stabilization (Donges, 2009).

- **Navigation** describes the process of planning the route from the starting point to a destination. It involves selecting a route based on the available possibilities and an assessment of the estimated travel time. This can be influenced by information about incidents, such as traffic jams, accidents and might require the driver to dynamically adapt the planned route. The actual effort of the navigation subtask depends on whether the driver is in a known traffic space. In unknown spaces, it requires a process of conscious planning, e.g., when finding an address in a foreign city. In contrast, in well-known space there is no planning needed, e.g. when driving the daily route to work. (Donges, 2009)
- **Maneuvering** basically consist of controlling the lateral and longitudinal movement of the vehicle based on the planned route from the navigation subtask, and the current traffic situation. To do this, it is particularly important to perceive the movement of the own vehicle and surrounding objects, which are in constant movement and therefore require a continuous assessment of the situation. (Donges, 2009)
- **Stabilization** describes corrective interventions of the driver to ensure that intended maneuvers, based on the maneuvering subtask, are performed correctly. (Donges, 2009)

Secondary Driving Tasks

Secondary driving tasks are not directly necessary for controlling of the vehicle, but they enhance driving performance and road safety. Drivers must react on a variety of road and environmental conditions and traffic situations. Exemplary actions for secondary tasks are turning on the headlights, using the indicators, turning on windshield wipers, or using the defrost function to remove ice on the windows.

Tertiary Driving Tasks

Finally, tertiary driving tasks are not related to the driving task, but the aim to increase the driver's comfort, provide information or entertain the driver and passengers. Typical tertiary tasks are the use of the climate controls, operating radio and music controls, or interacting with the in-car navigation system. In literature, the term in-vehicle information system (IVIS) has been established.

Alternative Differentiation of the Driving Task

The division of the driving task into primary, secondary, and tertiary tasks is widely recognized. An alternative approach divides the driving task only into two subtasks: primary and secondary tasks (Peissner & Doebler, 2011). The primary task is defined in the identical way as in the definition above. It includes all actions that are necessary to maneuver the car safely. Secondary tasks basically include all other tasks that are not directly related to the primary task. This includes both, secondary and tertiary tasks from the definitions above. In this work, we use the term secondary task in the latter meaning, as a term referring to all kinds of tasks in the car that are not part of the primary task of driving.

2.2.2. Driver Distraction

Driver distraction can be defined as *"the diversion of attention away from activities critical for safe driving toward a competing activity."* (Regan, Lee & Young, 2009) in (Peissner & Doebler, 2011). Drivers have a limited amount of resources for performing primary and secondary driving tasks. By assessing the amount of resources claimed by the primary driving task and the secondary driving tasks, it is possible to identify possible sources of driver distraction. For a more differentiated assessment, the driver resource pool can be divided into visual, manual and cognitive resources (W. W. Wierwille, 1993).

Visual resource

There are two forms of visual resources. First, there is the foveal visual subsystem. It provides a high resolution and thereby allows drivers to gather detailed information about their environment. Since the human eyes are tightly coupled and cannot be used separately, humans usually have only one foveal resource. It is not possible to use the foveal visual resource for the primary driving task and the secondary driving task at the same time. Thus, time-sharing is the only way to gather detailed visual information of two separate locations, e.g., observing the driving scene and reading instructions from a navigation system in the center stack. This tradeoff is crucial to consider for the design of secondary in-vehicle tasks. The second form of the visual resource is the peripheral visual subsystem. It has a lower resolution, but nonetheless it is very important. It enables humans to gather information of the environment even without directly focusing it. In the driving context, peripheral vision provides motion impressions and enables the driver to detect potential hazards that are in focus on the foveal visual subsystem. (W. W. Wierwille, 1993)

Manual resource

The manual resource is determined by the availability of the driver's hands. For the primary driving task, the driver needs at least one hand to control the steering wheel for lane keeping and smaller maneuvering tasks. In some situations, such as greater corrections, turns at intersections, or safety critical maneuvers the use of both hands on the steering wheel is generally required. These situations fully demand the driver's manual resource. Other manual tasks, such as buttons and switches must be delayed until manual resources are freed. Therefore, most secondary tasks are performed in less demanding situations, in which only one hand is needed to control the vehicle and the other hand is available. Moreover, many manually demanding tasks in the vehicle cannot be treated in isolation, but they are linked with other driver resources. For example, when operating buttons on the center stack, eye-hand coordination is required to carry out a targeted hand movement in the direction of the desired button. This creates additional demands on the visual resource. (W. W. Wierwille, 1993)

Cognitive resource

Both primary and secondary tasks place demands on the driver's cognitive resources. The amount of cognitive resources claimed can vary greatly depending on the tasks currently being performed. Driving on a straight road with little traffic requires very few cognitive resources, especially for experienced drivers. The same applies to simple secondary tasks such as changing a radio station or adjusting the volume. On the other hand, there are other cases where the cognitive load is high, e.g. when attempting to interpret confusing direction signs or when attempting to recall the correct setting for changing route options in a navigation system. (W. W. Wierwille, 1993)

Though not as obvious as visual and manual distraction, cognitive distraction is very important to consider for the design of in-vehicle infotainment systems. It can result in inattentional blindness by narrowing the driver's foveal vision to a specific area or device (Strayer, Watson, & Drews, 2011; W. W. Wierwille, 1993). In this regard, a major challenge is to measure the amount of cognitive load. While visual and manual load can be assessed based on measurable factors, such as glance behavior and hands on wheel detection, cognitive load cannot be measured directly. Some approaches base on biometric signals, such as heart rate signals, galvanic skin response, or eye-glance behavior (Mehler, Reimer, & Coughlin, 2012; Pfleging, Fekety, Schmidt, & Kun, 2016; Shi et al., 2007). Another option is to assess the cognitive workload based on subjective measures. There are several questionnaires that allow drivers to directly rate their perceived level of cognitive load. We describe some of these measures in Section 2.2.6.

Sources of Driver Distraction.

The previous paragraphs described the three types of driver resources (visual, manual, and cognitive) that are claimed by primary and secondary driving tasks. Distraction arises when the driver has to handle multiple tasks that compete for these resources, i.e., when the same resource claimed by the primary driving task is also demanded by a secondary task. For example, the

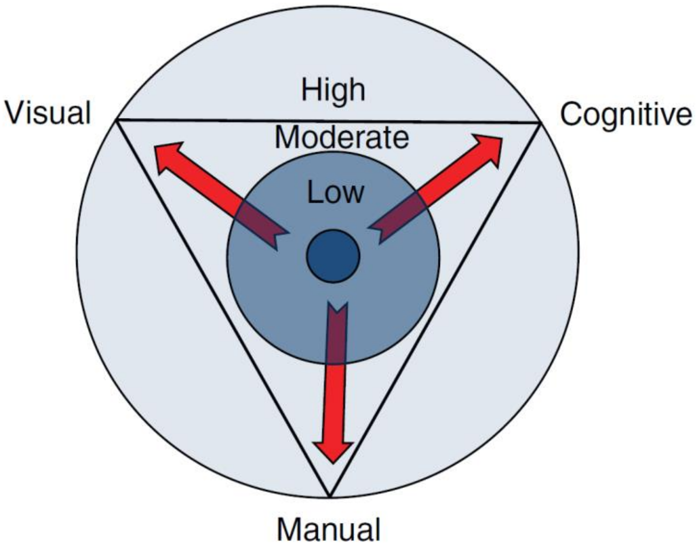


Figure 2.5: Visualization of the three sources of driver distraction. The three circles illustrate low, moderate and high levels of demand on the visual, manual, and cognitive resource (Strayer et al., 2011).

primary driving task demands a great part of a driver's visual resource. A secondary task that requires reading a text on an in-vehicle display also demands for a part of the visual resource and therefore leads to visual distraction. Manual distraction arises when secondary tasks put additional load on the manual resource, e.g., when the driver needs to take the hands off the steering wheel to manipulate the volume control or to select items on a touchscreen. Cognitive distraction occurs when the driver performs secondary tasks that are cognitively demanding, such as having a (hands-free) conversation over the cell phone or recalling and formulating a destination address for route guidance. These competitions for driver resources lead to interferences, which results in three sources of driver distraction. In-vehicle secondary tasks can result in competition on the visual, manual, or cognitive resource resulting in visual, manual, or cognitive distraction. Moreover, the occurrence of these sources is not mutually exclusive, but the interaction with different secondary tasks can lead to competition from one, two or all three sources (Strayer et al., 2011). This is illustrated in Figure 2.5.

2.2.3. Natural Input Modalities

This section summarizes interaction modalities that can be found in current vehicles or research work about in-vehicle information systems. We focus on natural input modalities, touch, speech, gestures, and gaze. For each modality, we give a short general introduction about its role in human-computer interaction before focusing on its application in automotive infotainment systems and potentials and challenges of its application in the automotive domain.

Direct-Touch Input

Touch is one of the five primary senses that humans use to interact with their environment. Humans explore their environment not only by seeing and hearing but also by touching objects in their environment i.e., by orienting a finger towards a target and touching it. In this sense, pointing at and touching objects of interest is a very natural means of human communication. Touch-sensitive screens enable this natural form of interaction also for human-computer interaction. (J. A. Pickering, 1986)

We use the term *direct touch input* when the display and the touch sensitive surface are positioned directly on top of each other in one device. Therefore, all touch-sensitive sensing methods, independent of their technology, are concerned with the detection of an object on a surface and determining its position (J. A. Pickering, 1986). Using the position of the finger, users can easily reference (pointing) and select (touching) content that is displayed on the underlying screen. The tradeoff of this interaction principle is that users are restricted to operate those functions of a system that are currently displayed on the screen. Functions without a selectable representation on the screen cannot be selected.

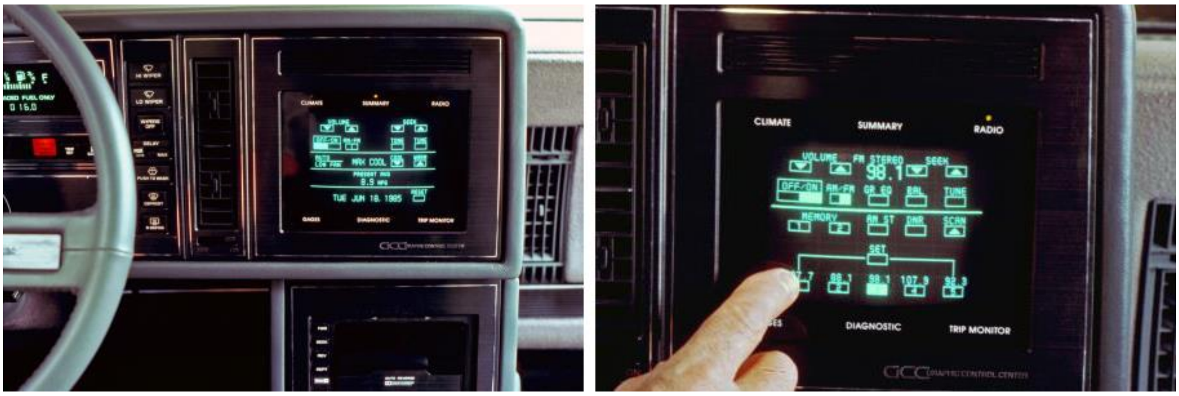


Figure 2.6: Buick offered direct-touch input in the 1980s with the so called Graphic Control Center in the Buick Riviera and Buick Reatta (Printz, 2017).

Besides the location of the finger, touch input offers some additional input parameters that can be used for interaction. The pressure of the users' finger on the screen can convey additional information. It enables systems to differentiate between different user actions. A pressure sensitive touch system can detect whether the surface was just slightly brushed or if it was actually pressed. This additional information can support touch-based interaction in many ways, e.g., by preventing unwanted selections or by enabling additional functionality depending on the applied pressure. Moreover, touch-gestures offer additional input options for touch input. They are controlled movements of the fingertips on a touch sensitive surface. A system, which reacts to touch gestures does not only react to the location of a finger at a certain point in time, but it interprets the change of the finger location over a certain duration. Multi-touch gestures use the location information of multiple fingers over time to convey even more information.

Direct-Touch Input in the Automotive Domain

One of the first cars that featured a touch-sensitive screens was the Buick Reatta (1988-1989) (Ortega, Barker, Wilson, & Kruse, 1987). This car offered a touch screen in the middle of the center stack to the right of the steering wheel. Figure 2.6 illustrates the touch interface of a 1988 Buick Reatta. It allowed to change many tertiary functions of the vehicle, such as controlling the volume, changing the radio station, or turning up the ventilation by selecting the respective touch buttons on the screen. However, this touch-based concept was discarded only a couple of years later and replaced by hardware switches again. It took quite a long time until car

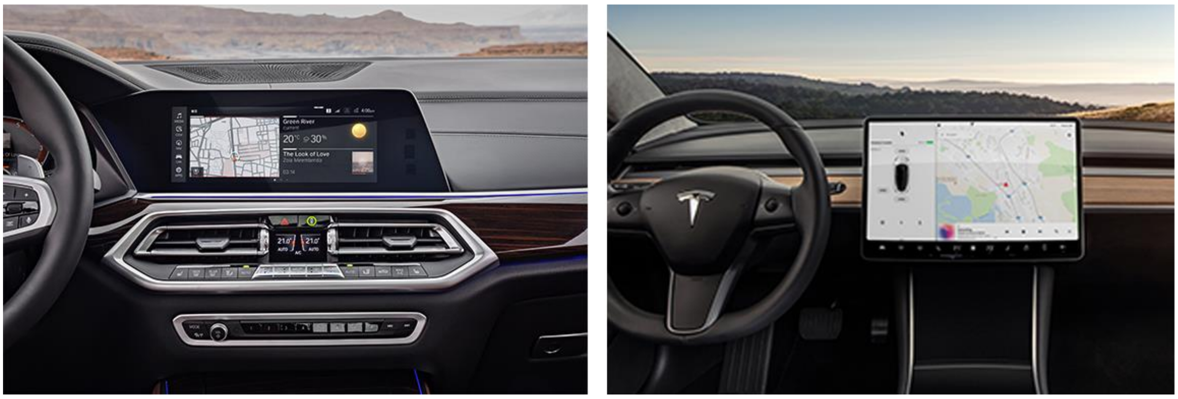


Figure 2.7: Touch interfaces in 2019 from the BMW 3 series and the Tesla Model 3 (official press photos).

manufacturers rediscovered touchscreens to control secondary functions. The growing number of functions in the vehicle made it unfeasible to map one haptic input element to each function. The more flexible menu-based interaction in combination with larger screens, as well as improved and widely accepted touchscreen functionality finally led to the rediscovery of touch interaction in the car. For example, the first BMW that featured a touch based central information display (CID) was the BMW 7 series which was released in 2015, 30 years after the first Buick Riviera. Today, touch interfaces have become state-of-the art in many modern vehicles (see Figure 2.7).

Touch input can be efficient for certain tasks in the car compared to alternative input modalities. May et al. compared direct touch input to input with mid-air gestures for in-vehicle menu navigation. Touch clearly outperformed gestures in terms of task time and errors. (May, Gable, & Walker, 2014). However, the efficiency of touch input depends on the driving workload. It has been shown that task times for touch-based text entry were significantly longer when driving as opposed to a parking car (Tsimhoni, Smith, & Green, 2004). Moreover, road conditions can affect the performance of touch and result in erroneous selections (Ahmad et al., 2015).

According to several studies, text entry with a touch screen keyboard is among the most visually distracting secondary tasks for the driver (Kujala, 2017; Tsimhoni et al., 2004). In comparison to speech input, touch leads to longer eyes-off the road time and a greater frequency of long glances off the road (Garay-Vega et al., 2010). However, this is not necessarily valid for all forms of touch input. The use of *touch-gestures* can significantly reduce eye glances off the road for simple secondary task interactions compared to a touch interface with touch buttons (Ba h, Jæger, Skov, & Thomassen, 2008).

Summary

Touch input on a touch sensitive screen allows to cope with growing number and complexity of features of modern vehicles. The touchscreen can show all interaction options, so that users do not need to remember specific voice commands or gestures for each function. This comes at the tradeoff of visual and manual distraction, as well as the limited accuracy of touch inputs in a moving vehicle.

Potentials	Challenges
Fast interaction	One hand off the steering wheel
Intuitive selection on screens	Eye hand coordination (eyes off)
Visibility of interaction options	Vulnerable to road bumps and car movement
	Little information bandwidth
	Physically exhausting (reaching)

Speech Input

Speech recognition describes the transformation of human-spoken words into machine-readable content, usually text. This allows the machine to identify the words a person is speaking and subsequently process an according command. It is also known as *automatic speech recognition* (ASR), or *speech-to-text* (STT).

One major benefit of speech input compared to traditional input modalities is, that the user can control functions of the system without having to manually control it, but only based on verbal input. This advantage makes speech input especially helpful when users find themselves in multitask situations. Speech input does not require visual attention and users have their hands free for other manual tasks. Therefore, it has been widely applied in various domains. For example, in medicine, surgeons use speech recognition to control equipment in the operating room, e.g. a surgical robot assistant (Zinchenko, Wu, & Song, 2017), while having their hands occupied. In healthcare, people with physical disabilities can user speech recognition to access computers. (Hua & Ng, 2010)

Speech Input in the Automotive Domain

Speech recognition has been also applied in the automotive domain. Nowadays, most modern vehicle manufacturers offer a speech recognition system that allows to control information and entertainment functions in the car. The primary driving task is primarily visually demanding because most of the information that drivers have to process is perceived visually. This is followed by manual demands that arise from controlling the steering wheel and operating secondary functions. Speech recognition does not claim visual nor manual resources. Even while making speech input, drivers can leave their eyes on the road and their hands on the steering wheel.

Earlier speech recognition systems were based on a predefined vocabulary, to improve the recognition accuracy and due to limited processing power. Besides technical limitations, this approach has some user-specific drawbacks. In order to trigger a function, users had to be aware of the available list of commands and speak the correct word to call a function. Learning and recalling all individual voice commands potentially leads to a high cognitive workload (C. a. C. a. Pickering, Burnham, & Richardson, 2007) making especially hard to use for novice users.

Natural language understanding (NLU) is a method to enable computers to extract meaning from the words that have been recognized by the ASR. A NLU engine converts spoken human language into formal representations that can be further processed by computer applications. One big advantage of speech recognition with NLU is the system understands the user's intent even if he did not use a specific formulation or predefined keyword. A user interface that uses ASR and NLU is also called a natural voice user interface (NVUI). NVUIs allow users to speak to the system the same way as they would when speaking to another person. This natural use of voice eliminates the need for remembering a specific lists of commands (Alvarez et al., 2011).

The differences between speech input and touch-based input for text entry has been shown in a number of previous studies (e.g. (Crandall & Chaparro, 2012; He et al., 2014; Kujala, 2017)). Studies compared speech input with touch input for address entry or music retrieval while driving. Speech input can outperform touch input in terms of total task time (Garay-Vega et al., 2010; Tsimhoni et al., 2004). In this regard, single turn voice interfaces are more efficient than system-paced speech interfaces based on dialogs (Garay-Vega et al., 2010). It was further shown that the total task times for speech input were not significantly affected by a moderate driving workload compared to operating the system when parking (Tsimhoni et al., 2004).

In contrast to that, Gärtner et al. (Gärtner, König, & Wittig, 2001) found that speech input took considerably more time than manual input for simple (2-3 steps) as well as for complex tasks (6-8 steps). The reason is that most of the speech inputs and followed by time consuming speech output, and verifications that needed additional time. Besides recognition errors (error rate 20.6%) they also observed different kinds of user errors. The most frequent ones were vocabulary errors, followed by orientation errors and "Push-to-Talk errors". The authors conclude that speech input is especially beneficial for complex tasks.

Furthermore, speech input can reduce glance times away from the road, improve cognitive workload and lead to improved measures of vehicle control (lateral and longitudinal control) compared to touch-based or manual interaction. (Gärtner et al., 2001; Tsimhoni et al., 2004). But speech interfaces are not necessarily entirely free of visual distraction either. It was shown that people turn toward the screen because the voice system was perceived to be *in* the device screen (Reimer, Mehler, Dobres, & Coughlin, 2013).

It was hypothesized that the potential of speech input lies in more complex tasks that require the user to transfer greater amounts of information in one sentence, since speech is capable of conveying much more information than touch input (Chun-cheng Chang, 2016). However, though not visually or manually demanding, the interaction demands for cognitive resources, which affects the driving performance (J. D. Lee, Caven, Haake, & Brown, 2001). More complex tasks are directly connected with a higher cognitive workload and are perceived as more distracting than simple tasks (Chun-cheng Chang, 2016; J. D. Lee et al., 2001). As a result, speech-based interaction with the in-vehicle system can lead to increased reaction times. Designers should try to create interactions with simple speech commands and no more than three chunks of information for any individual voice task (Chun-cheng Chang, 2016). Large et al. (Large, Burnett, Anyasodo, & Skrypchuk, 2016) found that natural in-vehicle speech interaction with a digital driving assistant elicited similar cognitive demand as a hands-free mobile phone conversation. However, the complexity of the speech commands was low, with no more than two chunks of information (based on examples in the paper).

Summary

In-vehicle speech input has advantages regarding distraction, cognitive demand and efficiency, but they should be used with caution and must be designed carefully with a minimal complexity of the underlying system (J. D. Lee et al., 2001).

Potentials	**Challenges**
Hands on the Wheel	Cognitive distraction
Eyes on the road	No immediate feedback
Little visual and manual distraction	Possibly annoying to codrivers
Efficient text input	Vulnerable to noisy environments
High information bandwidth	Time consuming speech output

Mid-Air Hand Gesture Input

With emerge of new technologies that are capable of detecting human motion without markers, such as the Microsoft Kinect[2], novel input modalities allow consumers to interact with computers using their body as a controller. For example, the Microsoft Kinect is capable of detecting poses and movements of the body, head, hand, and fingers. In comparison to touch

[2] https://developer.microsoft.com/de-de/windows/kinect, accessed on Sept. 5th 2018.

gestures, mid-air gestures are not restricted to be performed on a specific surface, but the user can perform free non-contact gestures in three-dimensional space.

Note: If not denoted otherwise, we refer to *mid-air hand gestures* whenever the term *gestures* is used in this work.

Mid-air gestures have been repeatedly presented as a promising method for interaction with secondary functions in the vehicle while driving ((May et al., 2014; Ohn-Bar, Tran, & Trivedi, 2012; C. A. Pickering, Burnham, & Richardson, 2007)). In particular, gestures have the potential to increase safety by reducing visual demand on the driver (Riener, 2012). In this context, it is very important to distinguish, which form of gesture interaction is used, because they vary in regards of usability and demands on the driver. Existing classifications differentiate between deictic gestures (i.e. pointing) and other forms of gestures, such as iconic, or metaphoric gestures (McNeill, 1992).

Classification of Gestures

For human-vehicle interaction, latter ones can also be classified as symbolic gestures. Symbolic gestures are pre-learned gesture shapes that users have to know beforehand to operate certain functions in the vehicle. An advantage of symbolic gestures is that they can be performed blindly, although it has been shown that control glances occur to check the correct posture and position of the hand. Another downside is the effort for memorizing gesture commands. It becomes more difficult to remember the entire gesture set as the number of gesture-controlled functions increases (C. A. Pickering et al., 2007). An increased learning effort might be acceptable for expert users, but not for the majority of drivers. This limits the number of in-vehicle functions that can be efficiently supported with symbolic gestures.

Pointing gestures, on the other hand, do not need to be learned by the user. Pointing creates a simple deictic reference to all kinds of real and on-screen objects. Users are enabled to interact with a wide range of vehicle functions without having to learn new gestures, which is particularly helpful for novice users (Ahmad & Langdon, 2018). During the execution of a pointing gesture, users must localize a pointing target and make a coordinated pointing movement with their hands. Compared to symbolic gestures, this requires a greater amount of the users' visual attention. However, regarding the advances in autonomous driving and the increasing number of driver assisting functions in modern vehicles, increased visual attention might be acceptable, when user experience, efficiency, and ease-of-use for operating secondary functions are increased in return.

In order to assess the potential of gesture interaction in the car it is important to distinguish the type of gesture interface used. Pickering et al. (C. A. Pickering et al., 2007) distinguishes three types of gesture interface types.

- *Natural hand gesture interfaces* respond to dynamic naturally occurring hand gestures. The benefit is that users require no training and the movement feels natural. Gestures can be easily remembered and therefore require a minimum cognitive workload.
- *Symbolic gesture interfaces* require the user perform pre-learned gestures shapes. Each gesture is usually mapped to a specific command. This allows to directly call a certain function without looking. The downside is that user have to lean the specific set of gestures that the system understands.

- *Sign language interface*s can be used to express semantic information by translating hand gestures into words. These types of gestural interfaces are usually connected with a very high training effort to learn and execute the gesture required by the system and thereby not suited to control in-vehicle secondary functions. Therefore, sign language interfaces will not be further considered in this work.

Another classification of gestures in an application context (C. A. Pickering et al., 2007):

- Direct mapping of a gesture to a function
- Mapping of gestures to physical controls
- Selective mapping of gestures to different functions depending on the context (Recommended due to its smaller, easy to learn gesture set, but requires menu navigation)

Besides the potential for reduced visual distraction (based on the type of gesture interface used) gesture interaction allows flexibility regarding the location of execution of the gestures. This contrasts with device-bound interaction such as remote controllers, or touchscreens. This flexibility enables drivers to perform gestures within comfortable control reach from a normal driving position (ISO 3958). Riener et al. (Riener et al., 2013) investigated execution regions and spatial extend of in-car gestures. They found that the majority of gestures were performed in the region between steering wheel, rear mirror and gearshift, which is in accordance with the suggested interaction area by the ISO.

Gesture Interaction in the Automotive Domain

In recent years, many researchers have presented concepts for the use of symbolic mid-air gestures to control secondary functions while driving. Akyol et al (Akyol & Canzler, 2000) presented an in-vehicle application for controlling messages of different categories, such as traffic updates or emails. They used a set of six symbolic gestures to skip pause and reset the automatic playback of messages. The gesture set is illustrated in Figure 2.8. The system further provided spoken audio feedback of the current message number to provide an orientation aid and to enable interaction with the system without eye-contact. Subjects rated the interaction with the system as natural and intuitive, while the authors further note that dynamic gestures could further “increase the impression of naturalness” (Akyol & Canzler, 2000). Alpern and Minardo (Alpern & Minardo, 2003) also used a limited set of eight symbolic gestures in combination with simple directional gestures and numeric gestures to control a graphical user interface. However, 89% of users tried to interact with the interface by pointing at objects. (The authors later added additional visual cues in the interface to promote the use of the intended

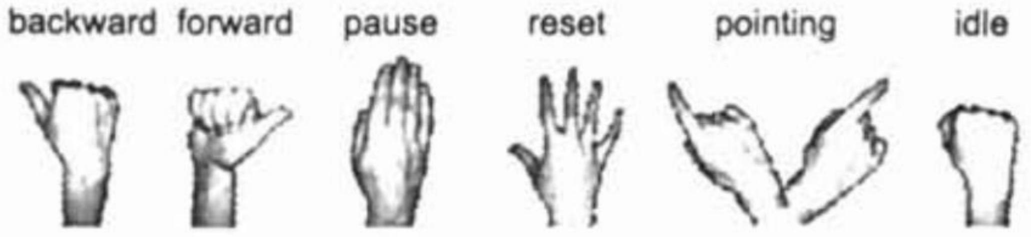


Figure 2.8: Symbolic gesture set for the control of in-vehicle functions (Akyol & Canzler, 2000).

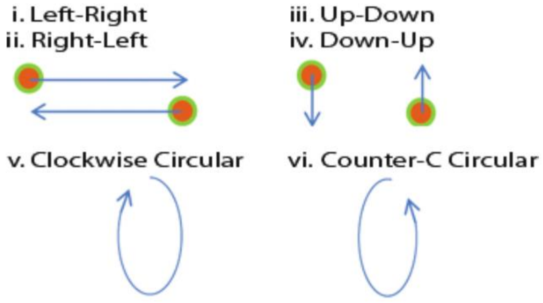


Figure 2.9: A set of six dynamic hand gestures (Ohn-Bar et al., 2012). Circular gestures are also used for index-finger gestures on the steering wheel.

gesture set). This shows that interaction with symbolic gestures is not self-revealing (C. A. Pickering et al., 2007). Visual interfaces rather make people point at objects on the screen.

Ohn-bar et al. (Ohn-Bar et al., 2012) used a gesture set of six dynamic gestures: swipe left, swipe right, swipe up, swipe down, circular clockwise, and circular counter-clockwise (see Figure 2.9). They achieved relatively high recognition rates but also describe confusion between gestures with opposite movement directions. For example, left and right swipe gestures can be misclassified when moving the hand to the initial position of the gesture, or when moving the hand back to the steering wheel after execution of a swipe gesture. Endres et al. (Endres, Schwartz, & Müller, 2011) used a similar set of gestures, but enabled the driver to keep both hands on the steering wheel by only using the index finger gestures. They propose to selective mapping of dynamic finger gestures on vehicle functions.

May et al. (May et al., 2014) present gestures controls for a graphical menu interface based on four dynamic hand gestures. The gestures are all based on movements with the open hand (see Figure 2.10): pushing forward (select), swiping right (go back), holding the hand above and below an idle position (scrolling up, down). Compared to a direct touch interface, this form of gestures interaction led to slower task completion, increased cognitive and physical workload. The latter can be explained by the effort to constantly hold the hand in the air for the whole task time.

Cairnie et al. (Cairnie, Ricketts, McKenna, & McAllister, 2000) introduced finger-pointing to operate secondary controls in the car. The use a screen that is mounted directly behind the display. The driver uses his index finger to move a cursor on the display over the desired GUI element. The selection is achieved with a haptic button on the steering wheel. In another experiment with a very similar setup, users made a forward movement with the index finger to select the desired element (Brand, Meschtscherjakov, & Büchele, 2016). However, this

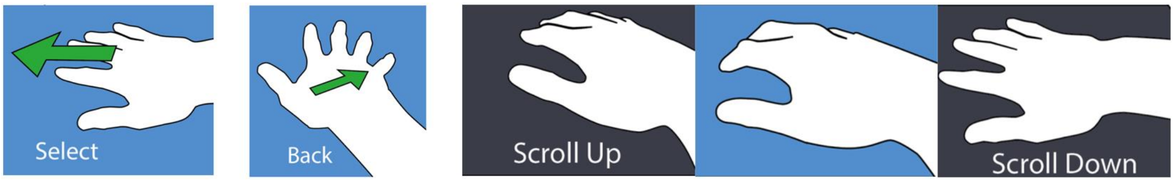


Figure 2.10: Dynamic hand gestures that allow selective mapping of functions depending on the context (May et al., 2014)

selection technique has the limitation that users unintendedly change the pointing direction of their finger during the forward movement, and thus select the wrong element.

Riener and Wintersberger (Riener & Wintersberger, 2011) use an indirect pointing approach. They track movements of the user's extended pointing finger with the hand on the gearshift to control a cursor on a display in the center stack. The also propose a selection by using an additional haptic button, e.g., on the gearshift.

Author	**Gesture Set**	**Recognition**	**Year**	**GUI**
Akyol et al.	6 Symbolic	Visual	2000	yes
Cairnie et a.	Pointing	Electric field sensing	2000	yes
Alpern and Minardo	8 Symbolic, 4 dynamic, 5 numeric	Visual	2003	yes
Riener and Wintersberger	Pointing (indirect)	Proximity	2011	yes
Endres et al.	10 Dynamic	Electric field sensing	2011	no
Ohn-Bar et al.	6 Dynamic (swipes, circles)	Visual	2012	no
May et al.	4 Dynamic	Visual	2014	yes
Brand et al.	Pointing	Visual	2016	yes

Table 2.1. *In-vehicle gesture interaction concepts in literature.*

Summary

Gestures have the potential to increase safety by reducing visual demand on the driver (Riener, 2012). This depends on the type of gestures used. Symbolic gestures are not natural and especially not self-revealing. They need explicit promoting using visual cues (C. A. Pickering et al., 2007). Moreover, most of the presented papers introduce concepts that control functions on a graphical user interface. However, this diminishes the potential safety benefit of using gestures blindly (Endres et al., 2011). Although the (symbolic) gesture itself might not require the user's visual attention, the complete interaction usually does. The lack of tactile feedback, and the replacement with visual cues on the screen further increase the risk for visual distraction. It is important to distinguish between the types of gestures. Strengths and weaknesses are therefore illustrated independently.

Symbolic gestures:

Potentials	**Challenges**
No visual attention needed	Limited set of gestures
3D interaction	Missing tactile feedback
Flexible interaction area (no grasp range problems)	Physically exhausting
Contains meaning	Not self-revealing
	No visibility

Pointing gestures:

Potentials	Challenges
Natural indication of objects	Only deictic information
3D interaction	Missing tactile feedback
No effort for remembering gestures	Physically exhausting
Flexible interaction area (no grasp range problems)	Eye-hand coordination needed
	Limited accuracy

Gaze Input

With the availability and evolution of cameras and eye-tracking technology, it is possible to determine the gaze of a driver on in-vehicle displays. This allows to use the driver's gaze as a "hands-free" input modality.

Many of the presented gesture and speech interfaces have a graphical user interface. While the actual interaction may be performed without looking, users often have to look at the user interface to determine a function they want to control. Therefore, one can argue that most of the speech interaction and gesture interaction concepts are completely "eyes-free". Touchscreens even require the driver to look at the screen, determine a target, take one hand off the screen, and finally reach for the target (not "eyes-free" nor "hands-free").

Selection in gaze-based systems

Another major challenge that gaze-based systems have to deal with is the selection of objects. An explicit selection technique is needed for explicit gaze-based interaction. Natural fixations mainly serve for inspection of the gazed object and are therefore not feasible for object selection, as this would lead to the so called *Midas Touch* problem (Jacob & K., 1991). This means that everything that a user looks at would be selected.

Huckauf and Urbina (Huckauf & Urbina, 2011) summarize techniques for object selection in gaze controlled systems. One possibility is the use of a dwell time. It describes the duration that a user must fixate an object before it is selected. The duration of a dwell time is critical for a successful application. A duration that is too short leads to unwanted selections, whereas a long dwell time slows down the interaction speech. Therefore, dwell times can be an easy solution for novice users, but annoying for experienced users and experts. Despite these downsides, the authors conclude that selection based on dwell times are an effective method for object selection in various applications and settings. However, they did not explicitly refer to applications in the automotive domain. The high visual and temporal demands of the primary driving might impact the suitability of dwell time-based selection in the car.

Gaze Input in the Automotive Domain

Kern et al. (Kern, Mahr, Castronovo, Schmidt, & Müller, 2010) proposed a concept that uses explicit gaze information to identify target GUI elements on an in-vehicle screen. The selection of the targeted GUI element is achieved with a haptic button on the steering wheel. The benefits of this approach are a faster selection compared to speech input and a comparable performance with touch input. Therefore, gaze input might be useful modality when touch screens are not feasible (e.g., because not in hand reach distance).

Müller et al. (Müller et al., 2011) suggest to use gaze input in combination with other input modalities such as speech input, or conventional haptic input elements. Gaze information could be used to identify objects in the vehicle, which could be then manipulated using the additional input element. Ecker (Ecker, 2013) used gaze information to determine the focus of the users' attention for interaction in multi-display environments. Thereby, it is possible to control multiple displays with only one haptic input element. With this approach, task completion times could be reduced, and attractiveness of the system increased compared to manually switching the focus. The implicit use of the driver's gaze did not negatively affect the gaze behavior. On the other hand, Poitschke et al. (Poitschke, Laquai, Stamboliev, & Rigoll, 2011) examined a very similar concept. Gaze input was used to point out a desired item and a barrel key on the steering wheel to manipulate it. The results of their experiment showed that gaze interaction resulted in higher reaction times, higher visual distraction and reduced lane keeping quality. The authors also mention that all results were better for gaze interaction for two expert users.

Summary

Gaze input provides a very fast way to provide deictic information just by looking at objects. However, the use of gaze input as a pointing modality is typically accompanied by a high degree of visual distraction. Moreover, like touch input and pointing gestures, it does not convey information itself.

Potentials	**Challenges**
Hands on the steering wheel	Visual distraction
Efficient pointing	Suitable selection techniques
Passive integration with other modalities	Unnatural to use gaze for controls
	Unfamiliar interaction requires training

2.2.4. Multimodal In-Vehicle Concepts

The last section described the individual potentials and challenges of touch, speech, gesture, and gaze input. The following section presents published research work that integrates at least two of different input modes. The concepts differ in the types of integrated input modalities, as well as their temporal demand and the fusion of modalities. An overview is given in Table 2.2. The subsequent paragraphs briefly describe and each concept and classify them according to the design space for multimodal application, the CARE properties (see Section 2.1.4), and the targeted main benefit (see Section 2.1.3). For easy reference in the following text, each concept is assigned an ID.

ID	Reference	Summary	CARE	Main Benefits
C1	Ford Model U Concept (Pieraccini et al., 2004)	Flexible Choice of speech and touch input.	Equivalence	Flexibility, Efficiency (Naturalness)
C2	A multimodal In-Car Dialogue System (Kousidis et al., 2014)	Flexible choice of speech or nodding.	Equivalence	Flexibility
C3	SiAM – Situation-Adaptive Multimodal Interaction for Innovative Mobility Concepts of the Future(Mitrevska et al., 2015)	Selecting objects by looking at or naming them and modify them using speech or gestures.	Equivalence, Complementarity	Flexibility (Naturalness)
C4	Natural and Intuitive Multimodal Dialogue for In-Car Applications: The SAMMIE System(Becker, Poller, et al., 2006)	Using speech and controller as alternatives or combine them by creating deictic references.	Complementarity, Equivalence	Flexibility (Naturalness)
C5	Multimodal Interaction in the Car - Combining Speech and Gestures on the Steering Wheel Bastian (Pfleging, Kienast, & Schmidt, 2011)	Select with speech and modify with touch gestures.	Complementarity	Efficiency (Naturalness)
C6	Making Use of Drivers' Glances onto the Screen for Explicit Gaze-Based Interaction (Kern et al., 2010)	Gaze input as a pointing device in combination with a push button for confirmation.	Complementarity	Efficiency (Naturalness)
C7	Gaze-based interaction on multiple displays in an automotive environment (Poitschke et al., 2011)	Combination of gaze pointing with a barrel key to confirm and manipulate UI content.	Complementarity	Efficiency (Naturalness)
C8	Der verteilte Fahrerinteraktionsraum (Ecker, 2013)	Combination of implicit gaze input with a barrel key to interact with multiple displays.	Complementarity	Efficiency (Naturalness)
C9	Multimodal Inference for Driver-Vehicle Interaction (Sezgin et al., 2009)	Combination of facial expressions with speech, to increase recognition rates.	Redundancy	Robustness
C10	A Leap for Touch: Proximity Sensitive Touch Targets in Cars(Aslan, Krischkowsky, Meschtscherjakov, Wuchse, & Tscheligi, 2015)	Proximity gestures support touch.	Redundancy	Robustness

Table 2.2. *Summary of multimodal in-vehicle interaction concepts in literature.*

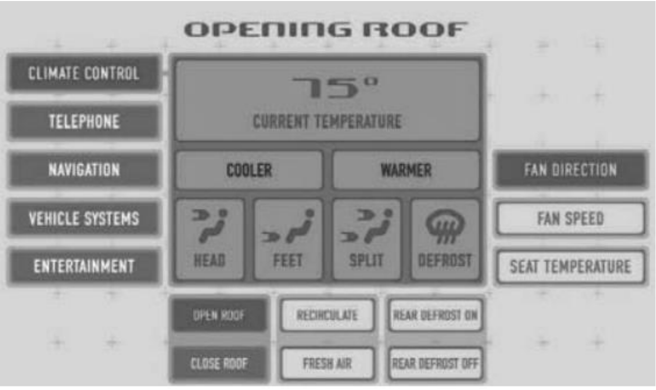


Figure 2.11: In the Ford Model U concept, the interface augments speech interaction by providing the user with additional information such as the state of the dialog, suggestions on what to say (Pieraccini et al., 2003).

Providing Equivalent Alternatives

Concepts *C1, C2, and C3* primarily target an increased flexibility of interaction by providing equivalent modalities. All concepts include speech input and integrate different input modalities in sequential order. C1 and C3 provide independent equivalent input modes (i.e., speech or touch, speech or nodding). For example, C1 enables drivers to switch between speech input and touch input at any point in time of the interaction. The GUI is used to augment speech interaction and provide additional information to the user (see Figure 2.11). C2 demonstrates a system that can interrupt speech output when it detects increased driver workload. The driver can resume the system speech output later with a voice command or by nodding with the head. Participants of the evaluation study were much more comfortable speaking than nodding and were much more forgiving for recognition errors with the speech recognition than for than for head nods (Kousidis et al., 2014). C3 also enables the driver to choose between equivalent input modalities (Mitrevska et al., 2015). At the same time, the concept also uses complementary input of different modalities to build up a command. The driver determines the context of the interaction by gazing at an object in the vehicle he wants to interact with, or by naming it. This is combined with speech commands, which are interpreted based on the set context. For example, the driver can gaze at an object in the environment, such as side rear-view mirrors, and then utter a speech command to alter their position. The equivalence of modalities seems to be the main idea in this concept, since not all interactions require complementary input. Still, this example shows that one concept can be assigned to multiple CARE properties.

Complementary Input

Concepts C4, *C5, C6, C7, and C8* target the benefit of efficiency and naturalness. They all use complementarity of both input devices, whereas one input modality is referencing the object of interest and the other modality is used to modify the selected object. According to this the temporal relation is sequential (selection before modification). This is an approach that is often used is the complementary use of different input modes. The goal is to exploit the individual strengths of the integrated modalities.

C4 provides complementary use of speech and a rotary controller. The controller can be turned and pushed down and sideways in four directions. The functionality is demonstrated at the example of an MP3 player application. The driver can for example start a song by highlighting

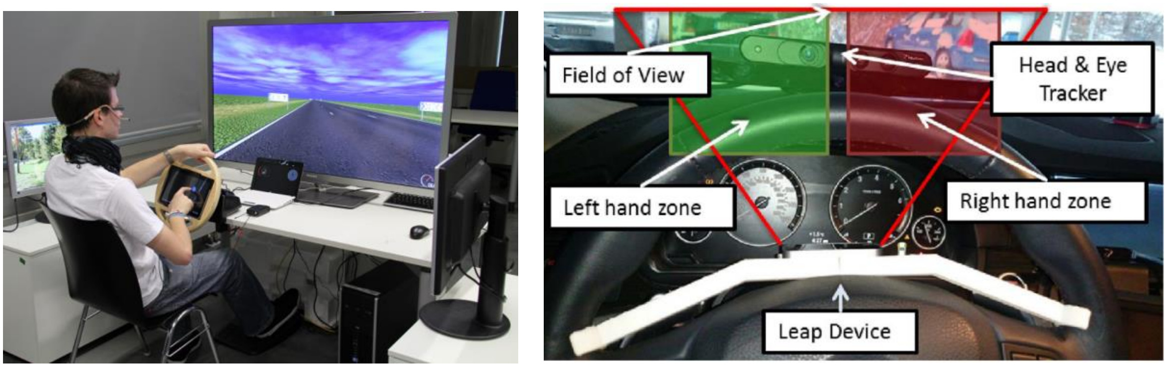


Figure 2.12: *Left*: C5 uses complementary input of speech with touch gestures on the steering wheel (Pfleging et al., 2011, 2012). *Right*: Sensor setup for combining gaze input with gestures on the steering wheel in C3 (Mitrevska et al., 2015).

the song with the controller and then speak a deictic command (e.g. "Play this song") (Becker, Blaylock, et al., 2006)**.** C5 combines command-based speech interaction with touch-gestures on the steering wheel (see Figure 2.12). Speech is used to create a quick reference to interactive objects in the vehicle interior, for example, windows or mirrors. Touch gestures on a tablet on the steering wheel are then used to specify a modification of the referenced object, e.g., swiping down to open a window. The authors mention that the naming of objects is a limiting factor, since it may be hard for drivers to find the correct name for each object they want to manipulate. Therefore, they propose to extend the approach with other modalities, in particular gaze. (Pfleging et al., 2011; Pfleging, Schneegass, & Schmidt, 2012).

Gaze has been used in multiple automotive concepts. Due to the problem of selection in gaze-controlled systems (see Section 2.2.3), it was often complemented with haptic input on the steering wheel. C6 combines explicit gaze input with a haptic push-button on the steering wheel. First, an object on the central information display (CID) is highlighted by looking at it with a spatial and temporal (0.04 seconds) threshold. The driver then uses the button to confirm the highlighted element. An element stays highlighted when the driver glances back to the driving scene before confirming it. This allows the driver to confirm the element even when not looking at the screen. Furthermore it makes it easier for the driver to orientate faster when looking back on the screen (Kern et al., 2010). Similarly, C7 combines gaze input with the barrel key on the steering wheel to interact with different UI elements on the HUD and the FPK. Drivers preselect a UI element by looking at is, push the barrel key to activate it, and then adjust values by turning the barrel key (Poitschke et al., 2011). C8 integrates gaze information for controlling multiple

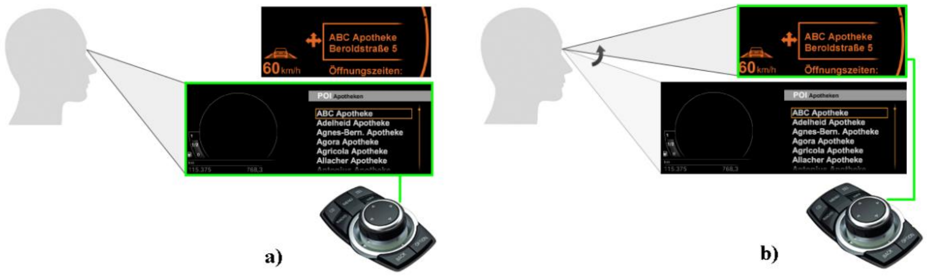


Figure 2.13: The assignment of a single haptic input element changes according to the driver's gaze-based focus of attention (Ecker, 2013).

screens in the vehicle, such as the HUD and the CID using only one haptic input element. A barrel on the MFL is then used to interact with the content on the selected display (Ecker, 2013). This concept is illustrated in Figure 2.13.

Using Redundancy

Finally, multimodal input modes can provide redundant information. Concepts *C9 and C10* focus on the potential of increases robustness of speech and touch input by incorporating other channels and using redundancy of information. Both concepts describe synergistic systems. This means that there is a fusion of two modalities, which occur in parallel or at least temporally overlap. C9 combines speech recognition with passive facial-expression recognition to increase intention recognition accuracy. Both modalities independently predict the driver's intention and are then fused on the decision level. This results in an overall increased robustness of the recognition in noisy environments (Sezgin et al., 2009). C10 fuses gestures proximity information to change touch targets on the screen. For example, touch targets could grow with decreasing proximity of the driver's finger position in mid-air. Thereby, smaller targets can used when the driver is not interacting to create a clean look, but still enable a robust targeting performance for touch input (Aslan et al., 2015).

2.2.5. Goals in the Automotive Domain

The last section described how concepts in literature combine input modalities to exploit the benefits of multimodal interaction in the vehicle. Modern vehicles feature a wide range of devices and sensors, such as haptic buttons, touchscreens, microphones for speech recognition, and cameras to detect the driver's posture, head- and eye-movements or gestures. This provides the driver with a variety of different channels to communicate with the IVIS. This section discusses how drivers can benefit from the technical potential that is offered by the variety of integrated sensors from a human-centered perspective. Based on the benefits of multimodal interaction (see Section 2.1.3), we describe the potentials of multimodal interaction in the context of driving. Figure 2.14 illustrates how the benefits of multimodal interaction could

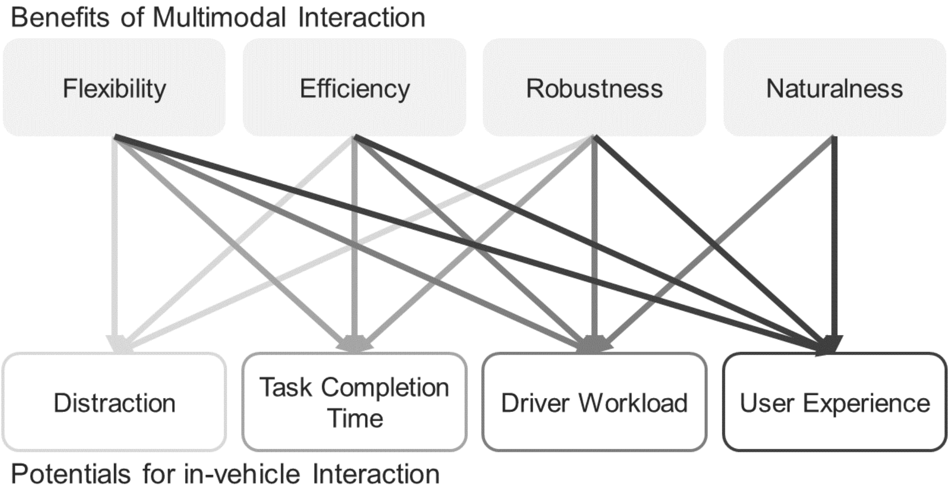


Figure 2.14: The benefits of multimodal interaction have the potential to improve in-vehicle interaction regarding distraction, task completion time, driver workload, and user experience.

improve in-vehicle interaction regarding distraction, task completion time, driver workload, and user experience.

In general, *Flexibility*, *Efficiency* and *Robustness* for multimodal in-vehicle applications are strongly interrelated and enable each other. *Flexibility* to choose alternative input modalities promotes efficient interaction by choosing the fastest input modality for each task type. It can also influence robustness by avoiding error-prone input modalities in certain situations. *Efficiency* of alternative input modes is a prerequisite for offering valuable alternatives and thereby a flexible use of the system. Furthermore, it allows greater robustness by allowing drivers to perform secondary tasks in a short times of low driving demands. *Robustness* of each available input channel enables a flexible choice of full-fledged alternative input modes, whereas the avoidance of time-consuming error correction has a positive effect on efficiency.

Reducing Driver Distraction

Keeping driver distraction low is a key criterion for any IVIS. Multimodal input can contribute to reducing driver distraction in several ways.

Increased *flexibility* allows the driver to choose the least distracting input modality over a variety of different tasks and situations. In this context, it is important to consider visual, manual, or cognitive distraction, as all three sources can have significant impact on driver distraction. Furthermore, multimodal input can increase the *robustness* of interaction by allowing the user to select the most reliable input modality for different situations, but also by fusing different channels to eliminate ambiguities. In principle, greater robustness of the interaction reduces the risk of incorrect input. This is especially critical regarding driver distraction because erroneous input is often connected with confusion caused by unpredicted system behavior and time-consuming corrections. Multimodal error correction provides a faster and more accurate alternative to unimodal error correction (Suhm et al., 2001). However, the potential of increased robustness in the automotive domain should focus on avoiding error and preventing additional driver distraction. For example, certain situations, such as loud background noise in the car might increase the risk of recognition errors. Although speech input might be little distracting in normal conditions, in this situation it could make sense to use another input modality to avoid distracting corrections of erroneous input.

In general, safety critical primary functions, such as steering, accelerating or braking should never be controllable with potentially error-prone input modalities. But also for the control of secondary and tertiary functions, unintended or missing system reactions can cause confusion and distract the driver, as humans naturally tend to look for the impact of their actions or try to correct them (Riener et al., 2013).

Reducing Task Completion Time

Another benefit for in vehicle interaction is reduced task completion time. It has been shown that the potential of multimodal interaction to increase efficiency is mostly limited to very specific task types, such as the manipulation of graphical information (e.g., maps). In particular, multimodal interaction with pen and speech resulted in a 10% reduction for visual-spatial tasks in comparison to speech input only, while the increase was not significant for verbal or quantitative tasks (Sharon Oviatt, 1997). Consequently, researchers concluded that the increase of efficiency is relatively low and thus not considered as one of the main benefits of multimodal interfaces (Sharon Oviatt, 1999).

However, for driver-vehicle interaction, efficiency is significantly more important than in other non-safety critical domains. In the driving context, the duration of an activity is not only a matter of greater productivity and convenience, but also affecting driver distraction (Strayer et al., 2011). The longer the driver is concerned with a secondary task, the longer is he visually, manually, and cognitively distracted. Shorter interactions allow drivers to predict the driving demands during the duration of the interaction, so that they can interact with the system when they predict low driving demands (e.g., while stopped at a red traffic light). With increasing duration of the interaction, the driver's ability to predict the driving demands for the duration of the interaction decreases, as the conditions might change considerably compared to the beginning of the interaction (Strayer et al., 2011). For example, a driver who is stopping at a red traffic light can perform a short interaction, such as selecting a different radio station, within a couple of seconds and complete the entire interaction while still standing. Longer interactions, such as a phone call, can be started while waiting at the traffic light, but as soon as the light changes to green the driver could find himself in the middle of the traffic junction and has to handle both the increased driving demands as well as the phone call. Accordingly, the 15-seconds rule specifies that a maximum time for drivers to complete navigation-related tasks involving visual displays and manual controls in a moving vehicle (Green, 1999). The 15-seconds rule is also adopted and published in SAE guideline (SAE J2364) (Society of Automotive Engineers, 2015).

It can be summarized that efficiency may not be one of the main benefits of multimodal interaction, however reducing the time on task is disproportionately more important in the automotive domain as it is directly linked to the distraction of the driver.

Reducing Driver Workload

The level of the driver's workload represents the perceived difficulty of the task he experiences (Pauzié, 2008). Controlling secondary functions while driving poses demands on the driver's workload. The driver has the choice between two options: a) an increase of cognitive effort resulting in a better dual-task performance, or b) no increase of cognitive effort resulting in a lower level of performance. As users tend to interact multimodally when their own cognitive load increases, multimodal input in the car has the potential to contribute to maintaining task performance at reduced driver workload or enabling a better performance at the same level or workload. Moreover, driver workload is multidimensional with dimension, such as effort of attention, visual and auditory demand, and situational stress (Pauzié, 2008). Accordingly, driver workload in the car is determined by a mobile environment with continuously changing environmental conditions, which influence the different dimension of workload. For example, heavy rain or fog, or conversations with passengers poses demands on the visual and auditory dimensions. Therefore, a *flexible* use of modalities allows to avoid additional load on those dimension that are already claimed and thereby avoid peaks of driver workload. Increased *efficiency* of interaction helps to keep task completion times short and therefore to minimize the duration that drivers are occupied with both tasks. A greater *robustness* avoids the need for error corrections, which can be especially confusing and thus putting high demands on the driver's workload. Finally, *naturalness* of interaction enables an intuitive interaction style, so that drivers do not have any effort for remembering.

Increasing User Experience

Finally, the multimodal interaction has the potential to increase the user experience of in-vehicle interaction. The international standard on ergonomics of human-system interaction defines user experience as "*a person's perceptions and responses that result from the use or anticipated use of a product, system or service*" (ISO, 2008). *Flexibility* allows drivers to choose input modalities depending on their personal preferences, as well as the current task. This often includes *natural* interaction modalities, such as speech, gestures, or gaze input, which can create a positive experience due the novelty and naturalness of interaction. The integration of novel input modalities is especially important regarding the growing variety of in-vehicle tasks. Consequently, drivers face multiple novel use cases. For example, drivers are enabled to control their smart home devices while on their way home, or they can book a table in the closest nearby restaurant. These new uses cases require novel forms of expressive interaction that go beyond classic button- and touchscreen-based approaches. For all use cases, it is crucial to maintain sufficient *robustness*, as erroneous selections can quickly ruin a positive user experience.

2.2.6. Evaluation in User Experiments

This section gives an overview over methods used for the evaluation of interactive in-vehicle information systems. It starts with a description of different test environments and points out advantages and limitations for each of them. Moreover, we present objective and subjective measures that are typically used to evaluate IVISs.

Test Environments

User experiments in the automotive domain can be conducted in several different test environments. Researchers have used laboratory settings, driving simulators, or real-world driving studies depending on the focus of the experiment.

Laboratory Studies

Laboratory studies are the simplest form of test environments in the automotive domain. They do not require a real driving task and therefore enable a low-cost evaluation. The AAM suggests the following approach for early stages of development in a laboratory setup. Participants are put in a dual task situation. There is a primary task and a secondary task. This setup requires the participants to concurrently perform two tasks. The primary "may loosely mimic the visual demands of monitoring a driving-like forward view" (Alliance of Automobile Manufacturers (AAM), 2006), while the secondary task addresses the interaction with the system of interest. A typical verification procedure of such divided attention test is the monitoring of the participants' eye glance behavior. Laboratory studies do not assess driving performance as well as more realistic test environments which provide a more realistic primary task closer to a real driving task.

Driving Simulator Studies

Driving simulators provide a safe and highly controllable environment that allow to investigate the driver interaction in a more realistic context. In comparison to a laboratory environment, driving simulators typically provide a realistic driving task as well as a mock-up that represents the vehicle.

Low-fidelity simulators often consist of a simple desktop setup in combination with a gaming steering wheel and pedals. This provides a simple dual-task environment, which allows to assess primary and secondary task performance. A limitation of low-fidelity simulators is the lack of a realistic geometrical representation of the mock-up. Therefore, ergonomic aspects, such as seating position and hand reach distances to interactive devices and controls can often not be considered.

Mid-fidelity driving simulators feature a more realistic car mockup with controls and a large screen to display the driving simulation. Additional displays behind the mock-up allow provide a wide field of view. In combination with realistic graphics and sound as well as a sophisticated driving model, mid-fidelity driving simulators can create a more immersive driving experience than low-fidelity simulators.

High-fidelity driving simulators include highly realistic vehicle mock-ups and a close to 360° field of view (see Figure 2.15). Dynamic motion systems can additionally provide kinesthetic feedback that creates a more immersive experience for the participants.

In general, driving simulators allow to create a more realistic setting for a user experiment than a laboratory setup. However, even high-fidelity simulators do not reach a completely realistic feeling of driving. Even participants with many years of real driving experience must get used to driving in the simulator, which may result in learning effects. Furthermore, the lack of kinesthetic feedback in fixed-based driving simulators, but also the provision of non-realistic kinesthetic feedback can lead to simulator sickness. Another problem is that participants know that driving errors and accidents do not lead to serious consequences. This can lead to a lower prioritization of the driving task than in the real world. As a researcher it is important to know about these limitations and to consider them during the study design of user experiments (i.e.,

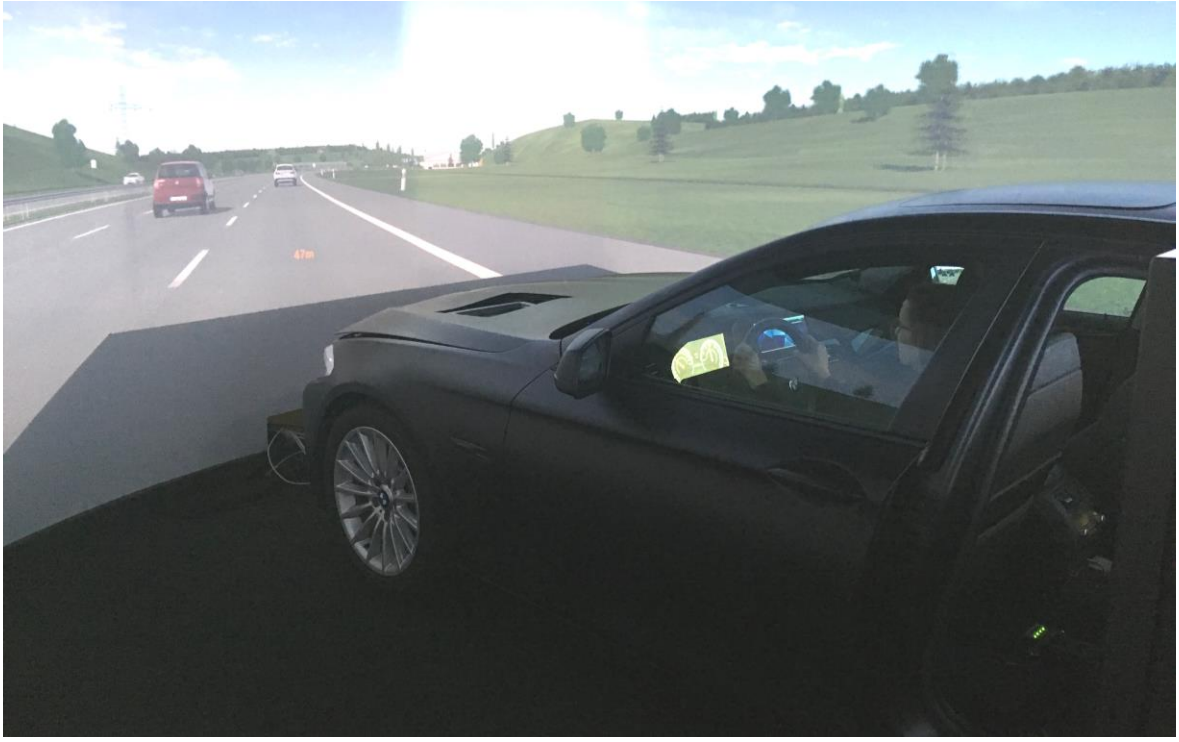


Figure 2.15: This fixed-based simulator with provides a realistic vehicle mock-up of a BMW 5 series in combination with a realistic driving task and a 220° field of view.

by selecting an appropriate environment for the research question), and for the interpretation of empirical results.

Real World Studies

Finally, real world studies are a very realistic way to evaluate IVISs. It can be useful to explore different and new factors that might influence the performance of a system and allows to assess how the system will perform in everyday use. The tradeoff for this realism is that the environment of real-world studies is much more difficult to control. Many factors, such as other traffic on the road, or lighting and weather conditions may influence the evaluation and can hardly be kept consistent across multiple participants. Furthermore, researchers need to consider safety risks. Failure of the system, but also high distraction, or bad usability can lead to critical situations that might have serious consequences for the study participants, or other traffic participants. Therefore, real world studies are not applicable to investigate specific aspects of early stages of a prototype, especially when prototypes investigate potentially distracting interaction techniques.

Measuring Primary Task Performance

The assessment of the driver's primary task performance is important to ensure that interaction with IVISs does not significantly affect the driving task. Lateral control and longitudinal control are two essential safety-critical criteria.

Lateral control describes the quality of lane keeping describes of the vehicle. This reflects the participants ability to anticipate the future vehicle path and make precise corrections (Alliance of Automobile Manufacturers (AAM), 2006). There are several ways to measure lateral control. Some examples are the mean lane position, the time to line crossing, or the number of lane exceedances that occur during one test trial. A lane exceedance is defined by the condition that one of the vehicles tires fully crosses the outer lane markers. A common way to measure lateral control is the standard deviation of lateral position (SDLP) (Green & Paul, 2013). The SDLP is calculated for each trial of one condition. In the results of a user experiment, the SDLP typically represents the mean standard deviation of lateral position over all exposures, i.e., over all trials from all participants for one condition.

Longitudinal control describes the ability to maintain a safe distance to one's own vehicle to other vehicles. Typical longitudinal measures the distance or time gap, and the time to collision (TTC) between the ego vehicle and a leading vehicle. In order to measure longitudinal and lateral control, the AAM suggests using a standard driving context for data collection. Driving should take place on a divided roadway, at a posted speed of about 45 mph, in daylight, on dry pavement and low to moderate traffic density (Alliance of Automobile Manufacturers (AAM), 2006). They also pronounce the importance to explicitly instruct to give highest priority to the primary task of driving (Alliance of Automobile Manufacturers (AAM), 2006).

Critical Tracking Task

The *Critical Tracking Task* (CTT) provides another assessment of a driver's primary task performance. The CTT was first described by Jex et al. as a task "in which a human operator is required to stabilize an increasingly unstable first-order controlled element up to the critical point of loss of control" (Jex, McDonnell, & Phatak, 1966). It was originally intended to be used to measure human performance in a single task setup. The level of instability increases

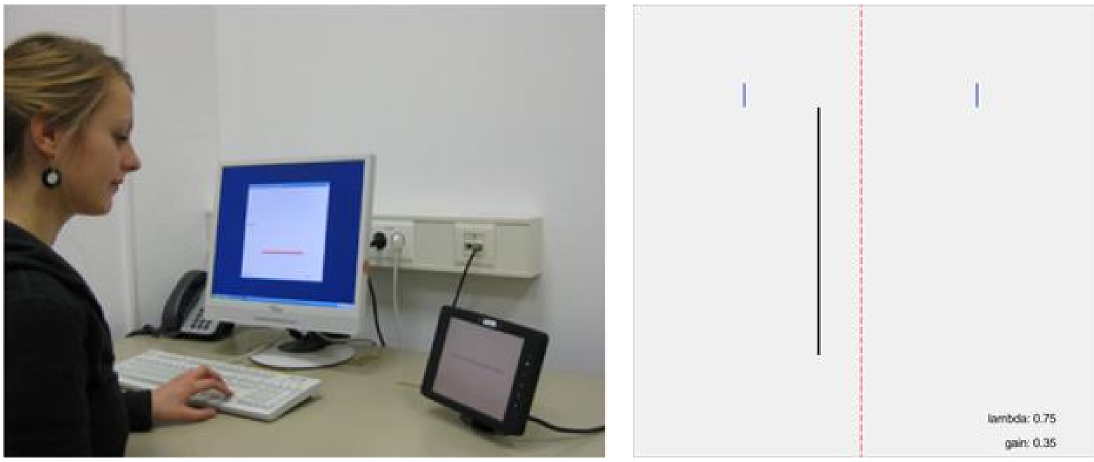


Figure 2.16: Left: The CTT in a simple desktop setup (from Petzoldt et al., 2014). Right: Screenshot of the CTT implementation used in this thesis. The user's task is to keep the black vertical line in the middle of the screen (dotted line).

constantly until a value at which control is lost. This value determines the performance of a participant. It is typically used as a primary task in divided attention setup in laboratory studies. The CTT can further be used as a primary task in combination with a secondary task. In this setup, the level of instability of the task is set to a constant subcritical level in order to assess the workload of the secondary task (Jex, 1967).

Regarding the automotive domain, the CTT and the driving task share similar characteristics. The task poses similar demands as to basic driving skills such as steering, accelerating and braking, which reflect the operational level of driving (Petzoldt, Bellem, & Krems, 2014). Just like those skills, the CTT requires a sufficient amount of attention in order to be operated properly. The withdrawal of attention from the CTT, e.g., due to a highly demanding secondary task, will result in higher deviations from the center position. This suggests using the CTT as a primary task in divided attention setups, to assess the performance and workload of secondary tasks. Petzoldt et al. conducted several experiments to evaluate the validity of the CTT as a means to measure driver distraction elicited by IVISs. They found that the task, reflects manipulations in distraction of secondary tasks. It allows to assess differences between visual and cognitive distraction, and differentiates between IVISs of varying (visual) demand (Petzoldt et al., 2014). However, due to the abstraction of the task, results will not be able to give a direct estimation of real world distraction, but they can represent relative patterns that should be similar in all experimental settings (Petzoldt et al., 2014). Overall, they conclude that the CTT can serve as a valid method to assess driver distraction.

Figure 2.16 shows a screenshot of the implementation of the CTT that was used for experiments described in this dissertation. The task is to keep the black vertical line in the middle of the screen, while it is constantly moving away from the dashed line in the center. It gains speed the further it is away from the center. Participants use their left hand to control two buttons on the left side of the steering wheel to move the vertical line to the left or to the right. This way they can always control the CTT, even while using their right hand for interaction with the IVIS. CTT values range from -100 (the vertical bar is at the left border) to 100 (the vertical bar is at the right border). The CTT performance is measured as the mean absolute position of the vertical bar during one experimental trial.

Eye Glance Behavior

The driver's eye glance behavior is an expressive means to assess visual distraction from the primary task of driving. In accordance with the guidelines presented in the previous section, a typical measure for visual demand is the dwell time. It describes the sum of the duration of all glances within a defined area that contains the visual target (NHTSA et al., 2016). In this work, we use the term *total glance time* (TGT), because the term dwell time is also used in a different meaning later. Other measures are the glance frequency, which describes the number of glances to a target within a defined time sample, or the time off road scene, which is the total glance time away from the road (Society of Automotive Engineers, 2000).

The two most frequently applied approaches of how to measures the users' eye-gaze behavior are head-mounted eye-trackers and remote eye-trackers. Head-mounted eye-trackers (e.g., *Dikablis Live Essentials*) provide a high accuracy independent of the head position of the user, which is especially important for the analysis of gaze behavior from experiments. The tradeoff is a relatively bulky, uncomfortable invasive device. Remote eye trackers (e.g., *SmartEye, Tobii 4C*) do not require the user to wear a special device. They track the users' gaze from a distance, which is more comfortable, but also more sensitive to head movements. Both types have been applied in automotive research.

Measuring Secondary Task Performance

The driver's performance in the interaction with secondary task while driving is depending on the level of interference of primary and secondary task. Typical measures are the task completion time (TCT), efficacy, and error rates. Especially the TCT is a widely adopted measure for efficiency in the evaluation of IVISs, since it is an easy measure to determine crash risk, which can be assessed early in development and thus support the iterative design of an IVIS (Green, 1999).

Measuring Subjective Factors

Some aspects of the interaction between driver and vehicle cannot be captured with the objective measures described above. Therefore, researchers need to consider additional subjective measures, such as driver workload and user experience of the system.

Workload

Driver workload can be assessed with the *Driver Activity Load Index* (DALI). It is derived from the *NASA-TLX* (Task Load Index) method, which was originally designed to assess pilot workload in the aviation domain, with the goal to better reflect workload in the automotive domain (Pauzié, 2008). Several researchers have used the DALI to assess workload for the novel interaction techniques in the vehicle (e.g. (Ecker, 2013; Kern et al., 2010; Pfleging et al., 2012)). Participants rate the perceived demand on seven dimensions: effort of attention, visual demand, auditory demand, tactile demand, temporal demand, situational stress, and interference with the primary task. This multidimensional approach makes it possible to determine the origins of driver workload. An additional global dimension is calculated by the mean value of all other dimensions to represent the overall trend. Participants rate each dimension on a six-point Likert-type scale from 0 (no demand) to 5 (high demand).

User Experience

The ISO 9241-210 defines use experience as *“a person's perceptions and responses that result from the use or anticipated use of a product, system or service“*. A sufficiently high user experience is necessary to create successful products (Schrepp, Hinderks, Thomaschewski, & Germany, 2017). The User Experience Questionnaire (UEQ) is a method that allows a fast and immediate measurement of user experience (Laugwitz, Held, & Schrepp, 2008; Laugwitz, Schrepp, & Held, 2006). It contains 26 items for 6 dimensions (attractiveness, perspicuity, efficiency dependability, stimulation, and novelty). Attractiveness is a pure valence dimension. Perspicuity, efficiency, and dependability are pragmatic quality aspects. Stimulation and Novelty are hedonic aspects.

Additionally, there is a short version of the UEQ (UEQ-S) that only uses 8 items (Schrepp et al., 2017). As a tradeoff the UEQ-S does not differentiate between all 6 dimensions, but only between hedonic and pragmatic quality. It can be used in experiment scenarios when participants are asked to rate the user experience for several variants of a prototype in one session. Keeping the time to complete questionnaires to a minimum avoids that the participant gets stressed, which could affect the quality of answers.

2.2.7. Summary

This section gave an introduction into the automotive domain with a focus on driver-vehicle interaction. While driving, users are in a dual-task situation. In parallel to the primary task of driving the vehicle they control secondary tasks, which may lead to distraction from the primary task. Driver distraction may not only arise from visual distraction, but also manual or cognitive distraction can seriously impair the driving task. Touch, speech, gestures, and gaze are natural input modalities that have been used for in-vehicle interaction. Each modality has individual potentials and limitations. Several multimodal in-vehicle interaction concepts can be found in literature, which use combinations of natural input modalities. The concepts differ in how they combine speech, touch, gestures, and gaze input. In regard of the specific requirements to interaction in the automotive domain, the main goals of multimodal IVISs are reducing driver distraction, increasing task completion times, minimizing driver workload, or enhancing user experience. The evaluation of IVISs is typically conducted in the form of user experiments in laboratory setups, driving simulators, or real-world driving studies. The results give insights about impacts on primary task performance, as well secondary task performance and subjective data.

3. Design Support for Multimodal In-Vehicle Interaction

The previous chapter provided an overview over the two main research fields relevant for this thesis. In multimodal human-computer interaction, many years of research have generated a profound knowledge base with the goal to facilitate the incorporation and application of this knowledge for developers of multimodal systems. While a part of this knowledge is derived from theoretical models, it is mainly based on empirical experiments (such as (Bolt, 1980; Sharon Oviatt, DeAngeli, & Kuhn, 1997)). These experiments serve as a foundation for the development of best practices, principles, and design guidelines, which represent design experience in a structured and condensed form. The automotive domain, however, poses specific requirements to interaction. The user (the driver) is not exclusively dedicated to the secondary task since the primary task of driving must be prioritized. Therefore, drivers have limited possibilities to interact with the system. Consequently, existing design knowledge from multimodal HCI is not directly applicable for in-vehicle interaction. Instead, it must be critically reflected and adapted based on the specific requirements of the automotive domain.

The following chapter describes methods how to provide design knowledge to support the development of multimodal in-vehicle applications. International organizations have published standards and guidelines to guide the development of IVISs. In HCI, design knowledge is rather presented in form of design requirements based on user experiments, which are typically described with the term *implications for design*. We describe different types of such implications for design, their generation process and summarize existing design support for multimodal applications. Another approach to provide design support are design patterns. We give an introduction into the role of design patterns in HCI, different pattern forms and organizations, and present existing pattern collections related to this thesis.

3.1. Automotive Standards and Guidelines

In the automotive industry, official standards and guidelines issued by international organizations are used to support the development and evaluation of novel interaction concepts. The International Organization for Standardization (ISO) has published a number of standards for various aspects of transport information and control systems. Table 3.1 gives a brief, non-complementary impression of some relevant ISO standards for the design of automotive user interfaces.

Standard	Year	Description
ISO 3958	1996	Describes operating distances for drivers and specifies a hand-control reach, as the possible space for manual interaction for the driver.
ISO 11429	1996	Specifies a system's acoustic and visual signals for danger and information.
ISO 15007	2002	Describes measurement of driver visual behavior with respect to transport information and control systems.
ISO 17287	2003	Describes ergonomic aspects of transport information and control systems, including procedures for assessing suitability for use while driving.
ISO 15006	2012	Provides requirements and compliance procedures for in-vehicle auditory presentation.

ISO 15005	2017	Specifies dialogue management principles and compliance procedures to reduce driver workload and to ensure effective and efficient use of transport information control systems.
ISO 15008	2017	Describes specifications and test procedures for in-vehicle visual presentation of information. This includes regulations for the minim dimensions, brightness, and contrast of written text in the vehicle.

***Table 3.1**: This table gives an exemplary, impression of ISO standards covering topics that influence the development of in-vehicle information systems.*

In accordance with these standards, several international organizations have published guidelines for the interaction with non-driving-related tasks in the car. The most influential organizations are the Alliance of Automotive Manufacturers (AAM), the US National Highway Traffic Safety Administration (NHTSA), the Japan Automobile Manufacturers Association (JAMA), the Society of Automotive Engineers (SAE), and the European Statement of Principles (ESOP). Although there are some differences between the specific guidelines of these organizations, they all address similar goals: They were developed as tools for designers of in-vehicle information systems in order to minimize the potential for driver distraction (Alliance of Automobile Manufacturers (AAM), 2006).

The following paragraphs will describe some of the guidelines of the ESOP and the AAM in more detail. They address the following sections (Alliance of Automobile Manufacturers (AAM), 2006; European Statement of Principles (ESoP), 2008):

1. Overall System Design Principles
2. Installation Principles
3. Information Presentation Principles
4. Principles on Interaction with Displays/Controls
5. System Behavior Principles
6. Principles on Information About the System

Each section covers several concrete principles, while sections number three (information presentation principles) and four (principles on interaction with displays/controls) cover the most relevant principles for the topic of this thesis.

3.1.1. Information and Presentation

Guidelines in this section focus on how to present information to the driver in the most effective way and at the same time create a minimum of distraction from the driving task. Systems should not visually entertain the driver while driving but enable the driver to maintain sufficient attention to the driving situation while using the system. For example, the system *"should be designed in such a way that the driver is able to assimilate the relevant information with a few glances which are brief enough not to adversely affect driving"* (European Statement of Principles (ESoP), 2008). The AAM suggests two different approaches how to define the concrete criteria for this guideline. The first one defines that single glance durations should not exceed two seconds and that total task completion should not require more than 20 seconds of total glance time away from the road. Alternatively, the level of distraction can be assessed directly by measuring the driving performance in terms of lateral position on the lane, and distance to the leading vehicle (Alliance of Automobile Manufacturers (AAM), 2006).

Other principles in this section cover the use of icons, symbols, and abbreviations, or the importance to provide relevant information in an accurate and timely manner and to prioritize information with higher safety relevance (European Statement of Principles (ESoP), 2008).

3.1.2. Interaction with Displays and Controls

This section presents principles that focus on the active interaction of the driver with the vehicle.

- *"System controls should be designed in such a way that they can be operated without adverse impact on the primary driving controls"* (European Statement of Principles (ESoP), 2008). This principle intents to avoid interference between primary and secondary controls. This means the system should be controllable in a way that neither hinders, nor facilitates primary control input.
- The *"system should allow the driver to leave at least one hand on the steering control"* (Alliance of Automobile Manufacturers (AAM), 2006). Although the vehicle can be controlled in the safest and most effective way with both hands, having only hand on the steering wheel is momentarily acceptable in some driving situations. This hand should not simultaneously be needed for interaction on the steering wheel, except it is only a single finger input. At the same time, the hand away from the steering wheel must be available immediately if the current traffic situation requires a second hand for steering.
- Speech-based communication systems should *"include provision for hands-free speaking and listening"* (Alliance of Automobile Manufacturers (AAM), 2006), while further actions, such as starting, ending, or interrupting a dialog may be done manually. This principle primarily aims at reducing manual driver distraction of speech input by minimizing additional manual interaction before speaking and listening (e.g. putting on a wireless Bluetooth headset). Cognitive distraction that is caused by listening to the system and processing the output (e.g. (Chun-cheng Chang, 2016)), as well as formulating commands is not respected here.
- Systems should avoid *"uninterruptible sequences of manual/visual interaction"* (Alliance of Automobile Manufacturers (AAM), 2006). Instead, the interaction with system should be interruptible at any point. It should not lead to a loss of driver inputs and facilitate resuming the interaction after the interruption.
- *"The driver should be able to control the pace of interaction with the system"* (Alliance of Automobile Manufacturers (AAM), 2006). In particular, the system should not prompt the driver to make time-critical responses but allow to either not respond at all or to suspend the prompt.
- *"The system's response [...] should be timely and clearly perceptible"* (Alliance of Automobile Manufacturers (AAM), 2006). Any uncertainty (caused by missing or delayed system responses) causes the driver to make a second input and therefore draws attention away from the driving task. The system response time should therefore not exceed 250 milliseconds (according to ISO 15005). Voice-based systems are explicitly not considered as within the scope of this principle.
- Systems that provide dynamic and non-safety-related dynamic visual information must be designed in accordance with the principles above. Moreover, there is an increased risk of driver distraction by moving spatially visual information (i.e., animations), which may trigger the driver to (unintentionally) look at the display. Therefore, the system

should provide a mode in which this kind of information is not provided to the driver, e.g., by dimming or turning off the display.

3.1.3. Voice Recognition Inputs

The guidelines of the AAM and ESOP explicitly state that they do not address voice-controlled systems, due to a lack of comprehensive research and scientific proof (Alliance of Automobile Manufacturers (AAM), 2006; European Statement of Principles (ESoP), 2008). They point out the necessity to develop according guidelines for voice-controlled IVISs in future work. Accordingly, the NHTSA provides more recent guidance for the design of voice recognition inputs with the goal to *"implement speech-controlled in-vehicle systems that have minimal input constraints provide user feedback and have an error handling strategy"* (NHTSA et al., 2016). The general design goals are to reduce the driver's cognitive load to minimize distraction while performing secondary tasks, reduce interaction times, and increase task completion rates. The design should further allow users to form a mental model of the system behavior and enable effective interaction for experienced and new users. Moreover, user feedback should support the correct use of the system (NHTSA et al., 2016). They identify user requirements regarding conversation style, system feedback and error handling:

- The *conversational style* should allow a natural conversation flow between the user and the vehicle that can be paced by the driver. At the same time, the available input vocabulary should be minimized to be consistent with a terse interaction style to keep interaction times short, as research results showed that drivers often tend to issue very terse commands (Aldridge & Lansdown, 1999).
- *System feedback* should notify the user when the system is ready for input, or when input was received. A push-to-talk button followed by a listening tone is recommended based on research, as a clear initiation of the dialog (Lumsden, 2008).
- The proposed strategy for *error handling* is mainly to only use speech input when the consequences of recognitions errors are low, or when the errors can be easily identified and corrected (NHTSA et al., 2016)

3.2. Design Support in HCI

The previous section gave an overview over design support for in-vehicle applications. The following section focuses on general types, sources, and forms of generation for design support in HCI. A central ideal of HCI is that the design of interactive systems is derived from a user-centric design process. Typically, researchers derive implications for design from their empirical findings to support the design of future systems. A well-tried way is to present those implications in form of a set of design guidelines that incorporate the findings from empirical experiments, but there are also other types of design support.

3.2.1. Types of Design Support

There is no formal definition of how the implications for design derived from a study or experiment should be described. Accordingly, researchers interpret the term in different ways. Sas et al. give an overview over different types of implications for design in HCI based on an expert review (Sas, Whittaker, Dow, & Forlizzi, 2014):

- **Abstractions & Meta-Abstractions** capture general functionalities. They are usually shaped as suggestions for a class an existing or new class of technologies.

- **Instantiations of Abstractions** are presented in the form of possible design concepts that aim to support the understanding of the related abstraction and thereby stimulate designers to think about the abstract principles behind the concept and further explore the design space.
- **Sensitizing Concepts as Socially-oriented Design Concepts** capture abstract design knowledge that focuses on specific social needs. They are more high level and thus more open and inspirational than abstractions but less technologically actionable.
- **Descriptions for Communicating Core Findings** are the most frequent type of design implications. They summarize key empirical findings in the form of short descriptions. Thus, they often represent a first step before deriving other types of design implications. Descriptions are sucking and highly situated with the tradeoff of limited generalizability across different settings.
- **Prescriptions as Requirements for Specific implementations** are concrete suggestions for simple implementations that describe how system features can be implemented in particular way. Similar to descriptions, they are highly situated and actionable, yet lack power to generalize to other settings.

3.2.2. Sources for Generating Design Support

There is no defined process or widely recognized standard on how to generate implications for design. Accordingly, they can be derived from a variety of sources (Sas et al., 2014). The most common source are relevant findings in *fieldwork data* from experiments that aim to generate design-relevant knowledge. Design implications can also be derived from existing *design practice*, based on the evaluation of implemented solutions. Other researchers develop design knowledge based on *human science theories* (Stanney et al., 2004). These theories were usually developed outside the field of HCI and include, for example, social theories of human behavior or theories about human perception (e.g., those in Section 2.1.2). Another source for design implications are *social categories*, such as social values, space, and time. For example, implications might address the problem of using technology in public places, such as a crowded subway, in contrast to using it at home. The *technology context*, such as shortcomings and possibilities of a certain technology, is another important source. It enables researchers to frame fieldwork and social categories and puts them in the context of a technological landscape. Finally, *previously developed design implications* can be used by researchers as a basis to build on them and derive novel implications for design.

3.2.3. Design Support for Multimodal Interfaces

The previous two sections described different forms of design support and the variety of sources from which it can be generated. Now, this section presents design support examples for multimodal interaction from literature. They provide crucial know-how for the design of multimodal interactive systems, as the mere fact that a system is multimodal does not necessarily lead to the higher flexibility, efficiency, or robustness. It has to respect empirically validated design support, such as the following examples, to provide the benefits described in Section 2.1.3.

Insights from Empirical Research

Oviatt describes 10 “myths” that exist in the context of multimodal user interfaces (Sharon Oviatt, 1999). These myths describe popular misconceptions and assumptions exist in the context of multimodal systems. She explains and refutes these myths based on empirical key findings from multiple user experiments (e.g. (Sharon Oviatt, 1997; Sharon Oviatt et al., 1997)). The result is a set of valuable insights that lay a foundation for guiding the design of multimodal systems:

1. *If you build a multimodal system, users will interact multimodally*
   User will typically mix unimodal and multimodal expressions. Exceptions to that are task in the spatial domain. Here, the percentage of multimodal input was observed to be much higher (95%) compared to tasks, which do not have a spatial component.
2. *Speech and pointing is the dominant multimodal integration pattern*
   Pointing gestures account for less than 20% of all gestures in combination with human speech (McNeill, 1992). Furthermore, other forms of gestural input, such as symbolic gesturing and facial expressions can generate symbolic information that is more expressive than simple object selection.
3. *Multimodal input involves simultaneous signals*
   Although there is the possibility to give simultaneous input, speech and pen or gesture input do frequently not overlap. Deictic terms are also often not spoken, and if they are, they are not in synchrony with pointing movements. In human communication, spontaneous gesturing often precedes spoken language.
4. *Speech is the primary input mode in any multimodal system that includes it*
   Alternative input modes can not only enhance speech input but also convey important content, which is not present in spoken input at all, such as spatial information or the manner of action (McNeill, 1992)**.**
5. *Multimodal language does not differ linguistically from unimodal language*
   Multimodal language is different to unimodal speech input. It is briefer, syntactically simpler, and less disfluent than users' unimodal speech. Users also avoid speaking error-prone utterances, which can be replaced by other forms of input.
6. *Multimodal integration involves redundancy of content between modes*
   The dominant theme in users' natural organization of input is complementarity of content, not redundancy. Linguists have shown that human communication rarely involves duplicate information in speech and gesture (McNeill, 1992).
7. *Individual error-prone recognition technologies combine multimodally to produce even greater unreliability*
   Multimodal systems can support greater robustness in several ways. Uses can make intelligent decisions about when and how to deploy input modes effectively and thereby avoid error-prone input modalities. In case of a recognition error, they can switch to an alternative input modality to resolve the error in an efficient way. Finally, mutual disambiguation of two input modes can resolve recognition errors from partial information sources.
8. *All users' multimodal commands are integrated in a uniform way*
   Users have very individual integration patterns of multimodal information, e.g., regarding the temporal relation of input modalities. Systems should therefore be capable of detecting and adapting to integration patterns of individual users.
9. *Different input modes are capable of transmitting comparable content*
   Input modalities, such as speech, touch, gestures, and gaze differ in the type of information they transmit. For this reason, they are not fully able to transmit comparable content. Instead, modalities vary in the degree to which they capable of transmitting a

certain type of information. For example, speech and touch-based handwriting input are more comparable than speech and gaze input.

10. *Enhanced efficiency is the main advantage of multimodal systems*
    There are other and more significant advantages of multimodal systems than efficiency. Increased flexibility, reduced error rates, and a more natural form of interaction are important advantages of multimodal systems. Moreover, increases in efficiency have only been moderate and may be limited to certain domains.

This collection of myths provides knowledge from empirical research in a condensed form by providing *Descriptions for Communicating Core Findings* (see 3.2.1). These descriptions summarize empirical key findings from fieldwork data. Thereby, they are often highly situated and thus may have limited generalizability to other settings. For example, we have to respect that these ten myths derive mostly from user experiments that investigated the use of speech in combination with pen input in a desktop setting (Sharon Oviatt, 1997; Sharon Oviatt et al., 1997). It is not clear to which extend these insights can be directly translated to other multimodal applications and settings, due to the wide range of modalities and their very individual attributes and strengths.

Moreover, we must consider the definition of multimodality used. A system, which incorporates gaze information to generate context for speech input is a multimodal system. For example, users who interact with a map-based application will look the parts of the map they are currently referring to and thereby provide information about their current focus of interest. Although these users might not interact multimodally intentionally (user's perspective), the system will process information from different modalities (system's perspective). This applies to any other form of input that is passively integrated, such as head and body position, or lip movements. This example shows that we have to differentiate between the user-perspective and system-perspective of multimodal usage. From the user's perspective, a person might be interacting in a unimodal way with the system, while from the system's perspective the user is interacting multimodally.

Design Guidelines for Multimodal Interfaces

Reeves et al. describe a number of design guidelines for the design of multimodal user interfaces (Reeves et al., 2004). These guidelines represent another type of design support for multimodal interfaces with differs from the descriptive myths of the previous section. They are assigned to six categories:

- Requirements specification
- Designing Multimodal Input and Output
- Adaptivity
- Consistency
- Feedback
- Error prevention and handling

Requirements specification. Multimodal interfaces should support the broadest range of users and contexts of use. Designers have to become familiar with the requirements of different user groups (e.g., based on age, level of experience, cognitive/physical abilities), tasks and environmental conditions, which influence the suitability of different modalities. Based on this knowledge they can support the best modality or combination of modalities for the interaction.

Privacy concerns can have a strong influence on the requirements on the design of multimodal systems. For example, non-speech alternatives should be provided in a public context that allows others to overhear private information or conversations. Therefore, multimodal systems should not use default input and output modalities for certain actions, but rather adapt the communication channel based on the user's wishes in the current context.

Designing Multimodal Input and Output. Multimodal interfaces should respect human information processing abilities and limitations (e.g., attention, working memory, and decision making). Avoid unnecessary fission of output modalities, which requires users to attend to both modalities simultaneously to understand the presented information. Instead, reduce user's cognitive load by maximizing advantages of each modality depending on task and situation and match system output with well suited user input modalities. For example, couple system visual presentation with user's manual input for spatial information and system audio presentation with user speech input for issuing commands. Accordingly, match the system output with the user input style and ensure that multiple output modalities are well synchronized. Share the system state across modalities. Multimodal cues can make the system state more transparent to the user and support users in choosing alternative interaction possibilities.

Adaptivity. Multimodal interfaces should adapt to the needs of the user and to the requirements of different contexts of use, e.g., allow alternative input modalities, such as pen input or gestures to augment or replace speech input in noisy environments. Individual differences of both, users and display devices, determine the quantity and method of information representation. Capturing these factors in user and device profiles can help to support the adaptivity of multimodal systems

Consistency. Multimodal interfaces should provide consistent system output independent of the flexible use of input modalities. For example, a search request should produce identical results for typed input and for speech input. Presentation and prompts should equally share common features as much as possible across modalities, such as using the same terminology across modalities.

Feedback. Multimodal interfaces should make users aware of their current state. They should show alternative interaction options that are available without overloading the user with lengthy instructions. For instance, descriptive icons, such as microphones and speech bubbles can help guiding the user, and earcons can be used to notify the user to begin speaking. Also provide feedback for the whole user input, rather than for each modality individually.

Error prevention and handling. Multimodal interfaces should exploit their potential for error prevention and error handling by integrating complementary modalities that allow use the strengths of each modality to overcome the weaknesses of others. Users should be given the control over modality selection, so that they can use the modality that is least error-prone for the current task and switch to a different modality in case an error occurs. Generally, multimodal interfaces should provide easy options to undo actions and point out clearly marked exits.

These guidelines are an example for *Abstractions & Meta-Abstractions* (see 3.2.1). Compared to the 10 "myths" we described before, they provide a different form of design support by using general suggestions for the class of multimodal systems. They are not limited to specific settings and therefore provide a higher generalizability. The downside of these abstractions is the difficulty to transfer them to applicable design implications for a specific setup

3.3. Design Patterns

The previous section described typical forms of design support in HCI. It showed that they are often limited in regards of generalizability or applicability. For this reason, this section discusses Design Patterns as another method to provide applicable design knowledge for designers and developers. Design patterns are an effective tool to capture and share design knowledge to support the development of interactive systems by communicating solutions to common design problems (Landay & Boriello, 2003). Thereby, patterns aim to prevent developers from "reinventing the wheel" and instead refer to successfully applied solutions.

3.3.1. Development of Design Patterns in HCI

Design patterns originally emerged from the field of architecture. Christopher Alexander first proposed the use of a set of formal patterns as a tool for architects, but also as a way to communicate the design process to all involved stakeholders. For example, the "*beer hall pattern*" addresses the following problem (Alexander, 1977):

"Where can people sing and drink, and shout and drink and let go of their sorrows?"

The solution is:

"Somewhere in the community at least one big place where a few hundred people can gather, with beer and wine, music, and perhaps a half-dozen activities, so that people are continuously crisscrossing from one to another."

The concept of patterns was intended for the reuse of architectural design knowledge, but it found its way into software development and HCI. In 1995, Erich Gamma, Richard Helm, Ralph Johnson and John Vlissides (the of Gang-of-Four) proposed design patterns as a mechanism for expressing design experience in object-oriented software development (Gamma, Helm, Johnson, & Vlissides, 1995). This book inspired the use of design patterns and pattern languages in software engineering as well as in user interface design. The first CHI workshop on pattern languages took place in 1997. The participants explored the utility of pattern languages to support interaction design (Bayle, E., Bellamy, R., Casaday, G., Erickson, T., Fincher, S., Grinter, B., Gross, B., Lehder, D., Marmolin, H., Moore, B., Potts, C., Skousen, G. and Thomas, 1998). At the same time, Coram and Lee presented a pattern language for user interface design (Coram & Lee, 1996). Tidwell presented a comprehensive pattern language for user interface design with a focus on web design (Tidwell, 1997). Many patters and pattern collections are not available in scientific papers. Instead, pattern writers have developed online repositories. They allow to organize, maintain and access patterns more easily for the entire HCI community (Kruschitz & Hitz, 2010a).

HCI design patterns describe recurring problems together with proven solutions. They have also been used as a means to organize, present and structure insights from ethnographic studies and fieldwork (Martin, Rodden, Rouncefield, Sommerville, & Viller, 2001). Patterns have a well-defined form and should be used consistently across a pattern language or pattern collection and thus make it easier for pattern users to understand the problem, context, and solution of a pattern. They may have references to other patterns, when it is part of a collection or a pattern language (Kruschitz & Hitz, 2010b). Moreover, HCI design patterns are concrete enough to provide a actionable solution, but at the same time they are abstract enough to support reusable solutions (Seffah, 2010). Table 3.2 summarizes further benefits of design patterns compared to design guidelines based on the work of Gundelsweiler (Gundelsweiler, 2008).

Guidelines	**Patterns**
Often too simple, specific, or too abstract	Describe an abstract solution to a problem. Additional concrete examples allow to easily comprehend the solution.
Difficult to select	Descriptions of the context and the problem help to find suitable patterns. Moreover, related patterns, which provide solutions to similar problems, can be easily found.
Difficult to interpret	Solutions refer to a concrete problem. Therefore, only little interpretation is needed. Concrete examples point out possible applications.
Difficult to put into the right context.	Description of the context in combination with the examples helps to understand the context.
Possibly contradict each other	Consistent pattern languages should not contain contradictions, because there is only one solution for a problem in a specific context. Discussion or novel better solutions lead to an adaption of the solution.

***Table 3.2**. A comparison between design guidelines and design patters (Gundelsweiler, 2008).*

3.3.2. Pattern Forms.

An important difference between patterns and design guidelines is that patterns are written in a well-defined form. While there is no widely accepted standard, several pattern forms have established.

Alexandrian Form

The original pattern form that was used by Christopher Alexander in the field of architecture (Alexander, 1977). Many authors are using this form, possibly because it was the first form used to encapsulate design knowledge (Kruschitz & Hitz, 2010b).

Picture	An archetypal example of the pattern in use
Introductory Paragraph	The pattern in the context of other, larger scale patterns
Headline	A short description of the problem
Body	A detailed description of the problem
Solution	The solution of the pattern which is written as a design instruction
Diagram	A diagram that sketches the solution
Closing Paragraph	References to other patterns

***Table 3.3**. The Alexandrian pattern form.*

The Gang of Four Form

This form was used by Gamma et al. (Gamma et al., 1995) for describing software engineering design patterns:

Pattern Name and Classification	Short name that shows the patterns use and a classification to patterns with similar problems

Intent	Short statement about the intent and rationale
Also Known as	Alternative names of the pattern
Motivation	A Scenario in which the pattern is applicable
Applicability	The situation in which the pattern can be applied
Participants	Classes and objects that participate in the pattern
Collaborations	Description how participants collaborate
Consequences	Shows how the pattern supports its objectives
Implementation	The implementation of the pattern
Sample Code	Some code fragments of the implementation
Known uses	Implementations of this patterns in use that prove its value
Related Patterns	References to closely related patterns

***Table 3.4**. The Gang-of-Four pattern form.*

Tidwell Form

Jennifer Tidwell presented the first pattern language for UI design. She uses a minimalistic form to document patterns for GUI and web design (Tidwell, 2010).

Name	The pattern's intention and defines a unique identifier
Sensitizing Image	An image to sensitize the reader to the solution
What	A short problem statement
Use When	The context in which this pattern can be used
Why	The design rationale
How	The solution of the pattern
Examples	Screenshots of implementation of the patterns with a short description

***Table 3.5.** The Tidwell pattern form.*

These examples demonstrate how authors tailor the structure of design patterns in order to provide the most relevant information to the targeted audience. Still, most pattern forms share very similar basic structures, although the wording of the different content elements often varies. For example, all patterns contain elements describing the context, the concrete problem and a solution, while each example calls these elements differently. In the end, the decision which form to use is typically based on the authors' preferences (Kruschitz & Hitz, 2010a).

A minimum HCI pattern form

Kruschitz et al. (Kruschitz & Hitz, 2010a) analyzed pattern forms used in HCI. In comparison to software engineering design patterns, HCI patterns are not using content elements such as *Implementation* or *Source Code* (Kruschitz & Hitz, 2010b). The solution should rather suggest an efficient way to resolve the problem, which leaves the designer flexibility to find a concrete implementation tailored to the specific application context. The authors identified a minimal set of common content elements and suggest using this as a basis for HCI design patterns. It is illustrated in Table 3.6.

Pattern Name	A name that gives a hint to the content of the pattern
Sensitizing Image	A screenshot or sketch describing the pattern's idea
Short problem description	A short overview of the problem
Context	Detailed description of the context in which the solution can be applied
Forces	Requirements and constraints relevant to the pattern that help to understand the problem
Solution	The solution in form of a reusable design practice
Examples	Links and screenshots to working instances of this pattern
Related Patterns	References to other patterns

Table 3.6. *A minimum HCI pattern form.*

3.3.3. Organization of Patterns

There are different ways how to organize patterns. One way to organize patters is to group them based on their functional aspects or similar problems (Welie & Van Der Veer, 2003). For example, for web user interfaces these functional aspects may be *searching*, *selecting*, or *layout.* Basically, there are several pattern classifications that can be useful, as the relevance of a classification is depending on the application scenario of the patterns. Pattern groups can be displayed in tables. This provides a good overview over a large number of patterns and support designers to find patterns for a certain problem.

Another way is to illustrate patterns in a graph. An example is given in Figure 3.1. The advantage of a graph is that it better represents the relationships between patterns compared to a table. Each node of the graph represents a pattern. The edges between nodes are the references between patterns. Borchers et al. give a formal syntactic definition (Borchers & O., 2000).

- A *pattern language* is a directed acyclic graph $\mathbf{PL} = (P, R)$ with nodes $\mathbf{P} = \{P_1, \dots, P_n\}$ and edges $\mathbf{R} = \{R_1, \dots, R_n\}$.
- Each node $P \in \mathbf{P}$ is called a *pattern*.
- For $P, Q \in \mathbf{P} : P\ references\ Q \Leftrightarrow \exists R = (P, Q) \in \mathbf{R}$.
- The set of edges leaving a node $P \in \mathbf{P}$ is called its *references*
- Each node $P \in \mathbf{P}$ is itsel a set $P = \{n, r, i, p, f_1 \dots f_i, e_1 \dots e_j, s, d\}$ of a name *n*, ranking *r*, illustration i, problem p, with forces $f_1 \dots f_i$, examples$e_1 \dots e_j$, the solution s and diagram *d*.

The references of a pattern are usually described in the *related patterns* content element in the pattern form. It illustrates how the pattern integrates with other patterns in the collection or pattern language. Typically, this content element does not only name related patterns, but it further classifies their relationship. There are some fundamental relationship that resemble types of relationships know from Object Oriented Modelling (Welie & Van Der Veer, 2003):

- **Specialization**. Patterns can be specializations of other patterns by adding more attributes to the original pattern. This is also called a *"is-a"* relationship.

- **Aggregation**. Patterns can consist of more than one sub-pattern. An aggregation relationship is used to connect them. This is also called a *"has-a"* relationship.
- **Association**. Patterns can have unspecified connections to other patterns, which are not *"is-a"* or *"has-a"* relationships. These connections are also called *"related-to"*.

Graphs can also illustrate different layers of patterns, which go from high-level to low-level patterns. For example, the graph in Figure 3.1 has the levels *business goals, posture*, *experience*, *task,* and *action*. It also shows that a graph might not be the best representation for pattern languages with many patterns, because the graph gets cluttered with increasing amount of patterns and references (Welie & Van Der Veer, 2003).

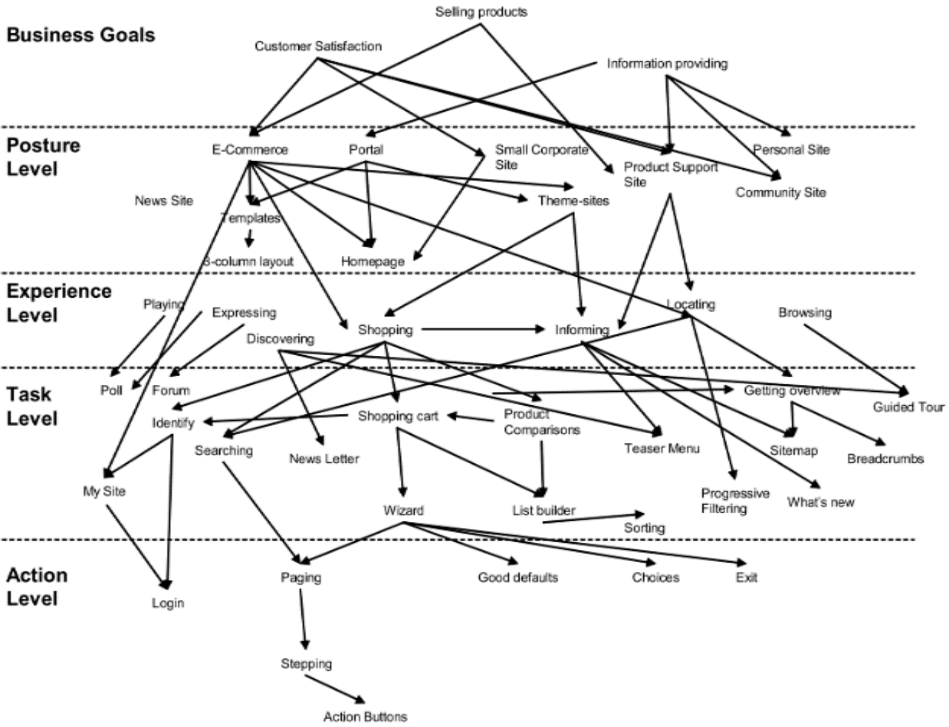


Figure 3.1: Exemplary organization of a pattern language in a directed acyclic graph.

3.3.4. Generating Design Patterns

Design Patterns are derived from proven solutions, which have been successfully applied in existing products or systems (Landay & Boriello, 2003). A typical pattern generation process is therefore a **bottom-up** approach. The pattern author looks for patterns in form of reoccurring solutions in existing products. This process is also called pattern mining. Ideally, a pattern is based on at least three examples that make use of the pattern (Landay & Boriello, 2003).

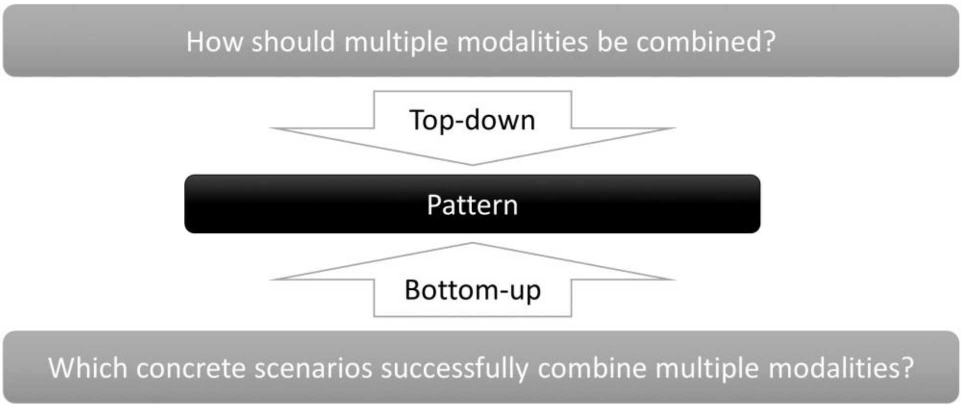

How should multiple modalities be combined?
Top-down
Pattern
Bottom-up
Which concrete scenarios successfully combine multiple modalities?

Figure 3.2: Processes for generating design patterns in HCI.

However, in some cases it is rather unusual to find the exact same information in two or more examples, especially when patterns are used to describe design knowledge that goes beyond the current state of the art (Mirnig et al., 2016). The process of pattern generation can therefore also be a **top-down** process. In this case, patterns are based on general design principles and implicit knowledge based on a researcher's own experience. For example, patterns can be generated based on the results from user studies (Martin et al., 2001). Still, it is important to respect that patterns are not created artificially, but can be traced back on successful products, or concrete know-how. (Seffah, 2010). In practice, both approaches can be combined to determine valid interaction design patterns (Ratzka, 2008).

Figure 3.2 illustrates a top-down and a bottom-up approach for pattern generation for multimodal interaction. The **top-down** process uses design principles and theories to derive a pattern. The **bottom-up** approach looks for several concrete examples of working solutions for reoccurring problems and then derive a pattern based on the common features of these examples.

3.3.5. Relevant Pattern Collections

Within the field of HCI, the predominant domains for design patterns are web design and interface design for desktop applications (Tidwell, 2010; Van Welie & Traetteberg, 2000). In contrast to that and to the best of our knowledge, design patterns have not been used to specifically address multimodal interaction techniques for automotive user interfaces. Yet, there are two relevant pattern collections we want to briefly introduce. The first collection presents design patterns for multimodal interaction in desktop and PDA settings. The second collection serves as an example for the use of design patterns in the context of automotive user interfaces.

Multimodal Interaction Design Patterns

Ratzka presents a pattern collection for multimodal interaction design with a focus on desktop and PDA-based applications (Ratzka, 2013). It is based on a thorough literature review of multimodal interaction, as well as of projects in research and industry. The pattern generation process consisted of both, a top-down and a bottom-up approach (Ratzka, 2013). The resulting pattern collection is illustrated in Figure 3.3. The goal of this collection is to support the development of fast, robust, and flexible systems by providing design support for developers.

It includes novel patterns, but also provides references to already existing related patterns from literature and shows the relationships between them.

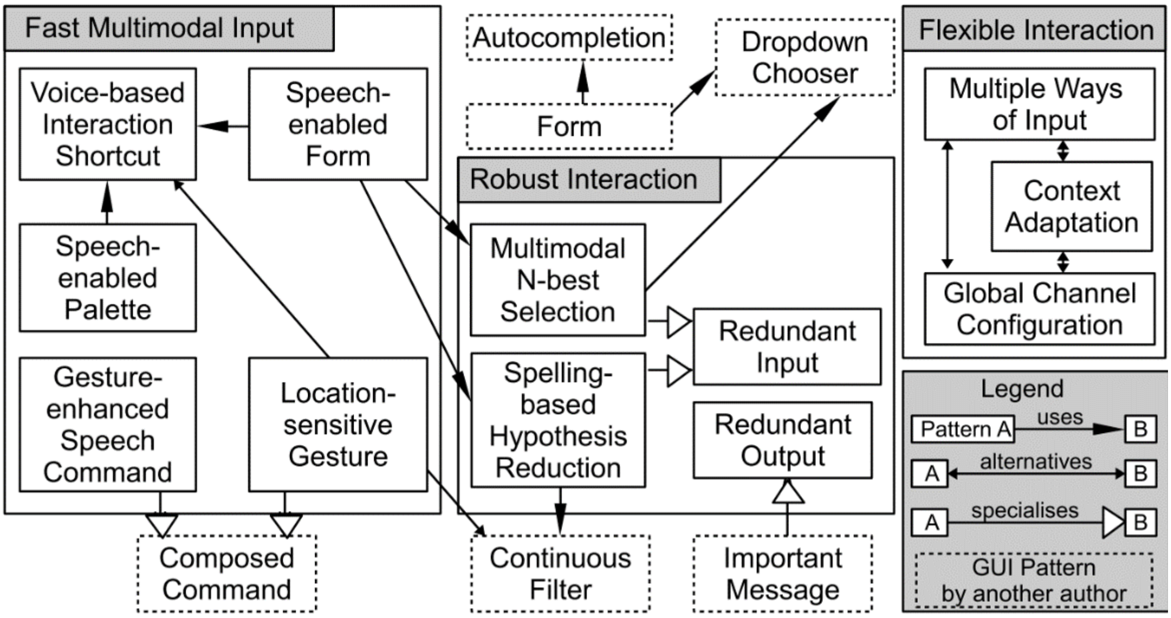


Figure 3.3: A pattern collection for multimodal interaction (Ratzka, 2013).

One small downside of this pattern collection is that it mixes very specific low-level patterns offering concrete solutions, such as “Spelling-based Hypothesis Reduction” and more general high-level patterns such as “Context Adaption”. These different levels should be illustrated in the structure of the collection to facilitate the use for interaction designers. Nevertheless, this collection provides a good example for the structural requirements of a pattern collection in the context of multimodal interaction. In terms of content, it is of limited use for automotive applications, as it focuses on desktop and PDA settings. The author concludes that further research results should be integrated to contribute to the application of design patterns in new fields of multimodal interaction, such as automotive and industry. The special requirements of these fields might identify new patterns or replace or extend existing ones.

Automotive User Interface Design Patterns

In the automotive domain, there is a wide range of norms, guidelines and standards that provide basic human factors and usability rules for in-car interfaces (see Section 3.1). However, these are often rather high-level recommendations with no concrete context of application, or very specific standards that do leave only little interpretation space for designers. Moreover, existing design guidance mainly targets a reduction of driver distraction. Mirnig et al. use design patterns for supporting user experience in automotive interfaces (Mirnig et al., 2016). These patterns were developed in cooperation with designers and engineers in research an automotive industry with a focus on contextual user experience. Their pattern generation process starts by identifying common and reoccurring problems in the automotive domain. Based on that, the researchers searched for existing publications, demos, prototypes, and other sources that present solutions for the identified problems and put into a structured pattern form and refined in an iterative process. Each pattern contains the following elements: *name* and number, *topics*, *problem*, *scenario*, *solution*, *examples*. They present eight patterns as one part of a larger pattern collection:

1. Menu Depth and Number of Options
2. Display Touch Field Size
3. Auditory Information and Warnings
4. Choosing the Best Modality for Warning Displays
5. IVIS System Response time
6. In-Vehicle Display Icon Size
7. Visual Display Color Choices
8. Physical Buttons versus Touch Screen Interfaces

This pattern collection specifically addresses the automotive domain and successfully integrates guidelines from AAM and NHTSA, as well as existing solutions in industry, which have proven to work. Each pattern is described in a structured form and part of a larger pattern collection. Thereby it is good example for the use of design patterns in the automotive domain. At the same time, it also illustrates the lack of current design support for in-vehicle user input. Like many guidelines, the presented patterns focus on the visual presentation of information in the car, while user input, especially natural user input, is hardly considered.

3.4. Summary and Research Directions

This chapter presented existing design support for multimodal application in the automotive domain and in the field of HCI. On the one hand, there are several internationally acknowledged guidelines and standards for the design of IVISs. They mainly focus on information presentation and classic input modalities such as hardware control elements and touchscreens. Multimodal interaction with natural input modalities, such as speech, gestures, and gaze is hardly addressed. On the other hand, there is design support for multimodal interaction. This knowledge has been generated mostly from HCI experiments in desktop settings and varies in form and level of abstraction, its generalizability, or applicability. This has two main limitations in the context of this thesis. First, it is not clear to which extend these guidelines from non-automotive settings can be transferred to multimodal interaction in the car. Second, they are often very abstract descriptions, or too concrete concepts that require to specify the context of use. Both forms do often not provide clearly actionable solutions for interaction designers. Design patterns address this problem by supporting a more structured presentation than typical guidelines, which allows an efficient reuse of the contained knowledge. Furthermore, it is a key aspect of the concept of design patterns that they do not exist individually, but that they are interconnected. These pattern collections describe relationships within patterns and thereby support a larger understanding of the problem space and the role of concrete solutions.

Research Directions

The aim of this thesis is to provide design support for novel multimodal interaction concepts. This chapter showed existing forms of design support for multimodal in-vehicle systems and its limitations. In regard of the presented benefits of design patterns, we argue that a pattern collection for multimodal in-vehicle interaction is a valuable tool to provide empirically validated design support. Regarding the form, organization, and generation process of design patterns, as well as the presented exemplary collections, we identified the following content-related and structural requirements:

Content related requirements:

- Focus specifically on the automotive domain
- Focus on the interaction with natural modalities, speech, gestures, and gaze
- Base on examples and solutions that have proven to work
- Consider existing patterns, guidelines, and industry standards
- Go beyond existing approaches in industry

Structural requirements:

- Describe patterns in a structured form
- Show relationships between patterns
- Cover different layers of abstraction
- Point out which benefits of multimodal interaction are targeted
- Include existing patterns

For the development of such a pattern collection it is essential to respect that the design patterns are not created artificially but derived from existing products or studies that have successfully demonstrated the validity of the solution. However, existing products in the automotive industry cover only a small part of the potentials of multimodal in-vehicle interaction, as current in-vehicle interfaces mostly rely on providing alternative input modalities without much interconnection. As a consequence, there is no comprehensive foundation of existing products or research work in literature that allows to derive design patterns to support developers in designing novel multimodal interaction concepts. The implementation of multimodal interaction concepts and the evaluation in the specific context is therefore an essential step for the development of novel design patterns. We need to understand whether the benefits of multimodal interaction from desktop settings can be applied in the context of driving, which of these potentials could provide the greatest benefit for the driver, and how input modalities can be combined to achieve this?

The following chapters describe the steps, we have conducted to answer these questions. We describe empirical studies to generate a broader understanding of the benefits of multimodality while driving and identify the problems and challenges the patterns should address. Based on that, we create working solutions and prototypes that serve as concrete examples for the patterns and evaluate them in automotive settings. Finally, we use the insights from the user experiments to extend existing design knowledge and to derive a design pattern collection, which fulfils the content-related and structural requirements described above.

4. Investigating the Potentials of Multimodal Interaction in the Car

One of the major advantages of multimodal interaction is an increased flexibility due to multiple alternative input modes. This enables drivers to choose input modalities depending on factors such as the task type or the environmental conditions. Such flexible use of different modalities also requires the driver to switch between the available input modes. In this regard, it is essential to assess the costs that emerge from the process of switching, to estimate the benefit for the overall interaction. Furthermore, flexibility allows to adapt the input mode to changing demands from environmental conditions. However, it is not clear how situational demands effect the input with alternative modalities. This chapter addresses these gaps with two empirical user experiments that focus on assessing the potential of flexibility for multimodal interaction in the car. The first experiment investigates the costs of modality switches between touch and speech input for two different tasks, while the second experiment examines the impact of different situational demands on speech, gesture, and gaze input.

- *Switching Between Input Modalities* investigates the costs of modality switches between touch and speech input for different tasks.
- *Effects of Situational Demands* assesses the benefit of flexibility by investigating the effects on situational influences on the performance of speech, gesture, and gaze input.

4.1. Switching Between Input Modalities

This section is based on the following publication:

Roider, F., Rümelin, S., Pfleging, B., & Gross, T. (2019). Investigating the Effects of Modality Switches on Driver Distraction and Interaction Efficiency in the Car. *Journal of Multimodal User Interfaces*, 1–9. https://doi.org/https://doi.org/10.1007/s12193-019-00297-9 (Roider, Rümelin, Pfleging, & Gross, 2019)

Over the last decades, touch and speech input have found their way into our cars, as the importance of driver-vehicle interaction that goes beyond the actual driving task has increased. The flexible availability of both can compensate disadvantages of a single modality by leveraging advantages from each modality (Ohn-Bar & Trivedi, 2014; Pfleging et al., 2012) and enable users to choose the modality that they assess most appropriate for interaction (Müller et al., 2011).

Parallel availability of different input modalities allows drivers to use the mode that is less occupied by the primary task and thereby reduce driver distraction (C. a. C. a. Pickering et al., 2007). As drivers have to process mainly visual information while driving, many interaction concepts encourage the usage of speech input. It allows drivers to keep the hands on the steering wheel and the eyes on the road while driving. A potential downside of speech input is its short term and sequential nature, which may put a heavy load on human working memory (Bradford & H., 1995). This can cause cognitive distraction, which also influences drivers' visual behavior resulting in inattentional blindness (Harbluk, Noy, Trbovich, & Eizenman, 2007; Strayer et al., 2011). Consequently, the best way to keep driver distraction low is to reduce every form of

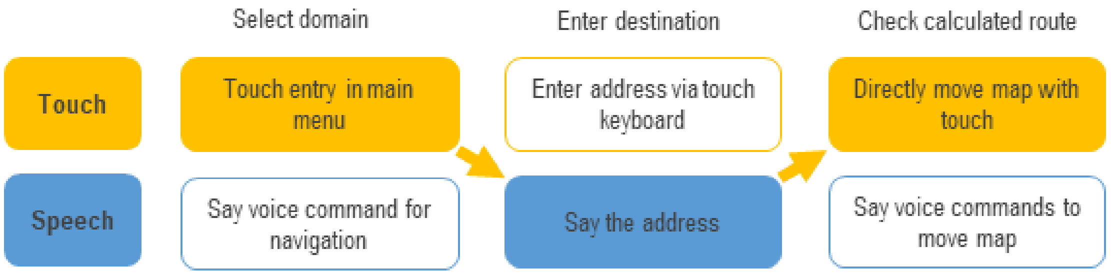

Select domain
Enter destination
Check calculated route
Touch
Touch entry in main menu
Enter address via touch keyboard
Directly move map with touch
Speech
Say voice command for navigation
Say the address
Say voice commands to move map

Figure 4.1: Input modalities in the car vary regarding their suitability for specific subtasks. In order to always use the best input modality, users have to switch.

non-driving-related interaction with the vehicle as much as possible. In this regards, increasing the efficiency of the interaction can reduce the amount of time the driver is distracted from the driving task (Green, 1999).

However, not all tasks in the car can be efficiently completed by using only speech. While speech is a good option in verbal descriptive tasks, it is less suited for expressing spatial information (Sharon Oviatt & Cohen, 2000). Touch and speech input may each be efficient options for some tasks, but less ideal or even inappropriate for others (Sharon Oviatt & Cohen, 2000; Wickens, Sandry, & Vidulich, 1983). Due to the fact that one task might be suited for touch input, but a consecutive task is better served using speech (or the other way around), the user might need to switch between both modalities in order to complete the entire interaction in the most efficient way. For example, starting a route guidance usually requires the user to perform several subtasks, such as selecting the navigation domain, entering the address of the destination and checking the calculated route. Figure 4.1 illustrates this example and how the subtasks could be performed using either touch or speech. While touch allows a quick and easy way to enter the navigation domain, it is easier to speak the address in the next step instead of typing it on a touch keyboard. Finally, moving the map to check the calculated route rather unhandy to do with speech, so it is better to switch back to touch input with allows to directly manipulate the map. In this context, it is essential to assess the costs (e.g., time and distraction) that emerge from the process of switching the input modality. High switch costs could lead to a loss of efficiency or increased distraction and therefore impair the benefit of using multiple modalities.

In this chapter, we present a user study that investigates the influence of modality switches between touch and speech input. We aim to assess the costs and benefits for switching between touch and speech input, by observing how users perform two tasks in various sequences.

4.1.1. Related Work

A number of studies have observed certain costs for switching between modalities. Gondan and colleagues have examined effects for switching between visual and auditory stimuli (Gondan, Lange, Rösler, & Röder, 2004). They found that reaction times were slower after the modality of the stimulus had changed. Similar costs have been observed when switching between input modalities. Zhang et al. (Zhang, Stellmach, & Sellen, 2015) presented an experiment in which they investigated the application of gaze and gesture input for two sequential interaction steps (hover and select). They compared gaze-hover and hand-select with hand-hover and hand-select and observed that the transition from gaze-hover to hand-select took longer than from hand-hover to hand-select.

Furthermore, Monsell has conducted several experiments to examine switch costs that emerge from changing from one cognitive task to another (Monsell, 2003): Participants conducted pairs of tasks of different types in alternating order. Responses on tasks that occurred immediately after a switch took longer and were usually more error-prone. He also notes that knowledge of the upcoming task and time to prepare for it usually reduced the impact of average switch costs. Moreover, there can be physical switching costs that are for instance modeled in the original Keystroke-Level Model, when moving ("homing") the hands between input devices like keyboard and mouse (Card, Moran, & Newell, 1980). Similar costs are reported for moving the hands between different parts of car interfaces, such as the steering wheel an buttons on the dashboard (Schneegaß, Pfleging, Kern, & Schmidt, 2011).

Experiments like these showed that both, the switches between different input and output modalities and switches between different tasks can be connected with a loss of efficiency. Therefore, the efficiency of individual input modalities in the context of a multimodal application has to be considered in regard of the appropriateness of used modalities as well as the costs that emerge from switching between them. While touch and speech are ubiquitous today, there is so far only little understanding for the costs of switching between modalities. We address this gap with our study where we examine the effects of switching between touch and speech input and, thus, contribute to understanding how to efficiently combine these modalities.

4.1.2. Experiment

We conducted a user study with 18 participants, to investigate the influence of modality switches between touch and speech input in a dual-task situation. The participants completed series of alternating tasks that define an entire interaction sequence. The focus of in this experiment was the assessment of costs (especially time) that emerge from the process of switching between efficient input modalities rather than a comparison of touch and speech. We propose three hypotheses to address our research question in a differentiated way:

H1. *Switching to a more efficient input modality leads to increased task completion times for individual subtasks* (i.e., switch costs). The cognitive process of switching the input modality is associated with certain costs. Therefore, task completion times will be longer for tasks that are preceded by a modality switch.

H2. *Switching to a more efficient input modality for changing subtasks increases the efficiency of the complete interaction sequence*. Despite potential costs due to modality switches, the use of more efficient input modalities for tasks will result in a shorter duration for the entire interaction sequence.

H3. *Switching to a more efficient input modality for changing subtasks improves primary task performance.* The use of appropriate input modalities can reduce cognitive workload, which may have a positive effect on the performance of the primary task.

Participants

A total number of 18 participants (13 male, 5 female) with a mean age of 35.2 years (SD = 7.7) participated in the study. All participants had experience with the use of touch-enabled screens. Their experience with speech interaction was lower in general, but only one participant stated that she had never used speech recognition before.

Study Design

We used a within-subject design in this experiment. Each participant conducted ten trials of 90 seconds each, since this duration has shown to provide sufficient information for meaningful CTT calculations without overloading participants (Petzoldt et al., 2014). Table 4.1 illustrates all trials, which differ in the tasks used, the modalities used, and the sequence of tasks. They are divided in four blocks that were permuted between participants to prevent ordering effects.

In the *baseline* block, the experiment contained four trials where the participants did neither change the modalities nor the task type during the run and, thus, repeatedly performed one type of tasks (the first four lines in Table 4.1). In addition, there were two *modality switch* runs during which the task type stayed the same, but the modality changed for each repetition. Also, another two *task switch* runs were included where the modality was fixed for the whole trial, but the task type changed for each repetition. Finally, there were two *combined switch* runs where both, the modality and the task, changed for every repetition: One combined each task with its most suited modality (Move with touch and Describe with speech), while the last one combined each task with the less suitable modality (e.g., speech with the move task). The latter was only included as a matter of completeness but did not reveal interesting insights and will, thus, not be further discussed.

The dependent variables were the CTT performance, task completion times, subjected demand, and suitability ratings. Participants rated the perceived suitability of both tasks when they are performed with touch or speech input at two different points in the study. The first rating took

#	Block	Tasks	Modalities	Sequence
1	Baseline	Move (M)	Touch (T)	MT – MT – MT …
2	Baseline	Move	Speech (S)	MS – MS – MS …
3	Baseline	Describe (D)	Touch	DT – DT – DT …
4	Baseline	Describe	Speech	DS – DS – DS …
5	Modality switch	Move	Touch ←→ Speech	MT – MS – MT …
6	Modality switch	Describe	Touch ←→ Speech	DT – DS – DT …
7	Task switch	Move ←→ Describe	Touch	MT – DT – MT …
8	Task switch	Move ←→ Describe	Speech	MS – DS – MS …
9	Combined switch	Move ←→ Describe	Touch ←→ Speech	MT – DS – MT…
10	Combined switch	Move ←→ Describe	Speech ←→ Touch	MS – DT – MS …

Table 4.1: *The trials differed regarding the conducted tasks and the used input modalities. The execution of a task with a specific input modality is described with both letters, e.g., MT refers to the Move task with touch input.*

place after the first practice of the secondary tasks. The participants rated their first impression of one combination of task and modality, without controlling the CTT at the same time. The second rating took place after all trials had been completed. At this point of the experiment, the participants had gained a better impression about the suitability of the input modes while performing a visually demanding primary task.

Experimental Tasks

Driving a car is an example for a typical dual-task situation. While driving, maneuvering the car is the primary task; other activities such as interacting with in-vehicle systems are secondary tasks (W. W. Wierwille, 1993). We used the Critical Tracking Task (Petzoldt et al., 2014) as *primary task*, which abstracts from the primary driving task. In addition, we used two specific *secondary tasks* to be performed parallel to the primary task. By using abstract tasks, we aim to understand the effects of switching modalities in a more general context.

Primary Task - Critical Tracking Task

The primary task in this experiment was the CTT, which is explained in more detail in Section 2.2.6. It has been shown to be sensitive to changes in the level of demand of the secondary task by imposing a constant visual-manual load on the user (Petzoldt et al., 2014). The CTT has two advantages that makes it very suitable for our study design. First, it produces a uniform demand, which is beneficial when splitting up the interaction sequence in short subtasks. Second, the CTT is a very simple task, which does not require long training phases for the participants. This way, we were able to keep explanations and training phases short and still avoid learning effects.

Spatial Secondary Task - Move

The *Move* task is illustrated in Figure 4.2. It is a spatial task and thereby potentially well-suited for touch input (Koons et al., 1998; Wickens et al., 1983). The goal is to move the two colored shapes from the center area to corresponding placeholders on the outside. A real-world equivalent for this abstract task could be the panning of the navigation map to find the desired location. When using touch input, participants moved the elements by dragging them with their fingers to the target position on the touch screen. For speech input, people selected the elements

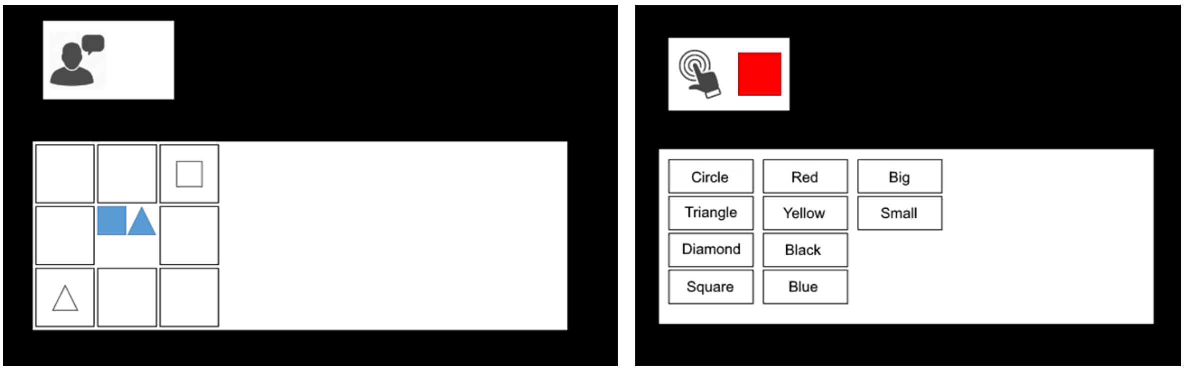


Figure 4.2: Move task (left): both shapes from the center area must be moved to corresponding fields (triangle to bottom-left and square to top-right). The icon on the upper left instructs the modality to use. Describe task (right): the element on top of the screen must be described by shape, color, and size.

by naming their shape and then described the location of the target fields to move them (e.g., "Circle to top-right"). Participants were free to move either the left or the right element first.

Verbal Secondary Task - Describe

Describe is a verbal task which exploits the strong descriptive capabilities of speech (Sharon Oviatt, 2012). Participants have to describe the element that is displayed on top of the screen by characterizing it by three attributes: shape, color, and size. A real-world equivalent is the input of an address for a navigation system or any other task where information has to be verbalized. Just like the colored elements, an address can be split into several attributes: the city, the street, and a house number. When using *speech input* for this task, participants could simply name the attributes of the element (e.g., “small yellow circle”). After a description was completed, the next task appeared. When using touch input for this task, the screen displayed a selection of possible attributes as illustrated in Figure 4.2. Each column represented the possible selections for one attribute. The participants had to touch the correct buttons in each column. While speech input was used, the interface did not display buttons with possible options, since we did not want the interface to influence what participants say.

Apparatus

The study was conducted in a laboratory setting consisting of a car seat, a steering wheel with buttons, and two displays as shown in Figure 4.3. A display at the position of the windshield showed the CTT screen. A touch-sensitive display to the right of the steering wheel displayed the secondary tasks. The size and position of the interaction area on this display were chosen according to typical arrangement of touchscreens in premium class cars. We implemented the tasks in am HTML5/JavaScript-based test framework with full functionality for touch input. Since speech commands for both tasks were relatively long and complicated, we decided to use a Wizard-of-Oz approach with a keyboard-based interface that allowed to quickly execute users’ voice commands. Based on original literature (Petzoldt et al., 2014), we implemented our own version of the CTT in Unity3D, which allowed us to trigger data logging for runs automatically via the framework. We recorded average CTT deviations over the whole duration of each trial as an indicator for primary task performance. The framework recorded the average task completion times (TCT) for a task depending on the input modality and the sequence of tasks in which it occurred.

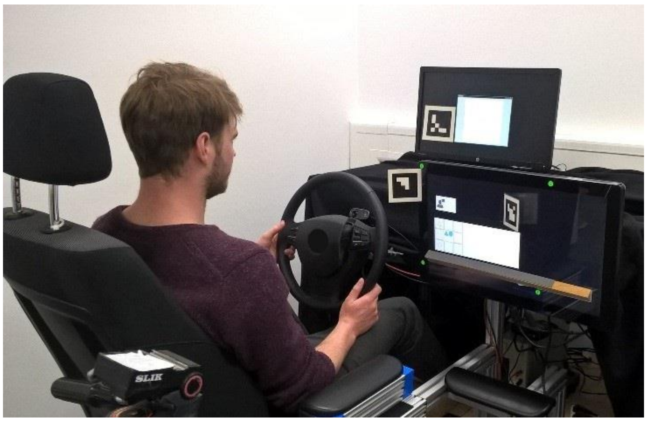


Figure 4.3: The experimental setup for the user study included a primary task in front of the user and a secondary task on the display to the right of the user.

Procedure

After a short introduction, the participants filled out a consent form and a questionnaire capturing demographic data and adjusted the seat position. The examiner presented the CTT as the primary task and explained its functionality. The participants conducted a trial with the CTT (without a secondary task) to get used to the controls. Then the examiner introduced the secondary tasks and explained the interaction with both modalities. Participants had a few minutes to practice both tasks with both input modalities (without CTT). During the trials, participants were instructed to have their primary focus on the CTT. At the same time, they should try to complete as many secondary tasks as possible in the given time.

4.1.3. Results

Based on literature we assumed that the spatial move task could be performed more efficiently with touch, whereas the verbal describe task, would be more efficient with speech input. Therefore, in a first step, the analysis of the baseline trials successfully demonstrates the validity of these assumptions in the specific setup of this experiment. The baseline trials are defined by two independent factors: the task (Move, Describe) and the modality (touch, speech). This refers to lines one to four in Table 4.1. A two-way repeated measures ANOVA indicates a strong interaction between both factors ($F(1,17) = 295.24, < .001, \eta_p^2 = .95$). Follow-up t-tests show that touch input (*M = 3.99, SD = 1.12*) was faster than speech (*M = 6.82, SD = 0.72*) for the spatial Move task ($t = -10.27, p < .001$). Accordingly, speech input (*M=4.20, SD=0.41*) was better than touch (*M=6.50, SD=1.50*) for the verbal Describe task ($t = 7.84, p < .001$). This is illustrated in Figure 4.4.

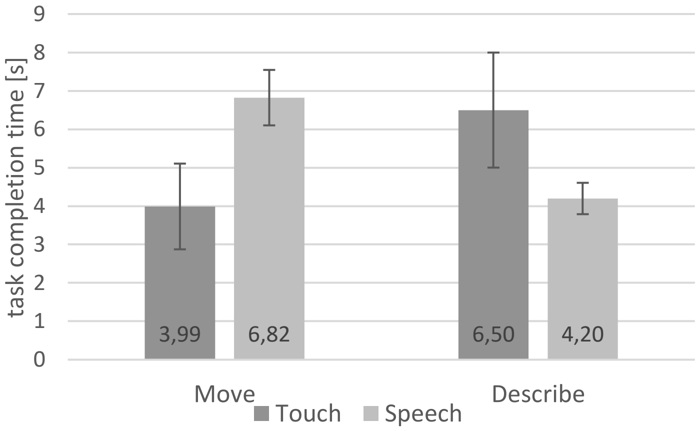


Figure 4.4: Task completion times from the baseline trials demonstrate that touch was the more efficient input for the Move task (left) and speech for the Describe task (right).

Switch Costs for Individual Tasks

The hypotheses in this experiment focus on the effects that emerge from switching between the most efficient input modalities for alternating tasks. Therefore, the remaining part of this section will concentrate on the performance of the Move task with touch input (MT) and the Describe Task with speech input (DS). We use the term tuple to describe these combinations of a task with an input modality. The next step is to determine the switch costs by investigating how these efficient tuples perform in different sequences.

The following analysis uses two independent factors: the **tuple** (MT, DS) and the **sequence** in which the tuple occurs (baseline, modality switch, task switch, combined switch of task and modality). In Table 4.1, this refers to lines one, five, seven, and nine for MT, and lines four, six, eight, and nine for DS.

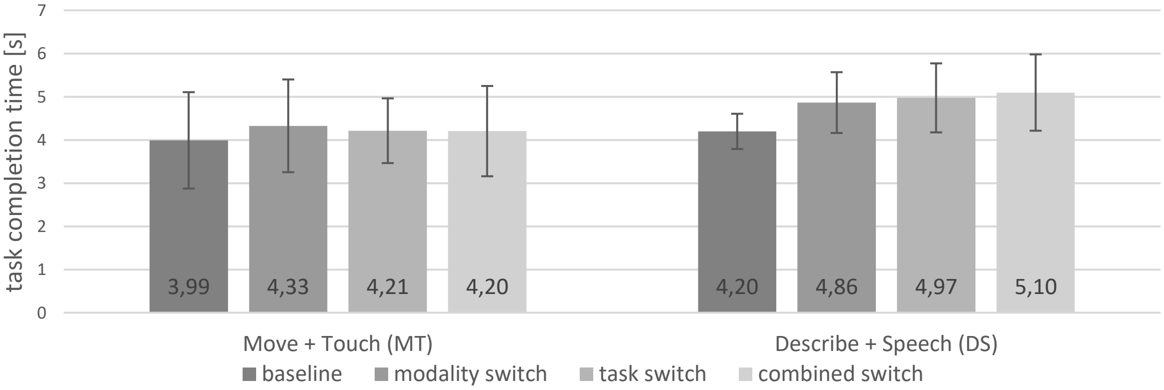


Figure 4.5: Task completion times for MT and DS. Modality switches led to an increase of TCT when the task stayed the same. In contrast, modality switches during a task switch did not induce additional costs compared to only task switches. Error bars indicate the standard deviation.

Figure 4.5 illustrates the task completion times for MT and DS in different sequences. It can be observed that the completion times for both tasks were shortest in the baseline trails compared to all other conditions. A two-way ANOVA showed that the task completion time was mainly effected by the tuple, $F(1,17) = 12.67, p < .01, \eta_p^2 = .427$, but also by the sequence of task in which the tuples were performed, $F(3,51) = 5.734, p < .01, \eta_p^2 = .252$. Bonferroni-corrected pairwise comparisons revealed a significant rise of TCT for modality-switch trails compared to baseline trials ($p < .001$). TCT was also increased in task switch trials in comparison to baseline trials ($p < .001$). Finally, the combined switch of the modality and the task type also resulted in a significant increase of TCT compared to baseline trials ($p = .010$). The processes of switching only the modality, switching between tasks with consistent input modality, and switching between task and input modality at the same time all led to impairments of efficiency compared to the execution in baseline trials. Surprisingly, the costs for modality switches and task switches did not add up when both occurred at the same time. TCT in combined task and modality switch trials were not longer compared to TCT in task switch runs ($p > .05$).

Interaction Sequence Efficiency

In the next step, we focus on the efficiency of completing an interaction sequence that consists of different tasks. We compare the summed task completion times for completing both tasks depending on the used modalities: only speech input, only touch input, or task-depending input of both (touch for Move and speech for Describe). This refers to lines seven, eight, and nine in Table 4.1. Figure 4.6 illustrates the summed task completion times for both tasks.

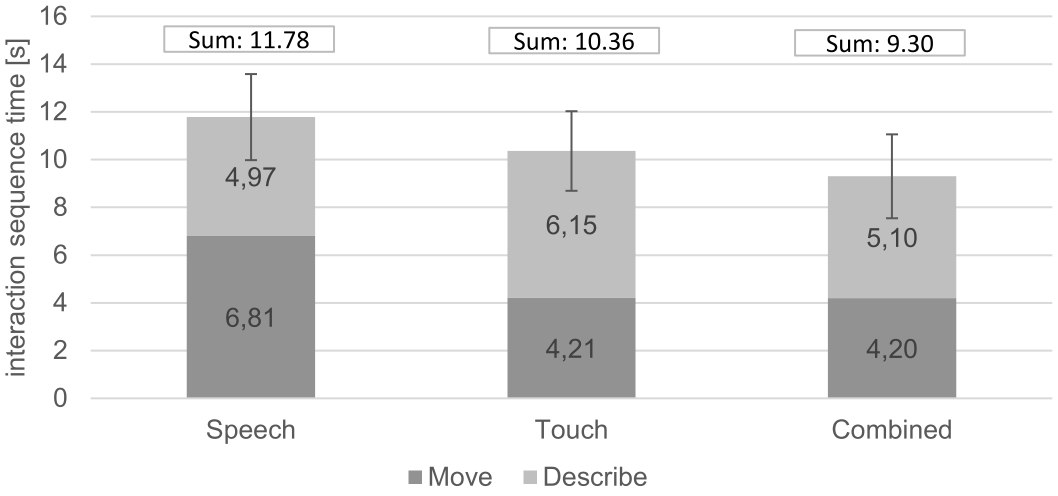


Figure 4.6: The use of both input modalities for a sequence of both subtasks was faster than using only speech input or only touch input. Error bars indicate the standard deviation.

Based on the completion times for individual tasks in each trial, we calculate the average time to complete a sequence with two different subtasks. The average time to complete both tasks was greatest using only speech input (*M=11.78, SD=1.80*), followed by touch input (*M=10.36, SD=1.67*). The combined use of both input modalities was the most efficient input form to complete both tasks (*M=9.30, SD=1.76*). A one-way repeated measures ANOVA showed that the effect of the used modalities is significant and can be classified as a strong effect, $F(2{,}34) = 19.990, p < .01, \eta_p^2 = .540$. All pairwise comparisons were significant. The combined use of both input modalities reduced interaction time by 21% compared to only speech input ($p < .001$), and by 10% compared to only touch input ($p < .01$).

Primary Task Performance

CTT Performance

We examine the CTT performance depending on the used modalities: only speech input, only touch input, or task-depending input of both. Figure 4.7 shows the average CTT values for those conditions with only speech (*M=4.76, SD=1.69*), only touch (*M=8.96, SD=4.20*), and the combined use of touch and speech (*M=7.26, SD=3.54*). A one-way repeated measures ANOVA shows that the used modalities had a significant effect on average CTT deviation, $F(2, 34) = 16.64, p < .01, \eta_p^2 = .495$. Pairwise tests (bonferroni-corrected) show that the combination of touch and speech was less distracting than only touch ($p < .05$), but also more distracting than only speech ($p < .01$).

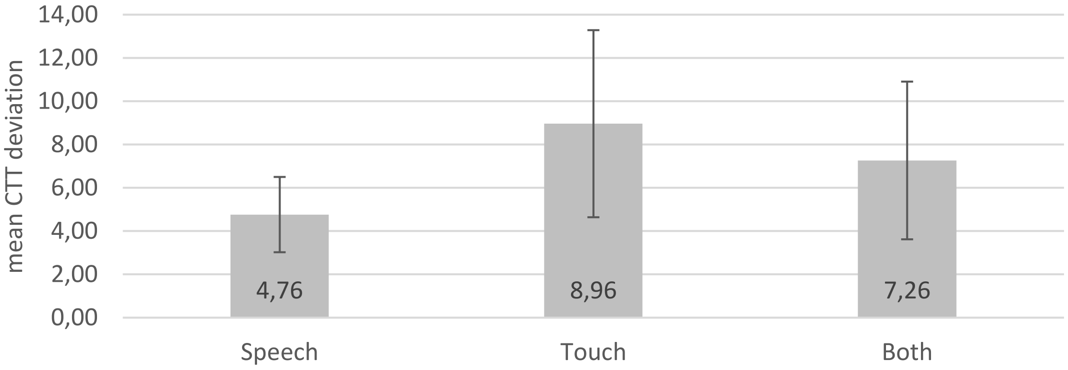


Figure 4.7: Mean CTT deviations for input with only speech, only touch, or touch and speech input when alternating between the two tasks. Error bars indicate the standard deviation.

Glance Behavior

Furthermore, we assess visual distraction based on the average total duration that participants glanced on the screen during the execution of a subtask. Figure 4.8 illustrates the total glance times for the different input modes over both tasks. Speech input resulted in the least amount of visual distraction (*M=1.34, SD=0.32*), while touch input results in longer glance durations on the screen (*M=2.76, SD=0.33*). The combined use of both modalities led to glance duration between touch and speech (*M=1.68, SD=0.38*). The used modalities have a very strong effect on the total glance duration, $F(2,34) = 133.926, p < .001, \eta_p^2 = .887$). In particular, the values for speech, touch, or the combined use of both modalities, all differ significantly ($p < .01$).

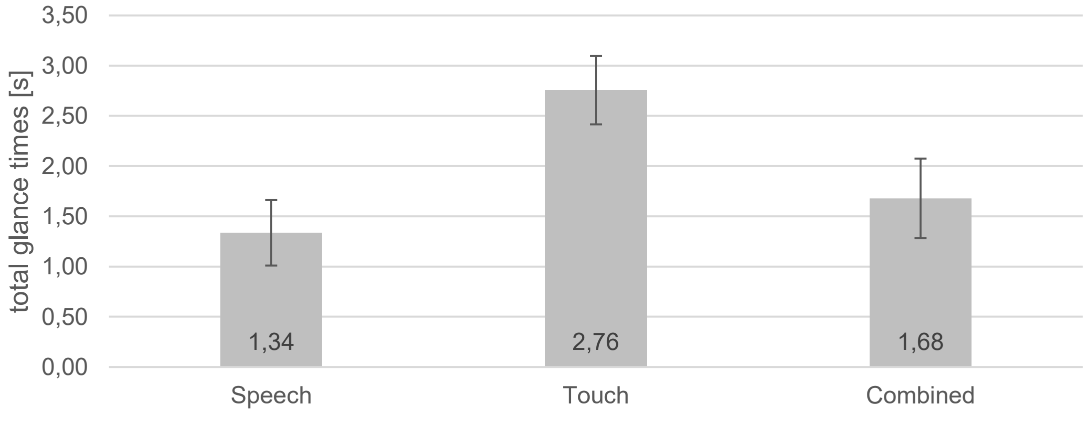


Figure 4.8: The average total glance time on the screen during one subtask was lowest for speech. Touch requires more visual attention. The combined use of speech and touch was between the individual modalities.

Subjective Demand

The results of the participants' ratings for each dimension are shown in Figure 4.9. The data is distributed normally. Accordingly, it shows that touch input (*global: M=2.56, SD=0.95*) led to higher perceived cognitive demands than speech input (*global: M=1.59, SD=0.90*). The only exception is the demand on the auditory dimension, for which speech input was rated most demanding. Combined use of touch and speech falls between both individual modalities (*global: M=2.08, SD=1.09*). A one-way ANOVA shows that the input modalities significantly influence the global dimension $F(2,34) = 15.602, p < .001, \eta_p^2 = .479)$. Speech input causes significantly lower demand than touch input ($p < .001$), but the advantage over the combined use of both modalities is not significant.

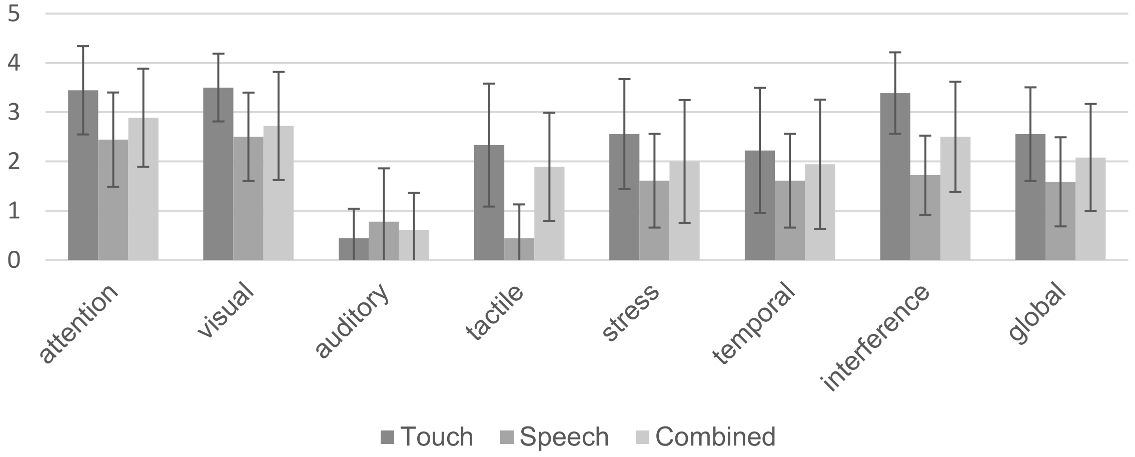


Figure 4.9: The DALI questionnaire assesses the participants' perceived workload. Touch input leads to higher ratings than speech input on all dimensions (except auditory demand). Error bars indicate the standard deviation.

4.1.4. Discussion

H1 addresses the costs of modality switches. We claimed that switching to a more efficient modality leads to increased completion times for individual tasks, due to the process of switching. Our results showed that modality switches caused an increase of task completion time compared to baseline runs. However, in an interaction sequence that requires the user to change between different tasks - which means that there are already costs for task switches - additional modality switches could be performed without a loss of efficiency. The process of switching between efficient modalities did not induce any additional time for task completion and we reject H1.

For *H2*, we assumed that switching to a more efficient input modality increases the total efficiency of an interaction sequence with different tasks. A simple example for such an interaction sequence is illustrated in Figure 4.6. The results show that touch was more efficient for Move and speech for Describe. Moreover, in the previous paragraph we concluded that switches between both modalities can be made without loss of efficiency. Consequently, switching between modalities compared to using only speech or only touch increases the overall efficiency and we accept *H2*. While this result was to be expected, the important insight is that the benefit regarding efficiency of the entire sequence compensates for any loss of efficiency due to the process of switching between subtasks.

In *H3* we claimed that switching to a more efficient input modality could reduce cognitive workload and therefore increase the performance of the primary task. Switching between touch and speech was more distracting than only speech input, but also less distracting than only touch input. This can be explained since participants spent about half of the duration of the trial with either modality. This implies that the process of switching the input modality did not result in additional distraction. The CTT deviation was mainly influenced by the used modalities and not by the modality switch itself. While the CTT results did not directly show that the efficient use of touch and speech reduces distraction from the primary task there might still be an indirect effect. Distraction might not be smaller when switching between modalities, but the time the driver is distracted from the primary task will be shorter. However, the best primary task performance was achieved using only speech and we reject *H3*.

Furthermore, *H3* based on the assumption that the combined use of touch and speech could reduce cognitive workload. However, the results of the DALI indicate that the use of touch input always led to higher ratings than speech input. The combined use of both modalities consequently led to higher cognitive workload than speech only.

The small effect of the modality switches in our setup might be also influenced by the fact, that average switch costs are usually reduced when knowledge of the upcoming task exists (Monsell, 2003; Müller et al., 2011). In our study, participants had knowledge about upcoming tasks and modalities. We assume that this knowledge had a similarly positive effect on TCT, according to knowledge of the upcoming task. Conversely, not knowing the modality for an upcoming interaction step might result in increased TCT, as the user has to determine the requested input modality first. A user interface could require the user to switch to a more efficient modality, without increasing TCT, if the user has prior knowledge of this switch.

The basic assumption for the hypotheses of this study is that some tasks are better suited for either speech or touch input, based on the specific requirements of a task and the different ability of touch and speech to transmit information (Wickens et al., 1983). The results from our study justify this basic assumption: Touch was the more efficient input modality for Move, whereas speech input was more efficient for Describe. The suitability ratings after a first impression of the tasks also reflects this relation. However, there are significant differences to the perceived suitability after all trails had been completed with parallel execution of the CTT. Overall, touch is still rated as the more suitable input for move and speech as more suitable for describe. Still the ratings illustrate the effect of a visually demanding primary task on the suitability of these input modalities to operate secondary tasks. They also show that drivers might assume that a touch input is a suitable modality for certain tasks in the car but change their minds after trying it out themselves: a potential for dangerous situations. Therefore, giving drivers to choose between alternative input modes freely might not always be ideal. Drivers could instead benefit from some sort of guiding of the system that promotes suitable input modalities.

Limitations

With respect to our results, it is important to respect the factors that our particular experimental apparatus entailed. Namely, the abstraction of the CTT for driving, as well as the Move and Describe tasks as reasonable stand-ins for subtasks within a particular interaction sequence. The findings may therefore not apply to different types of secondary tasks or to other input modalities, such as gestures or steering wheel controls.

The alternation between two tasks represents only an approximation to real-world interaction steps. In-car interactions are often composed of more than just two tasks, but still shorter than

90 seconds. We chose the duration of 90 seconds for runs in order to produce meaningful CTT values (Petzoldt et al., 2014). Alternating between two tasks for this duration enabled us to focus on effects of the switching process based on a greater number of samples. As all experiments that rely on a Wizard-of-Oz approach, the performance of the wizard is one factor to keep in mind. We minimized the wizard's influence with the help of extensive practice by the wizard, combined with a quick and simple hotkey interface for speech commands.

Summary

This experiment investigated the costs of modality switches regarding effects on efficiency, primary task performance and cognitive workload.

- Switching between input modalities led to an overall increased efficiency.
- Modality switches did not lead to a loss of efficiency for individual tasks when they occurred during a task switch.
- Efficiency, primary task performance, visual distraction and cognitive workload mainly depended on the used modalities and not on the process of switching.
- Speech input was the best option regarding driver distraction and subjective demand for both tasks, despite reduced efficiency compared to switching between both modalities.

4.2. Effects of Situational Demands

This section is based on the following publication:

Roider, F., Rümelin, S., Pfleging, B., & Gross, T. (2017). The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 94–102). New York, USA: ACM

In-vehicle interaction concepts integrate multiple input modalities such as touch, speech, gestures, or gaze to control the in-vehicle information system while driving. Concepts typically aim to improve the interaction regarding naturalness, efficiency, robustness, and flexibility. The latter can be achieved by providing alternative input modalities. They allow drivers to choose freely the input modality they prefer to use for interaction with the vehicle. One particular benefit in the automotive domain is that they enable the driver to accomplish interactions using the modality that is most appropriate to the current driving situation (Müller et al., 2011).

The following section investigates the effects of situational demands on speech, gesture and gaze input while driving. A user experiment was conducted to assess the impact on objective and subjective measurements with the goal to assess the benefit of adapting the input mode to the situational environment.

4.2.1. Related Work

Varying Demands on the Driver

Cognitive science showed that people have separate pools of attentional resources that refer to the different sensory modalities (Wickens, 1980). The most important resource for driving a car is the visual resource since the majority of all information that drivers use is obtained visually. This is followed by the manual resource, which is needed for steering the vehicle. Finally, auditory and speech resources can be classified least important for driving (W. Wierwille, 1993). For a more detailed explanation of the cognitive foundations, please refer to Section 2.1.2.

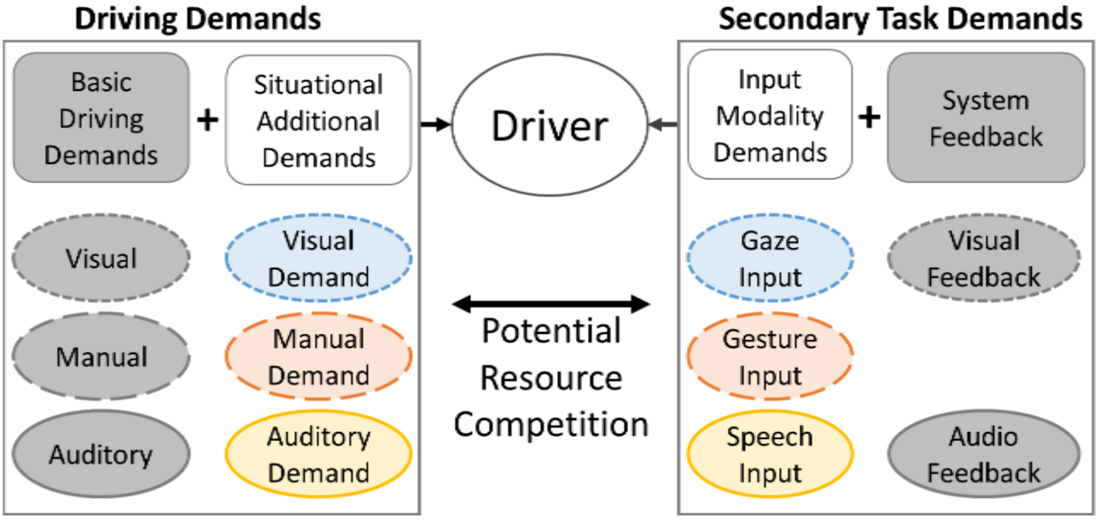


Figure 4.10: This figure illustrates demands of a driving task and a secondary task on the driver. Rounded elements represent demands on driver's visual, manual, and auditory resource.

Besides these basic demands of driving, on-the-road situations can impose additional demands on the driver's resources. Poor visibility, conversations with passengers, and curvy or bumpy roads are just some examples. They add up to the demands of basic driving and thereby change the proportion of demanded resources for the driving task. There are situations that put additional demands on the driver's visual, auditory, or manual resource. It is not common for a real-world situation to occupy only one resource, nevertheless there is usually one resource that is primarily in demand.

Accordingly, input modalities used for operating the secondary task also impact the proportion of demanded resources. Although the full range of potential demands of an interface needs to be considered, there is typically one sensory channel that is mainly addressed (Mehler et al., 2016). Gaze input mainly addresses the visual resource, gesture input the manual resource and speech input the auditory resource. We note that these channels match the demanded resources from the on-the-road situations described in the previous paragraph (see Figure 4.10).

Interference between Resources

There is a potential for interference between the demanded resources for the driving task and those for operating secondary tasks while driving. Strayer et al. identified three sources of driver distraction: visual, manual and cognitive interference (Strayer et al., 2011). In their experiment, cognitive interference is created by having a conversation over a cell phone, which is related to the auditory resource. According to their differentiation, interaction with different devices can lead to competition from one, two or all three sources.

The competition between resources is also addressed in multiple resource theory, which describes the influence of different resources for the prediction of dual task interference (Wickens, 2002). A more detailed explanation of the multiple resource theory is given in Section 2.1.2. In short, the theory implies that time-sharing two tasks with separate modalities (e.g., one visual task, one auditory task) has advantages over using the same modality for both tasks (e.g., both visual) due to resource competition. The amount of interference due to resource competition for a given resource is depending on the amount of competition for the resource and the relative importance of that resource in performing the task (Wickens et al., 1983).

Engström and colleagues presented results that show radically different effects on driving performance between visual and auditory secondary tasks (Engström, Johansson, & Östlund, 2005). The visual task led to a reduction of speed and large steering corrections, whereas the auditory task led to increased lane keeping quality. This variance of effects for visual and auditory tasks suggest that manual secondary tasks may also result in different impacts on the driver. Therefore, it is important to differentiate further modalities in order to cover specific influences. More concrete, the existing resource model could be expanded by another level in the modality dimension, which is related to tactile input (Boles et al., 2007; Wickens, 2008).

Several studies in the last years investigated the application of natural input modalities to operate in-vehicle information systems using either standardized tests, such as the Lane Change Task (LCT) (e.g. (Kern et al., 2010; Pfleging et al., 2012)), or miscellaneous driving simulations (e.g. (S. H. Lee, Yoon, & Shin, 2015; May et al., 2014)) to simulate driving demands. While these studies produced valuable results, they are related to one specific driving scenario. It is not clear how natural input modalities perform in driving situations with different demands. Being aware of the collective influence is essential to correctly assess results regarding secondary task performance because results might vary across studies if driving tasks impose different demands on participants. For example, gesture input (manual resource) is likely to perform worse in a driving simulation that requires frequent steering (also manual resource) compared to gesture input on a straight road while listening to the radio (auditory resource). Accordingly, a driving simulation that has many detailed and rich sounds of the vehicle and the environment might discriminate against speech input, due to resource competition.

In this experiment, we investigate the specific effects of using gaze, gesture, and speech input for operating a secondary task in different driving situations. In a driving simulator-based experiment, we measured the impacts on driving performance, secondary task performance and cognitive workload.

4.2.2. Experiment

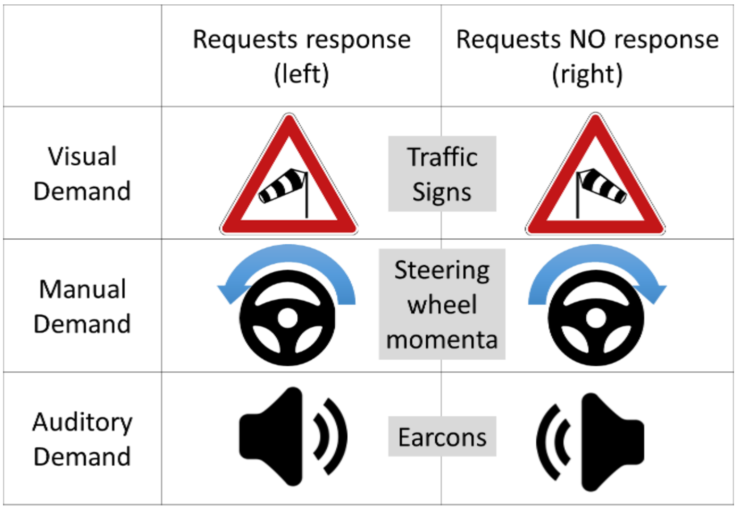
Design

We used a within-subject design for our experiment in the driving simulator. Participants experienced four driving scenarios in a balanced layout. Three of these scenarios contained additional events that increased the demand on either the visual, manual, or auditory resource of the driver, whereas one scenario represented the same driving scene without any additional demands. A more detailed explanation of how these demands were induced is given in the following subsection.

During each scenario participants performed three trials during which they performed a secondary task (selection of elements) via either gaze, gesture, or speech input in a balanced layout. During each trial of 90 seconds, they tried to make as many selections as possible. Consequently, the independent variables were the induced additional demand (visual, manual, auditory, none) and the used input modality (gaze, gesture, speech). We measure the effects of these two variables and their mutual influence with the help of the following dependent variables: driving performance (deviation of distance and lane keeping quality) and secondary task performance (secondary task completion time and subjective rankings). For subjective ratings, we used the driver activity load index (DALI) (Pauzié, 2008) and a 6-point Likert-type scale to the assess suitability of the input mode from 0 (not suitable for the interaction) to 5 (very suitable for the interaction).

Situational Demands

The different demands of traffic situations that occur on the road are the basis for the idea of applying additional demands in our study. However, we decided to use artificial events, which gave us good control over the study setup and allowed us to provide consistent demands for all participants. They also allow to separate the induced demands more clearly in comparison to more realistic events. Our goal was to design events that have a realistic connection to real-world driving scenarios on the one hand but address only one specific resource on the other hand.



	Requests response (left)		Requests NO response (right)
Visual Demand		Traffic Signs	
Manual Demand		Steering wheel momenta	
Auditory Demand		Earcons	

Figure 4.11: Visual demand was induced by traffic signs showing wind directions. Manual demand was created by applying a momentum on the steering wheel to either the left or the right side. Acoustic signals on the left side and the right side of the driver increased the auditory demand.

Figure 4.11 illustrates the events that were used to create the additional demands during the experiment: traffic signs at the roadside (visual demand), ear-cons in the cockpit (auditory demand) and temporary steering wheel momenta (manual demand). The latter was achieved by applying a force on the steering wheel that turned the wheel either to the left or to the right for one second. Participants would describe this event like a gust of wind that hits the car from the side. The distances between events were randomly distributed between 160 and 200 meters, which corresponds approximately to one event every six to seven seconds during the drive with 100 kilometers per hour. To make sure that participants really perceived and cognitively processed the meaning of the events, every event had two directions, left and right. Participants had to press a marked button on the steering wheel when a) the traffic sign shows to the left b) the sounds comes from the left c) the steering wheel turns to the left (see Figure 4.11). This way, events stayed consistent regarding their demand of cognitive processing and response and only differed in the perceptual modality.

Secondary Task

The secondary task was a simple selection task. There were three icons displayed on the central information display (CID) that represented typical categories of infotainment systems: contacts, navigation, and settings. An icon on the head-up display (HUD) instructed which category to select (see Figure 4.12).

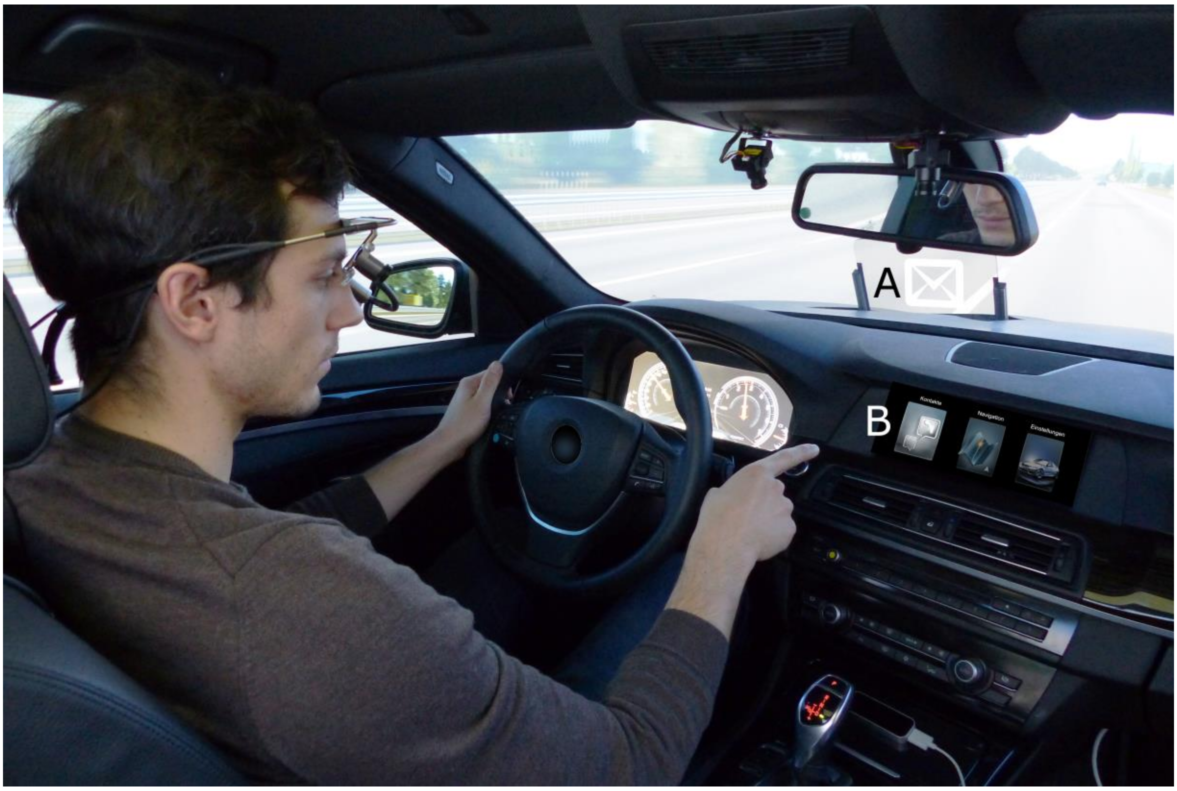


Figure 4.12: The setup used a head-up display (A) to instruct requested elements, which participants selected on the central information display (B) in the center stack.

Selections could be made via speech, gestures, and gaze. Selection via speech required participants to name the categories. Category names were displayed above the according elements on the CID. Selections via gestures were made by pointing in the direction of the category to highlight it and then make a tap in the air with the index finger to select it. Gaze input was implemented using a dwell time approach. Participants had to focus the icon of the requested category with their gaze. The focused icon was selected when participant's gaze continually stayed on the icon for 600 milliseconds. This duration was visualized by filling the icon with a lighter color over the duration of the dwell time. For all three modalities, the system provided visual and acoustic feedback when categories where hovered or selected.

The dwell time approach represents the standard approach in purely gaze controlled systems (Huckauf & Urbina, 2011). The combinations of gaze input with haptic buttons might produce better results in the automotive context (Kern et al., 2010). However, we deliberately decided against such an approach in order to keep input modalities clearly separated.

Participants

Originally, 36 participants took part in the study, but we had to exclude data of six participants. For those participants the eye-tracking system did not work properly, which resulted in non-representative secondary task times. Furthermore, one participant had to abort the experiment, because of simulator sickness. This resulted in a total number of 29 participants (21 male, 8 female) with a mean age of 25.3 years (*SD = 4.9*). All of them possessed a valid driver's license. The majority of participants was driving regularly, 18 of them reported to drive more than 20,000 kilometers per year. All but two had experiences with speech input either in cars or on their mobile phones. 24 participants reported that they had never used any form of gaze input before.

Apparatus

The experiment was conducted in a static driving simulator. The driving scene was projected on a 180-degree canvas in front of a vehicle mock-up. There were three displays in the cockpit (see Figure 4.12): the instrument cluster with a speedometer and the CID and HUD to display the secondary task. A *Leap Motion* gesture controller was placed below the center stack facing upwards to enable gesture recognition in the area in front of the CID. Speech recognition was achieved using a clip-on microphone that was attached to participants' clothing to provide good audio quality. Gaze recognition was accomplished using *Dikablis Live Essentials Eye Tracking Glasses*, a head mounted eye-tracker. The system includes one forward-facing camera and a second camera filming the participant's left pupil enabling an exact determination of users' gazes in the cockpit. The software for the secondary task was implemented in *Unity3D*. The examiner controlled the trials from an adjacent examiner room. Data capture was managed automatically by the software. Additionally, the software offered the examiner to act as a *Wizard-of-Oz* for speech input, in case that speech recognition did not work reliably for some participants.

Procedure

The examiner welcomed the participants in the simulator room and gave them a brief introduction about the course of the study. Participants signed a consent form and completed a questionnaire for demographic data. After that, they adjusted the seat position to their needs and put on the eye-tracking glasses. The examiner explained the secondary task and the interaction with gaze, gesture, and speech input. After the calibration of the eye tracking system, participants had a few minutes to practice the selection via gaze, gestures, and speech (without driving). They had another five minutes to familiarize with the driving simulator before they entered the motorway and followed a leading vehicle with 100 kilometers per hour at a distance of 50 meters. Participants were instructed to keep in the middle of the right lane and were not allowed to overtake the leading vehicle. After a few minutes of driving the examiner explained the additional demands and demonstrated them to the participants. It was made sure that participants experienced all additional demands (traffic signs, acoustic signals, and steering wheel momenta) and were able to distinguish between both directions (see Figure 4.11) before continuing with the trials. At the beginning of each trial, participants were informed about the upcoming type of additional demand and instructed to prioritize the primary task of driving when simultaneously executing the secondary task. The secondary task started with the instructions on the HUD. After the selection of a category, a new random icon was displayed. Participants continued selecting categories for the duration of the trial. Between trials they stopped the vehicle and completed the DALI questionnaire before continuing with the next trail. This procedure was repeated for all conditions.

4.2.3. Results

The main goal of our experiment was to investigate the influence of additional demands on different input modalities. However, absolute measures only represent our specific implementation for speech, gaze, and gesture interaction. Our dwell time approach for gaze input added a constant amount of time to every gaze selection. Times for speech input include the processing time of the system for automated speech recognition. Gesture input, in contrast, is more immediate and does not include any constant duration. Therefore, a more suitable approach to compare effects based in standardized values. Accordingly, Standardization provides a means to compare values from different normal distributions, by focusing on relative effects. We applied z-transformations on data from each input modality in order to create+ comparability across modalities. All statistical calculations in the following section were performed based on z-scores. At this point, we want to emphasize that, for absolute measures, there occurred significant differences between modalities and that the z-scores are only a means to provide a theoretical comparability here. The figures in this section display both, absolute values and standardized values, in order to support the awareness of actual differences.

Driving Performance

Figure 4.13 shows the standardized mean deviations of the requested distance of 50 meters to the leading vehicle. Highest mean deviations occurred during the visual demand conditions across all modalities. This indicates that it was most challenging for participants to keep the requested distance while they had to keep an eye on traffic signs on the same time. A two-way repeated measures ANOVA (additional demand: visual, manual, auditory, none; input modality: gaze, gesture, speech) was performed on these data. There was a significant main effect of the additional demand ($F(3,84) = 6.36, p < .01, \eta_p^2 = .185$). Bonferroni corrected post-hoc tests show that visual demand caused significantly higher deviation than all other conditions (all $p < .05$). There was no significant interaction between additional demand and input modality.

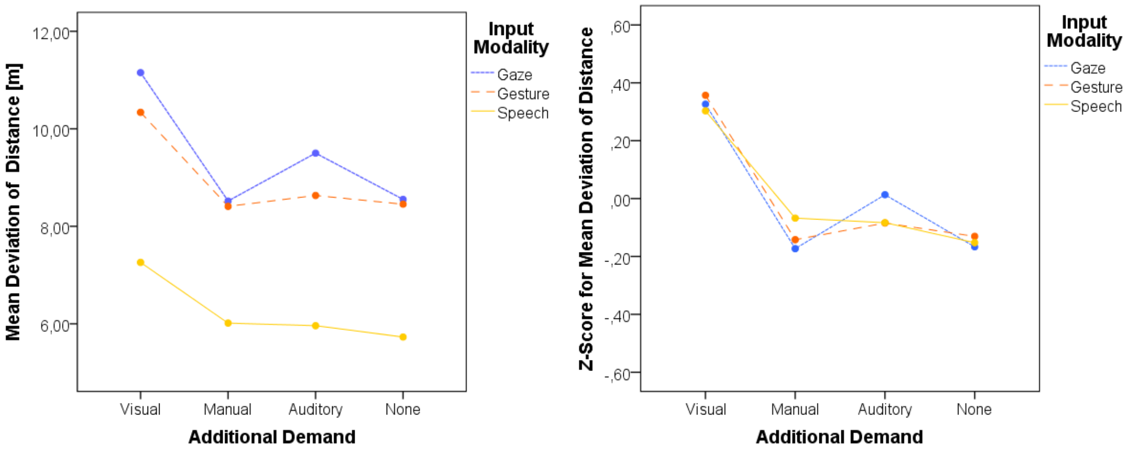


Figure 4.13: Additional visual demand caused the highest deviation of the requested distance of 50 meters over all input modalities.

We did not find significant effects on lane keeping quality. A two-way repeated measures ANOVA did not reveal any significant effects neither for the type of additional demand or the input modality and there was no interaction between both factors.

Glance Behavior

We assess the influence on the driver's visual distraction by analyzing the total glance time (TGT) on the display during the execution of a task. The results in Figure 4.14 show that the input modality has the greatest effect on TGT. Speech input hardly required any glances on the screen and gesture input also resulted in a relatively short TGT. Gaze input was by far the most visually distracting input modality with an average TGT of 1.539 seconds. Even when subtracting the chosen dwell time for gaze selections of 600ms, the gaze interaction required the most visual attention by far. Accordingly, there was a strong main effect of the input modality ($F(2{,}56) = 151.564, p < .001, \eta_p^2 = .844$), but also a weaker one for the demand ($F(3{,}84) = 5.194, p < .01, \eta_p^2 = .156$). Moreover, there was very weak interaction effect between both factors ($F(6{,}168) = 2.717, p < .05, \eta_p^2 = .088$). Significance levels have been corrected according to Greenhouse-Geisser.

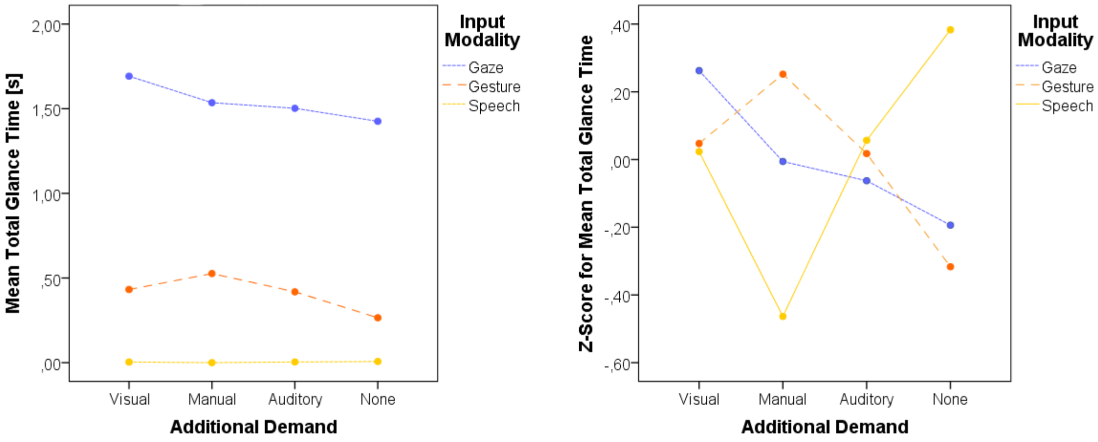


Figure 4.14: Total glance times were lowest for speech input, which hardly caused any glances on the screen. Gesture and gaze input resulted in longer glance times away from the

Task Completion Time

Figure 4.15 illustrates the z-scores of secondary task completion times. We only used data of correct selections for the calculations. The most noticeable impairment due to resource competition can be observed for gaze input. The impact on gestures was less pronounced and not clearly visible for speech.

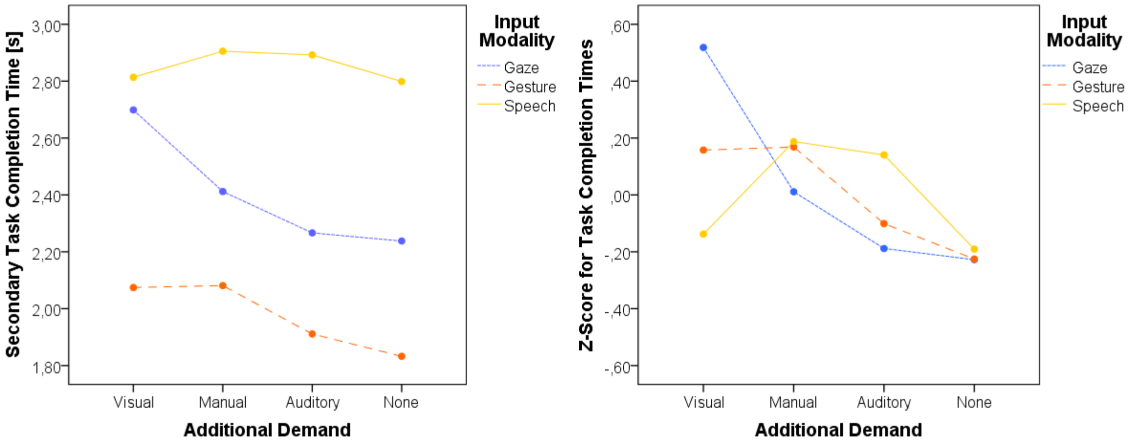


Figure 4.15: The mean task completion times show that the greatest impairments of competing resources could be observed for gaze input.

A two-way repeated measures ANOVA reveals that the additional demand had a significant effect on task completion time ($F(3{,}84) = 6.51, p = .02, \eta_p^2 = .131$). We applied correction of the significance level based on Greenhouse-Geisser, because the assumption of sphericity was not met for this test. Bonferroni corrected post-hoc tests show that selections during trials without additional demands were completed faster than during additional visual or manual demand (both $p < .05$). Additional auditory demand, in contrast, did not lead to a significant increase of task completion times. Moreover, there was a significant interaction between additional demand and input modality ($F(6{,}168) = 3.20, p = .02, \eta_p^2 = .103$). This shows that there is a specific influence of additional demands on task completion times depending on the input modality.

Error rates

Error rates were similar for gaze (*M = 6.1%, SD = 0.01*), gesture (*M = 6.4%, SD = 0.01*) and speech input (*M = 3.2%, SD = 0.01*). A two-way repeated measures ANOVA of z-scores did not reveal a significant effect for the type of additional demand nor was there an interaction between both factors.

Cognitive Workload

The dimensions of the DALI refer to cognitive, temporal and perceptive components (Pauzié, 2008). Results regarding the perceptive components would not add value at this point (e.g., there was higher auditory demand during additional auditory demand). More interesting are those dimensions, which refer to cognitive workload: effort of attention, interference and stress (Pauzié, 2008). We took the average ratings of those three dimensions to assess the cognitive workload during each trial.

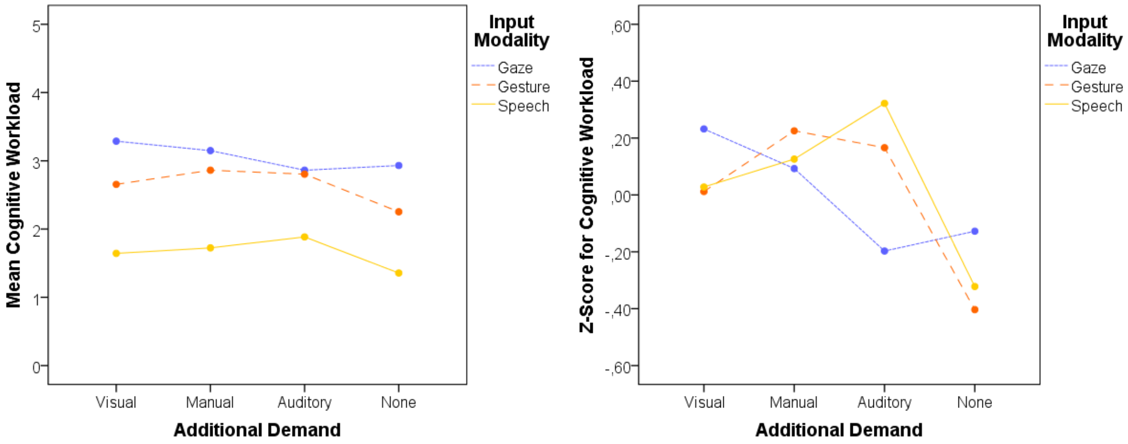


Figure 4.16: The ratings for cognitive workload are calculated from the average ratings for effort of attention, interference, and stress from the DALI questionnaire. Within each modality, competition of resources led to the greatest increase of the perceived cognitive workload.

Figure 4.16 shows the impact of different types of demand on cognitive workload. Participants perceived relatively highest cognitive workload when input modalities addressed the same modality as the additional demand. A two-way repeated measures ANOVA reveals a significant effect of the type of additional demand on cognitive workload, $F(3{,}84) = 9.90, p < .01, \eta_p^2 = .261$. Bonferroni corrected post-hoc test show that the ratings for each additional demand are significantly worse than without additional demand (all $p < .01$). There was a significant interaction between both factors (additional demand and input modality), $F(6{,}168) = 2.77, p < .01, \eta_p^2 = .119$. The dimension regarding temporal demand showed the same interaction pattern as cognitive workload, yet the effect was not significant.

Suitability Rating

The effects of additional demands on perceived suitability of secondary task interaction are illustrated in Figure 4.17. We observe the same interconnection as proposed in rankings of cognitive workload. Participants rated the suitability of the input modality lowest during trials in which the additional demand referred to the same resource. The impact on speech input best reflects this observation. During trials with additional auditory demand, speech input was rated significantly worse than during other trials (all $p < .05$).

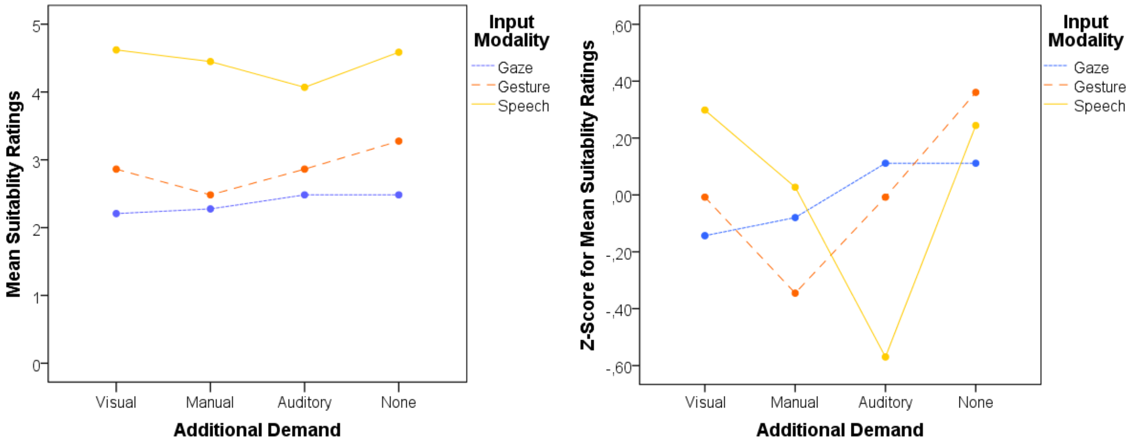


Figure 4.17: The mean rankings of suitability illustrate how resource competition led to the greatest impairments of perceived suitability for all modalities. Note that higher values are positive here since they indicate a higher suitability.

A two-way repeated measures ANOVA showed a significant effect of additional demand, $F(3{,}84) = 6.70, p < .01, \eta_p^2 = .193$. Bonferroni corrected post-hoc tests show that there was a significant difference between trials with no additional demand and those with manual and auditory demand (both $p < .01$). There was a significant interaction between additional demand and input modality, ($F(6{,}168) = 4.87, p < .01, \eta_p^2 = .148$). This supports the observation that participants rated each input modality least suitable when it matched the applied additional demand (gaze input during visual demand, gesture input during manual demand, speech input during auditory demand).

4.2.4. Discussion

For driving performance, we did not find an interaction effect between the additional demand and input modality neither on the deviation of distance nor on lane keeping quality. This can be explained by the fact that participants were instructed to prioritize the primary task. In a potential case of resource competition between driving task and secondary task, participants neglected the later one in order to fulfill the driving task. Therefore, their driving performance was not influenced by specific combinations of input modality and additional demand. Instead, we observed a significant impairment on distance keeping for all input modalities during additional visual demand. Since the basic task of driving is mainly visually demanding (over all trials), there is already a resource competition between visual additional demand and the basic driving task (W. Wierwille, 1993). Other studies observed similar behavior and concluded that additional visual distraction often leads to reduction of speed in order to reduce the primary task demand (e.g. (Antin, Dingus, Hulse, & Wierwille, 1990)). This explains greater deviations of requested distance to the leading vehicle during trials with additional visual demand.

In contrast to driving performance, participants' secondary task performance suffered from the effects of resource competition. The results section described interaction effects between additional demands and input modalities on secondary task performance and subjective rankings. This illustrates that speech, gaze and gestures input while driving is individually influenced by situational demands. This applies to task completion times, as well as to perceived cognitive workload and suitability. For each input modality, task completion times were longest, cognitive workload was greatest and suitability was lowest when the input modality and the additional task addressed the same resource. The estimated effect sizes f of interaction on task completion times ($f = 0.34$), suitability ($f = 0.42$) and cognitive workload ($f = 0.37$) indicated medium to strong effects of resource competition according to Cohen (Cohen, 1988). However, literature also suggests that demands of higher intensity (e.g. higher frequency of events) will result in an even greater disruptive effect of resource competition (Wickens et al., 1983). Accordingly, the potential advantage of using non-competing input modalities for interaction is likely to be even greater when the magnitude of demands on the driver increases.

Moreover, we found that the magnitude of impairments on objective and subjective measures differed between gaze, gesture and speech input. Therefore, it is crucial to include objective and subjective measures for investigation of natural interaction modalities in the car. For example, regarding task completion times only, there is no significant effect of additional demands on speech input (see Table 4.2). Concluding that speech input is not affected by additional demands of specific situations would be wrong, since there is a significant effect on subjective ratings that reflects the specific influence of additional auditory demand on speech input. Figure 4.16 shows that absolute ratings for cognitive workload of speech input was on a relatively low level compared to gesture and gaze. We assume that participants could maintain task completion times, even during resource competition, at the cost of increased cognitive workload and resulting impairments on perceived suitability. For gaze input, cognitive workload was already at a relatively high level without additional demand. Therefore, participants could not make this compensation at the cost of higher cognitive workload, which resulted in an impairment on task completion times. For gesture input, cognitive workload was rated slightly below gaze input. A part of the impact of resource competition could be compensated by increased cognitive workload, while the rest resulted in increased task times.

	TCT	**Suitability**	**Cog. Workload**
Gaze	0.425	Ns.	0.375
Gesture	0.344	0.383	0.631
Speech	Ns.	0.631	0.525

Table 4.2*: Estimated effect sizes f for the influence of additional demands on task completion time, perceived suitability, and cognitive workload for each modality. Dark gray indicates a statistically strong effect and light gray a medium effect.*

Limitations

All figures showing absolute values depict that the input modality had a significant effect on our measures. However, these measures strongly depend on our specific implementation of gaze-, speech- and gesture input. Other implementations (e.g., gaze input with a different dwell time, faster speech recognition systems) might lead to differing measures (e.g., generally shorter completion times), or even switched rankings (e.g., speech is faster than gestures).

While this is a point to keep in mind, we did not primarily compare the absolute performances of gaze, gesture, and speech input for the selection task. Instead, we illustrated the effects on task completion time in relation to the mean values for each input modality by using z-scores. This enhances the generalizability of results and achieves a comparability between the effects on individual input modalities.

Speech recognition did not work perfectly for some participants who did not speak loud enough to cover the sounds in the driving simulation. For those participants we included an additional

Wizard-of-Oz control that allowed the examiner to select the element himself, according to what participants said. The examiner imitated the typical response time of the system, to keep data for speech recognition consistent.

Summary

This experiment addressed the potential of multimodal interaction to use those input modalities that are least required by different situational demands.

- Speech, pointing gestures, and gaze input were individually influenced by situational demands.
- Speech input was rated as most suitable input modality for the secondary task during all situational demands although it was least efficient.
- Additional situational demands were to be neglected, in regard of the demands of the driving task. The benefit of adapting the input mode in a flexible way did not increase the performance of the primary task.

4.3. Summary

This chapter presented two user experiments that investigate the benefits of increased flexibility of multimodal in-vehicle input by providing alternative input modalities.

The first experiment in Section 4.1 investigated costs that emerge from switching between input modalities for secondary interaction. We observed that drivers can switch between touch and speech input without loss of efficiency and show how multiple input modalities can lead to an overall increase of interaction efficiency. Moreover, we observed that touch is better suited for spatial tasks and speech for verbal descriptive tasks in terms of efficiency, but that speech input is the better choice for both task types regarding distraction and cognitive load. We conclude that efficiency, primary task performance, visual distraction and cognitive workload mainly depended on the used modalities and not on the process of switching.

The second experiment demonstrated the individual effects of situational demands on natural input modes in Section 4.2. Each tested input modality - gaze, gestures, and speech input - performed worst during those situational demands that addressed the associated resources. In addition to those situational demands, the basic demands for driving a car, which is primarily a visually demanding task, must be respected. Situational influences add up to these basic demands, but they are comparably small and often do not change the proportion of demanded resources while driving. Consequently, throughout all demands speech input was generally favored, followed by gestures and gaze. Beyond that, we showed that gaze, gesture, and speech input differ in *how* they are influenced by additional demands in the car. The effect of additional demands on gaze input mainly led to an increase of task completion times, whereas effects on speech input were reflected only in subjective ratings. Effects on gesture input were split up between both factors. This knowledge is especially relevant for interaction designers that integrate gaze, gestures, and speech in automotive concepts. They need to understand how situational demands influence the efficiency and perceived demands of these input modalities.

In summary, both experiments show the great potential of speech input while driving. It was generally favored for a safe and convenient interaction in the car despite a low efficiency for some tasks. At the same time, the experiments also support the benefits of integrating alternative input modalities, such as gestures. Switches between speech and manual interaction can be performed without additional costs and thereby increase efficiency or avoid resource competition in certain situations. This enables a feasible interaction for a wider range of situations occurring on the road.

5. Supporting the Flexible Use of Alternative Input Modes

The last chapter pointed out strengths and weaknesses of the individual modalities, but also demonstrated benefits of using different input modes. Speech input has many advantages while driving regarding distraction and cognitive workload, but it is often inefficient and not transparent to users. Gesture input could be a valuable addition to speech input by providing simple interactions for quick selections and allowing to reference spatial information, but people often struggle with the correct execution of gestures due to a lack of adequate feedback. In this chapter we present two user experiments that explore solutions how to support speech and gesture input for in-vehicle interaction by tackling the described issues.

- *Leveraging Speech Input* shows how the use of speech input can be supported with the help of visual cues.
- *Visualizing Gesture Information* explores ambient light feedback to provide adequate feedback to support the use of gesture input.

5.1. Leveraging Speech Input

This section is based on the following publication:

Roider, F., Rümelin, S., & Gross, T. (2018). Using Visual Cues to Leverage the Use of Speech Input in the Vehicle. In *International Conference on Persuasive Technology* (pp. 120–131). Springer, Cham. https://doi.org/10.1007/978-3-319-78978-1_10

Touch input has become the state-of-the-art input modality for interaction with many devices over the last decade. More recently, speech input is about to emerge as a full-fledged alternative to touch input, supported by the success of voice-based systems such as Amazon Alexa, Apple's Siri, or the Google assistant. Besides mobile devices or smart home applications, touch and speech have evolved as the dominating input modalities in the automotive domain. The latest models of many manufacturers integrate large touch-based screens and intelligent speech-based systems, but there is no or only little interplay between both modalities at the moment. Touch is usually the primary input mode, while speech input is mostly a less used alternative path for specific use cases that work independently of the touch interaction. At the moment, there is no or only little interplay between both modalities. Touch is usually the primary input mode. Large touch-based screens are dominant in the driver’s field of view and increase the likelihood that he will use touch input. However, speech is only rarely used, especially when drivers start the interaction by touching. When they do not decide to use in the first place, they do not switch modalities easily.

There are several advantages of speech compared to touch input that support the driver's safety. Speech input reduces visual distraction, it allows drivers to keep both hands on the steering wheel, and it offers a fast and convenient way to achieve many tasks in the vehicle, especially those that require the driver to enter text in any forms (e.g., when giving destinations, searching for contacts, or composing text messages). Efficiency of the interaction is also a safety concern in the driving context. The shorter the driver is concerned with other tasks than driving, the better.

However, for some tasks, especially those that require the user to express spatial information, touch input is suited much better (Wickens et al., 1983). Furthermore, Strayer et al. have shown that speech input is not free of distraction either (Strayer, Drews, & Crouch, 2006) and Section 4.2 has shown that situational influences can impair the suitability of speech input. Finally, speech input still faces some technical challenges, such as understanding heavy dialect or recognition in noisy conditions. In order to cope with such problems, it is beneficial to integrate both input modalities in the car. The challenge is to find a seamless and efficient interplay between alternative input modalities, so that users can actually benefit from the many possibilities they have. Users should be made aware of alternative interaction options without being overloaded by instructions that distract from the actual task (Reeves et al., 2004).

In this section, we address the question if visual cues provide an effective, but unobtrusive way to leverage speech input while driving.

5.1.1. Related Work

Fogg describes the likelihood of influencing peoples' behavior as product of three factors. Besides sufficient motivation and the ability to perform a target behavior, effective triggers are necessary. There are three types of triggers: *sparks* motivate behavior, *facilitators* make behavior easier, and *signals* simply remind people to perform a behavior (Fogg, 2009a).

In the case of speech input while driving, reduced distraction and increased safety provide a strong motivation. Furthermore, we assume that people have the ability and know-how to use speech input. In this case, visual cues are signals that just remind people to use speech input now. But they can also serve as facilitators that make the target behavior easier to do. By displaying possible voice commands, they help reducing the effort for formulating words ourselves, reduce the thinking effort and thus increases the likeliness that speech input is used.

Effects of the Prompt Modality

Why do people rather interact via touch instead of speaking to current cars in regard of these benefits? A psychological explanation is the cognitive mapping of visual stimuli to manual responses (Teichner & Krebs, 1974; Wickens et al., 1983). Large touch-sensitive screens in current vehicles provide visual stimuli that provoke direct touch input. The other way around, auditory stimuli are most compatibly mapped to speech responses (Teichner & Krebs, 1974; Wickens et al., 1983). Accordingly, one way to remind users to use speech input is to prompt them with auditory cues, such as spoken prompts or earcons.

Yet, visual cues have some major advantages over spoken or auditory cues: Visual cues are faster. Users can benefit from preattentive processes that support rapid pattern recognition and thereby absorb information at one glance (Parush, 2005). Furthermore, auditory prompts are short term and sequential by nature and thus make heavy demands on human working memory (Bradford & H., 1995). Visual cues, in contrast, do not have this temporal relation and can be displayed permanently. At the same time, they are less disruptive than acoustic prompts. Playing a sound or spoken prompt whenever the user should use speech interaction can be very annoying.

Parush compared spoken and visual prompts for speech dialog interaction in multitasking situations such as driving (Parush, 2005). They found that speech interaction with spoken prompts took longer than with visual prompts, whereas the driving performance was better with spoken prompts. Their study also showed that the difficulty of the tracking task affected these results. They conclude that multitask situations must not always have spoken prompts. Especially novice users can profit from visual cues for speech interaction (Yankelovich, 1996). In multimodal systems, this allows to display the names of possible selections to suggest or explicitly indicate what users can say.

Implicit and Explicit Prompts

Explicit prompts stand in contrast to implicit prompts that help to direct user input in a more reserved way. Yankelovic proposes that those are not two distinct categories but spoken prompts rather fall along a continuum from implicit to explicit (Yankelovich, 1996). The most explicit form of prompts are directive prompts. They tell user the exact words they should say. Descriptive icons such as microphones or speech bubbles are one potential way to notify users to begin speaking (Reeves et al., 2004). Kamm concludes that directive prompts can facilitate the "ease of use" of voice interfaces (Kamm, 1995).

Explicitly telling people what to do can potentially result in the exact opposite behavior. Prompts that are perceived as restricting to one’s freedom (to choose the input modality) can arouse reactance (Dillard & Shen, 2005). Reactance is an unpleasant motivational arousal that serves as a motivator to restore ones freedom e.g. by not following what the system suggests (Steindl, Jonas, Sittenthaler, Traut-Mattausch, & Greenberg, 2015). The extent to which a message is perceived as threatening to one's freedom finally influences peoples' behavior to follow or not follow the advice of the message (Steindl et al., 2015).

Summary

User studies in literature, as well as experiments in Sections 4.1 and 4.2 have shown that speech interaction potentially leads to safer and often more efficient interaction. Therefore, user interfaces should support drivers to use their voice whenever this can provide a benefit. We assume that this could be changed by providing a suitable trigger. Typically, auditory cues are used to prompt drivers to talk, as humans tend to use speech to respond to auditory stimuli. However, visual cues have some advantages over auditory cues that make them a promising means for triggering speech input. They can range from implicit hints to very explicit directive prompts. The latter ones are potentially more effective, yet they might draw too much of the driver's attention or arouse reactance so that user will eventually not follow the system's advice.

5.1.2. Experiment

We conducted a user experiment that investigated the efficacy of visual cues to leverage speech input while driving. In order to address this question in a differentiated way, we propose five hypotheses:

H1. Visual cues increase the amount speech interactions.
H2. Explicit visual cues result in higher speech rates than implicit ones.
H3. Additional audio signals result in higher speech rates than only visual cues.
H4. Explicit visual cues cause higher visual distraction than implicit ones.
H5. Explicit visual cues induce a higher threat to freedom than implicit ones.

Participants

45 participants (17 females, 28 males) with a mean age of 30.2 years ranging between 21 and 58 years took part in the study. All of them were either native German speakers or had excellent

knowledge of the German language and none of the participants had motor impairments of the upper limbs, which would have shifted their decisions towards either touch or speech input. Participants' self-reported data showed about the same openness to touch and speech input with a slight advantage for touch. Tendencies to use rather speech or touch input while driving was balanced over all participants.

Experimental Design

The experiment used a within-subject design. Each participant completed 64 tasks that were displayed on a secondary display while they were driving. For every task, participants had to decide whether to use speech or touch input. Tasks varied in presence and explicitness of visual cues (none, implicit cues, explicit cues, implicit and explicit cues). In order to create a greater generalizability of our results, we additionally included two task types (selection, text input) and two driving scenarios (easy, difficult) and varied the presence of an additional audio signal (none, audio). Each specific configuration occurred twice to each participant. All tasks were counterbalanced in order to prevent ordering effects.

In both driving scenarios, participants followed a leading vehicle on a highway with three lanes and slight curves. In the easy scenario, the leading vehicle moved with 100 km/h, it stayed on the rightmost lane and did not overtake. There was only few traffic. In the difficult scenario, there was more traffic. The leading vehicle moved at 130 km/h and it used all three lanes to overtake slower cars. The audio signal was the standard earcon of a 2017 BMW 7-series for pressing the push-to-talk button on the steering wheel. Task types and visual cues will be explained in detail in the following sections.

Experimental Tasks

We used two task types in our experiment. The *selection task* is well suited to be solved with touch input, while the *text input task* is better solved using speech. By including these very different task types we aim to achieve a better generalizability of our results for a broader range of tasks. The speech recognizer was active as soon as a task appeared.

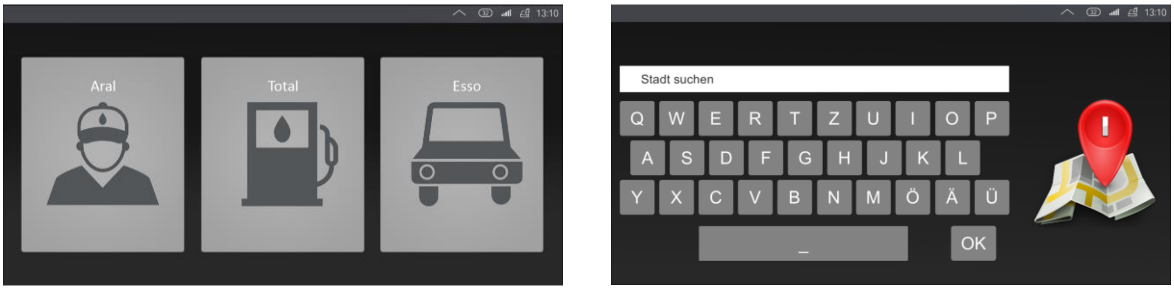


Figure 5.1: The selection task (left) displayed three big elements for gas stations (example in the figure) or restaurants. The text input task (right) displayed an input field and a virtual keyboard to search a destination city (example in the figure) or a contact name.

The goal of the *selection task* is to make a selection out of three elements. It is illustrated in Figure 5.1 on the left. The task displayed either three gas stations or three restaurants. Participants were instructed before the beginning of the trails, which elements to select for the gas stations ("Total") and the restaurants ("Seehaus"). Thereby we avoided that the modality of the instruction would influence the modality choice of the driver. Selections were made by saying the name of the instructed element or by touching the according tile whenever this screen would appear.

In the *text input task*, participants had to enter a short text in form of a contact name or a destination. It is illustrated in Figure 5.1 on the right. They were instructed to enter "Lisa" for contacts and "Jena" for the destination by either saying the requested entry or by typing it on the keyboard. Both instructed texts have four letters. We assume that current intelligent text input systems propose a small selection of possible words about three letters. They only require the user to tap a fourth time to select out the correct proposition.

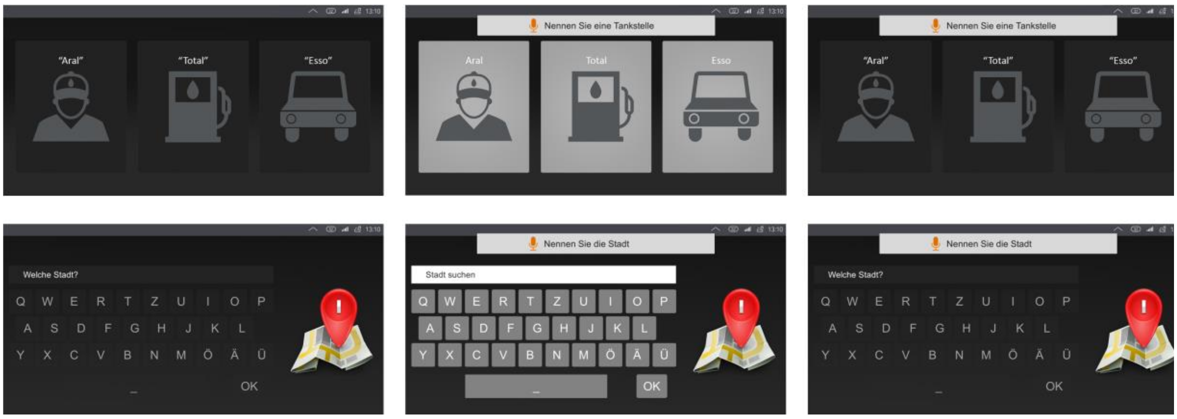


Figure 5.2: Both tasks with increasing levels of visual cues. From left to right: implicit, explicit, implicit and explicit.

Visual Cues

In a preceding brainstorming session, we identified interface elements in touch-based systems that users associate with the use of speech input. Identified elements were split in two groups: implicit cues and explicit cues.

Implicit cues are more subtle adaptations that refer to speech input without explicitly telling the user what to do. In the experiment, three types of adaptions were made when implicit cues were used. First, the highlighting of touch elements such as buttons was reduced. Touchable areas are often highlighted in brighter colors, which creates a visual stimulus that makes users more likely to touch them. Second, more emphasis was put in visible text on the screen by highlighting possible commands with quotation marks, making it easier for users to remember potential commands. Third, text was rephrased to be rather conversational and therefore promote a spoken answer. For example, instead of "Search city" the text input task displayed "Which city?".

Explicit cues, in contrast, are more noticeable and directly prompt the user to use speech input. Again, there were three adaptations made in conditions with explicit cues. First, a notification banner was displayed on the top of the screen to catch users' attention. Second, on the banner, there was a microphone symbol orange color. Third, there was a short text displayed, that prompted users to name the desired selection or text.

Figure 5.2 displays the application of implicit and explicit cues on the two experimental tasks. The most left picture illustrates the task with implicit cues (Imp). The next one shows the explicit cues (Exp). Finally, the rightmost picture integrates both, implicit and explicit cues

(ImpExp). Together with the basic version of each task (see Figure 5.1), both rows illustrate four rising levels within the continuum from implicit (left) to explicit adaptions (right).

Apparatus

The experiment was conducted in a static high-fidelity driving simulator illustrated in Figure 5.3. The driving scene was projected on a 180-degree canvas in front of the vehicle mock-up. There were two displays in the cockpit: the instrument cluster displayed a speedometer and rounds per minute, the central information display in the dashboard showed the experimental tasks.

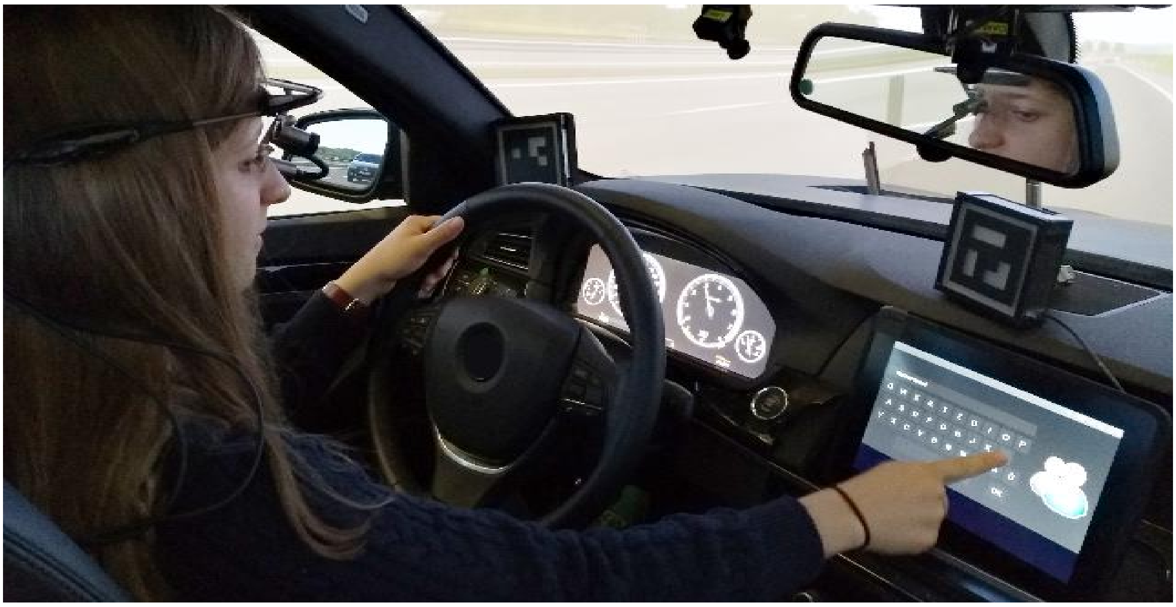


Figure 5.3: The cockpit in the experimental setup. The tasks were displayed on the central display and participant decided whether to use touch or speech input for the interaction. Glance behavior was recorded using head mounted eye tracking glasses.

The latter was a 10.1 inch *Faytech capacitive touch display*[3] with a resolution of 1280x800 pixels. The experimental tasks were integrated in a special application implemented in *Unity3D*. Speech recognition was achieved using the built-in speech engine in *Unity3D* which uses the Windows speech recognition engine in combination with a *Rode SmartLav+*[4] clip on microphone. The users' glance behavior was recorded with *Dikablis Essential*[5]eye tracking glasses in combination with infrared markers.

Procedure

Participants completed a short form covering demographic data before they were introduced to the experimental tasks. They were shown all tasks (selection-gas stations, selection-restaurant, text-contacts, and text-destination) in the basic version, without any visual cues and without an

[3] https://www.faytech.com/de/katalog/product/101-capacitive-touch-monitor-ft10wtmbcap/

[4] http://de.rode.com/microphones/smartlav

[5] http://www.ergoneers.com/eye-tracking

audio signal (as illustrated in Figure 5.1). They were instructed to memorize the correct selection for each of the four tasks. Participants were *not* told that there will be additional visual cues, but the examiner emphasized that participants *always* have the choice to use either touch or speech input. Tasks appeared automatically on the central display after a random wait time between 10 and 15 seconds. This varying wait time avoided that participants got used to a certain rhythm, and ensured that there was sufficient time between tasks, so that each task was handled independent of the previous one. As soon as the task was displayed participants looked at the screen to identify the task, decided whether to use touch or speech, and made their input. Tasks disappeared after the selection or text input was completed and participants turned back to driving until the next task appeared. After all tasks had been completed, participants were shown all possible combinations of task type, visual cue, and audio signal on a laptop display without driving. This way they could concentrate on the illustration of tasks. For each specific illustration, they rated the suitability of touch and speech input, as well as the threat to freedom. The order in which tasks appeared was counterbalanced.

Data

For each task, the application logged the users' choice of input modality along with the visual cue, the driving scenario, the task-type, and the presence of an audio signal. There was a total number of 2880 choices (45 participants * 4 visual cues * 2 scenarios * 2 task types * 2 audio signals * 2 choices per configuration) and 90 choices for one specific configuration. The Eye Tracking system recorded the total glance time per task, which is the average duration that a participant looked on the display while a task was active.

Finally, there were participants' self-reported assessments about the perceived threat to freedom that is caused by a specific illustration of a task. They are based on the ratings of four items, each on a 5-point Likert scale from -2 (strongly disagree) to 2 (strongly agree) (Dillard & Shen, 2005). It is based on the ratings of four items on a 5-point Likert-type scale: 1. the message threatened my freedom to choose. 2. The message tried to make a decision for me. 3. The message tried to manipulate me. 4. The message tried to pressure me.

5.1.3. Results

Choice of Input Modality

The choice of input modality was encoded in a binary variable (0 = touch input, 1 = speech input). Figure 5.4 illustrates the percentage of speech inputs depending on the visual cue, the driving scenario, the task, and on the occurrence of an audio signal. The percentage of speech input was lowest with no visual cues displayed and grew with increasing level of explicitness of the visual cues. The maximum increase was 16% for the selection task and 15% for the text task. This trend can be observed for both task types and both driving scenarios.

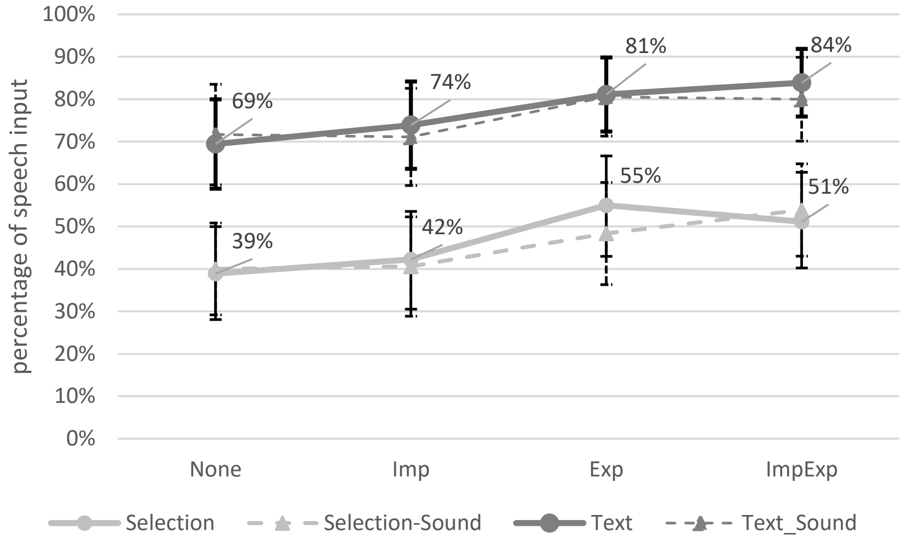


Figure 5.4: This figure illustrates the percentage of speech input used, depending on the visual cues, the task-type, and the audio signal.

The results of a Friedman test show that the visual cues had a significant influence on the choice of input modality ($\mathrm{X}^2 = 13{,}904, p = .003, r = 2.07$). Additional Wilcoxon signed-rank tests were used to compare implicit cues to those conditions with explicit cues. They show that only explicit cues (*Mdn=0.69*) did not result in significantly higher percentage of speech input than implicit cues (*Mdn=0.56*). Instead, implicit and explicit cues (*Mdn=0.69*) led to a significant rise of the speech rates compared to implicit cues ($Z = -2.48, p = .013, r = 0.37$). The level of significance was corrected according to Bonferroni.

Additionally, a logistic regression was performed to analyze the influence of all factors on the participants' modality choice. The results show that both the logistic regression model, $\mathrm{X}^2(4) = 350.00, p < .001$, as well as the individual coefficients (except the audio signal) were statistically significant. The model correctly classified 66.8% of the cases. Increasing the explicitness of visual cues by one level rises the relative probability to choose speech input by 17.4%. In the difficult driving scenario, the relative probability to choose speech is 54.2% higher than in the easy one. Finally, the task-type had the greatest impact. Choosing speech for text input was 290.0% more likely than for the selection task. R^2 (Nagelwerke R square) is 0.155, which indicates a strong effect (Cohen, 1992). In line with these findings, a Wilcoxon signed-rank test between conditions with acoustic signal (*Mdn=0.59*) and without acoustic signal (*Mdn=0.56*) showed no significant differences (*Z=-0.87, ns.*).

We can summarize that the task type was the most decisive coefficient, followed by the driving scenario. Visual cues play a smaller, yet decisive role in influencing the participants' decision. Moreover, the model shows that the probability to choose speech input rises with the level of visual cues. The additional audio signal did not influence the participants' decisions.

Glance Behavior

The average glance duration in both driving scenarios was between 0.83 and 0.89 seconds for the selection task and between 1.41 and 1.83 seconds for the text entry task. Figure 5.5 and Figure 5.6 illustrate the total glance times for both tasks individually. In the selection task, TGT remained constant and did not change with higher levels of speech input. In the text entry task, we observe a light tendency that glance times decrease with increasing level of visual cues for the text input task.

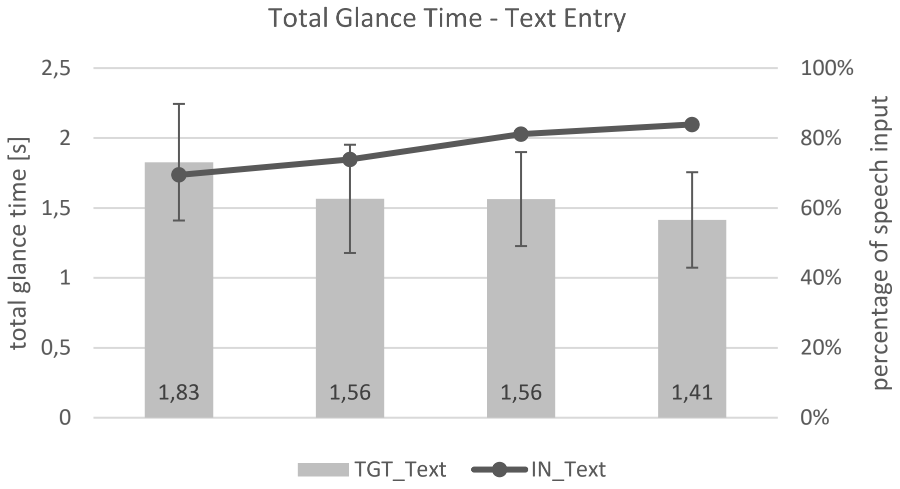


Figure 5.5: The total glance time for the text entry task was significantly reduced due to an increase of speech input with increasing levels of explicitness of visual cues.

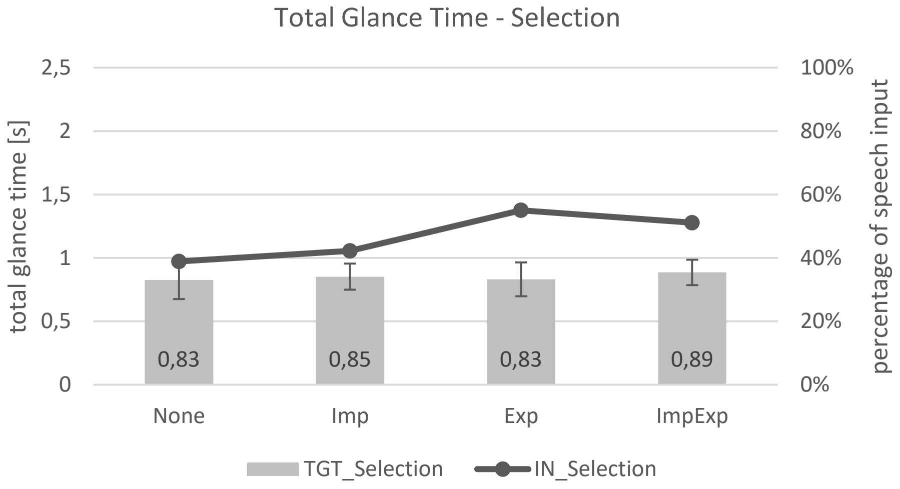


Figure 5.6: The total glance time for the selection task was not affected by the increase of speech input with increasing levels of explicitness of visual cues.

Glance data was not normally distributed. Results of a Friedman test showed that the average glance duration on the display was not significantly affected by the visual cues ($X^2 = 2.32, ns.$). Wilcoxon tests confirmed that the total glance duration without visual cues did not significantly differ between implicit cues (*Mdn = 1.03*) compared to Explicit cues (*Mdn = 1.13*) or implicit and explicit cues (*Mdn = 0.99*). The duration that participants looked on the display did not depend on the type or presence of visual cues.

Perceived Threat to Freedom

The perceived threat to freedom was measured with the mentioned four questions. Ratings were very diverse because some participants did not feel any restriction, while others reported that they felt like being influenced or urged to behave in certain way. The data was not normally distributed. A Friedman test indicted that the average ratings were not significantly affected by the visual cues ($\mathrm{X}^2 = 5.49, ns.$). Accordingly, results of Wilcoxon tests showed that neither explicit cues (*Mdn = 0.00*) nor implicit and explicit cues (*Mdn = 0.19*) were associated with a higher threat to freedom compared to implicit cues (*Mdn = 0.00*).

5.1.4. Discussion

The first hypothesis H1 proposes that visual cues increase the amount of speech input used. Our results show that the user's choice of input modality was mainly determined by the task type and the driving scenario. Still, the visual cues had a significant influence, which can be classified as a strong effect based on the estimated effect size of the Friedman test (Cohen, 1992). Furthermore, the results of the logistic regression model and Figure 5.4 indicate a trend that implicit cues can already increase speech usage and we can accept H1.

H2 claimed that explicit cues would be more effective to promote speech input than implicit ones. The logistic regression model supports this thesis, but additional pairwise comparisons showed that the increase of speech rates for explicit cues compared to implicit cues was not significant. Based on these findings we do not accept H2. However, *extending* implicit cues by explicit ones (ImpExp) led to a significant increase of speech interaction. This suggests that the effects of implicit and explicit cues complement one another. The combination of both led to overall highest speech input rates.

H3 proposed that additional audio signals increase the amount of speech input used. However, the results show that they did not have a statistically significant influence on the participants' decisions. This was surprising, given the fact that speech input is mostly prompted using audio signals. At the same time, it is in line with the disadvantages of the temporal and short term nature of audio (Bradford & H., 1995). The audio signals were played the moment the task appeared on the screen, but participants often needed a couple of seconds to control the vehicle before attending to the task. The trigger existed, but it was not well-timed, which is one possible reason why achieving the target behavior fails (Fogg, 2009b).

H4 proposed that explicit cues cause increased visual distraction compared to implicit ones. The results did not reveal significant differences between the four levels of visual cues. A deeper look into glance data shows different trends for glance behavior depending on the task. For the text input task, the usage of speech input rises with increasing level of visual cues while the average glance times for both, speech and touch decreases (Figure 5.5). This means that not the actual glance times per task changed, but rather the amount of (less visually distracting) speech inputs rose, which led to an overall decrease of the total glance time. The fact that this affects only the text input task shows that glance times for touch and speech input for the selection task were similar, since the higher percentage of speech selections did not reduce the overall glance duration. In summary, explicit cues did not result in longer or more glances on the display, but they reduced the overall visual distraction by increasing the amount of speech usage. For these reasons, we do not accept H4.

H5 assumed that explicit visual cues induce a higher threat to freedom than implicit ones, which increases the likeliness to show reactance and that participants will not follow the systems' advice. The average ratings from participants' self-assessed threat to freedom did not differ significantly between conditions. This indicates that the design of our visual cues did not have a big influence on the perceived freedom to choose themselves. A limiting factor might be that participants were explicitly told that they can always decide freely. Moreover, previous work in this field notes under-reporting as potential problem for participants' self-reported data. This might also contribute the missing variance in this case (Miranda et al., 2013). Therefore, we do not accept H5.

Summary

In this experiment, we explored the influence of visual cues on the users' choice whether to use speech or touch input. We conclude that visual cues are an effective means to influence the user's choice of input modality and thereby to support users by emphasizing suited input modalities. The system can guide users in an unobtrusive way so that they can benefit from the whole range of input modalities, without concerning themselves with the decision. Our study showed that visual cues increased the amount of speech input used, decreased visual distraction on the road, and thereby contributed to the driver's safety.

- Visual cues can significantly contribute to leverage speech input while driving across different task types and different driving scenarios
- Explicit cues were more effective than implicit adaptions of the user interface
- Visual cues did not cause increased visual distraction
- Overall glance time away from the street can be reduced for text input tasks by using explicit visual cues
- Neither implicit nor explicit visual cues made participants feel restricted in their freedom to freely choose the modality they preferred

5.2. Visualizing Gesture Interaction

This section is based on the following publication:

Roider, F., & Raab, K. (2018, June). Implementation and Evaluation of Peripheral Light Feedback for Mid-Air Gesture Interaction in the Car. In *2018 14th International Conference on Intelligent Environments (IE)* (pp. 87-90). IEEE.

Interaction with non-driving related functions while driving poses some special demands to the mode of interaction. For the driver's safety, it is important that the interaction is least distracting. In this regard, mid-air gestures are a promising form for in-vehicle interaction, because they can reduce the driver's visual attention (Riener, 2012). Existing systems in the 2017 VW Golf or the 2017 BMW 5 Series provide acoustic feedback to signal a successful execution of a gesture. However, there is usually no feedback during the execution of the gesture. If users execute a gesture not correctly and therefore fail to trigger any feedback, there is no information available to understand why. As a result they tend to look for any impact, which is likely to result in visual and mental distraction (Riener et al., 2013). Therefore, many users struggle to use gesture input effectively due to a lack of experience and tactile feedback. Adequate feedback should support gestures continuously during execution and not only in case of a successful gesture. It should be capable of reflecting the spatial freedom of gesture interaction. Finally, it must be least distracting from the driving task.

5.2.1. Related Work

Compared to a turn-push-knob, or a touch display, gestures are not locally constrained to a certain input device or interaction area. It is possible to perform them in a larger three dimensional interaction area in the vehicle cockpit (Riener et al., 2013). The variety of mid-air gestures in the vehicle ranges from simple confirmation or swiping gestures (e.g.(Gable, Raja, Samuels, & Walker, 2015; Ohn-Bar & Trivedi, 2014)) to more complex pointing gestures. Latter allow to expand the interaction possibilities to objects in the users environment (Rümelin, Marouane, & Butz, 2013). Due to this freedom and variety, supportive feedback is even more important to exploit the full potential of gesture interaction.

Past research mainly used acoustic and visual feedback for gesture interaction (e.g. (Cairnie et al., 2000; May et al., 2014; Roider, Rümelin, Pfleging, & Gross, 2017)). Visual feedback on a display does not fulfill the requirements of gesture feedback, because it is restricted to the display area and visually distracting (Shakeri, Williamson, & Brewster, 2017). In contrast, acoustic feedback has been shown to efficiently support gesture interaction while driving (May et al., 2014; Shakeri et al., 2017). It is hardly visually distracting, but its capability of reflecting spatial information in the environment is limited and it can be annoying for continuous use. Tactile mid-air feedback (e.g. via ultrasound) has shown great potential to improve in vehicle gesture interaction (Harrington, Large, Burnett, & Georgiou, 2018; Shakeri, Williamson, & Brewster, 2018) . However, the spatial coverage as well as the information bandwidth are still limited (Rümelin, Gabler, & Bellenbaum, 2017). In this regard, ambient interior lighting is an interesting option to compensate the downsides of purely acoustic feedback. It is already prevalent in many modern cars and has the potential to represent spatial information, which is inherent to gesture interaction. Moreover, it has been show that peripheral light feedback is less visually distracting than visual feedback on a central display and that there is no negative effect on lane deviation compared to acoustic feedback (Shakeri et al., 2017).

In summary, acoustic and peripheral light feedback are potentially suited for gesture feedback while driving. Peripheral light feedback could further compensate some limitations of acoustic feedback. This section focuses on the benefit of extending acoustic feedback by peripheral light feedback regarding the support during the execution of different mid-air gestures in the car.

5.2.2. Experiment

We conducted a user study to investigate peripheral light feedback for typical in-car gestures regarding efficacy, task times and subjective ratings.

Participants

21 participants (15 male, 6 female) between 19 and 55 years (*M=33.0, SD=10.5*) were recruited for the study. The majority of them were regular drivers, who reported to drive more than 20.000 kilometers per year.

Study Design

A within-subject design was used. The independent variables were the feedback type (acoustic feedback only (AF), acoustic feedback with additional peripheral light feedback (PLF)) and the gesture type (confirmation, swipe, pointing). The dependent variables were the efficacy (percentage of correct executions) and the task completion time (seconds) per gesture. Additionally, participants rated the overall impression of the gesture interaction, the support of the feedback (each 6-point Likert scale), and the subjective workload (Driver Activity Load Index) for the two feedback types.

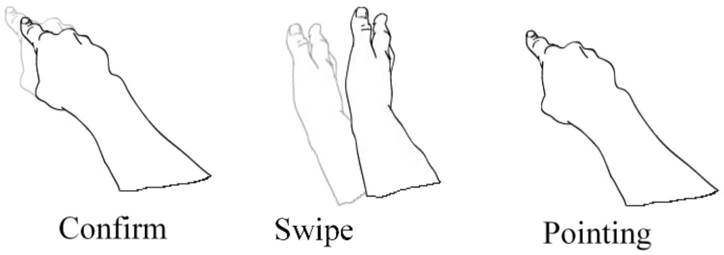


Figure 5.7: Three different gestures with different complexity were supported: confirm, swipe and pointing.

Gestures

There were three gesture types with different complexity: confirm, swipe and pointing gestures (see Figure 5.7). Confirm and pointing gestures were implemented by ourselves. For the swipe gesture we used the predefined gesture in the Leap Motion SDK. For all three gestures we implemented feedback variants on the LED strip (see Figure 5.8):

Confirm: The confirm gesture required the user to move the hand forward with the index finger outstretched. As soon as the finger passed a certain virtual plane the confirmation is triggered with acoustic feedback. PLF already started 10 cm before the plane was passed. A small light bar gets more opaque the closer the finger gets to the virtual plane, which gives the user a sense of how far he is from triggering the gesture.

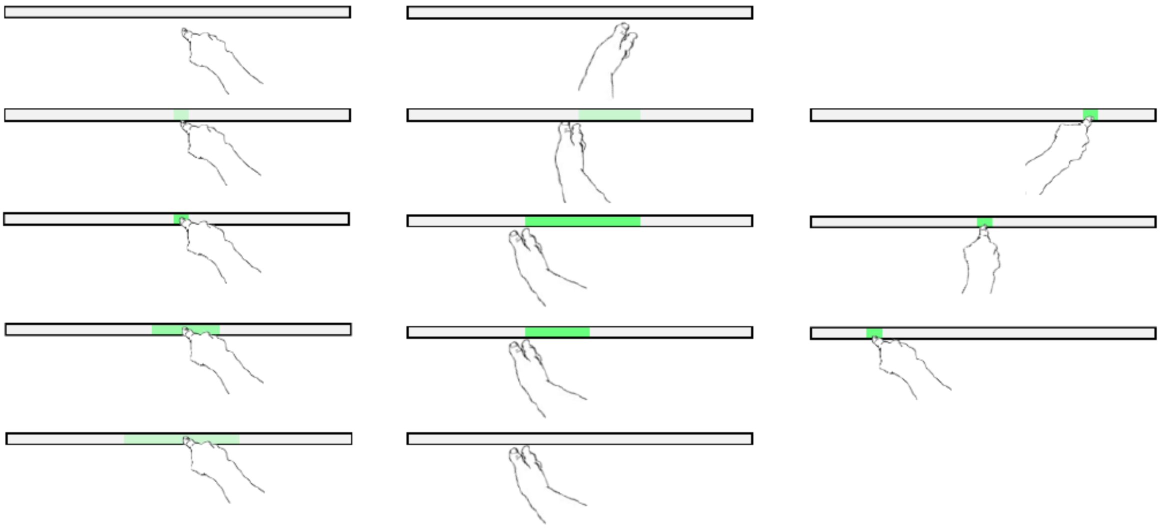


Figure 5.8: The light feedback implementations for confirm (left), swipe (middle) and pointing (right).

Swipe: The user moved his open hand from left to right (right swipe), or right to left (left swipe). Acoustic feedback was given when the swipe was completed. PLF started when the beginning of swipe was detected by fading in a light bar. Along with the progress of the swipe the light bar got more opaque and grew into the swiping direction. This way, the user could get information also in case a swipe was not recognized (e.g., the light bar did not fill completely because the required distance was not covered).

Pointing: The idea of the pointing gesture is to allow the user to select points of interest in his field of view by pointing at them. We used the absolute hand position in the vehicle in combination with the index finger pointing vector to calculate the pointing target on the LED strip at the windshield. A target was confirmed by pointing at it and saying "okay" (the examiner hit a hotkey), which triggered the acoustic feedback. PLF continuously visualized the pointing target with a small light bar at the target point. Due to inaccuracy in the sensor data, which caused flickering of the target point, we decided to define nine larger areas (each 13 cm wide) on the LED strip that could be targeted (see Figure 5.9 for an example of the pointing area).

Apparatus

The experiment was conducted in a static driving simulator. We used an in-house BMW driving simulation, which was displayed on a 55 inch display. Peripheral light feedback was implemented using an LED strip that was controlled using a *Pixelator Mini* [6] *controller*. A Leap Motion gesture controller[7] was placed at the position of the gearshift. The LED strip was 140 cm wide at a distance of 85 cm to the Leap Motion. The positioning of the light strip at the bottom of the windshield has two major advantages. First, it is close to the driver's line of sight reducing glances further away from the driving scene and allowing to create a reference to objects outside the vehicle. Second, compared to typical in-car displays, the light strip it is relatively far away from the driver, reducing the effect of accommodation. Moreover, we implemented a custom middleware that received gesture information from the Leap Motion, calculated the feedback for the gestures (e.g., the position of the pointing target) and controlled the LED strip.

[6] https://www.enttec.com/eu/products/controls/pixel-controller/pixelator-mini/

[7] https://www.leapmotion.com/

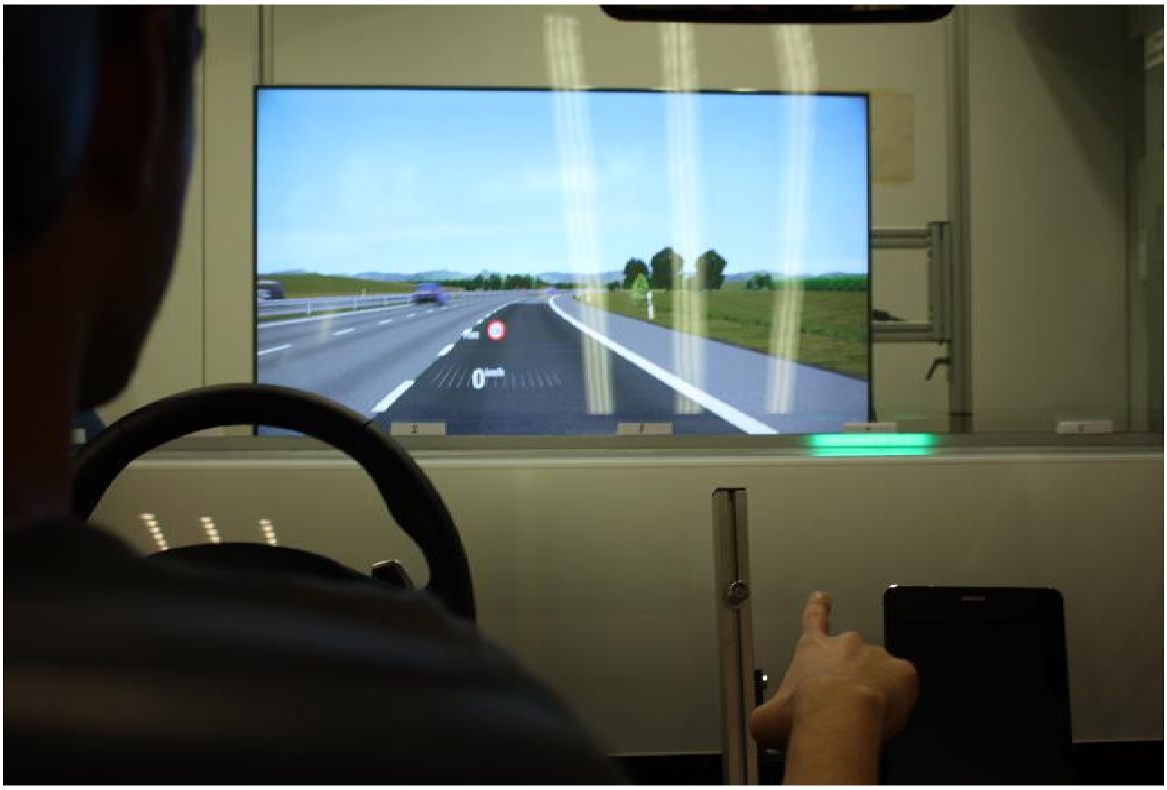


Figure 5.9: This figure illustrates the apparatus for the experiment. In front of the driver is the display with the driving simulation. Behind the steering wheel is the LED strip. This example shows the visualization for the pointing gesture.

Procedure

After being introduced to the topic of the study and the test environment, participants completed a questionnaire covering demographic data. Then they practiced the execution of the three gestures with both feedback types until they had a feeling for the right speed and scope of the movement. Each participant performed two test trials with AF and PLF in counterbalanced order. They drove on a three-lane highway with slight curves and light traffic. After a few minutes of familiarization with the driving simulator, participants were instructed to perform gestures through text-to-speech output. The instructions were "confirm" (occurred five times during one trial), "swipe to the left/right" (five times each), and "point at one/two/three/four/five" (two times each). The numbers of the pointing targets referred to small labels on top of the LED strip (see Figure 5.9). There was a ten second break between the instructions. The order of instructions was counterbalanced over all participants. After both trials, participants gave an overall rating for the gesture interaction, the support of the feedback, and the subjective workload for the feedback types.

5.2.3. Results

Efficacy

We define efficacy as the percentage of correct executions for each gesture. During the trials, only the instructed gesture type was recognized by our software. The confirm gesture was considered correct when the participant triggered it. A swipe had to be performed in the requested direction and a pointing gesture was considered correct, when the pointing target on the LED strip was in the area of the instructed element.

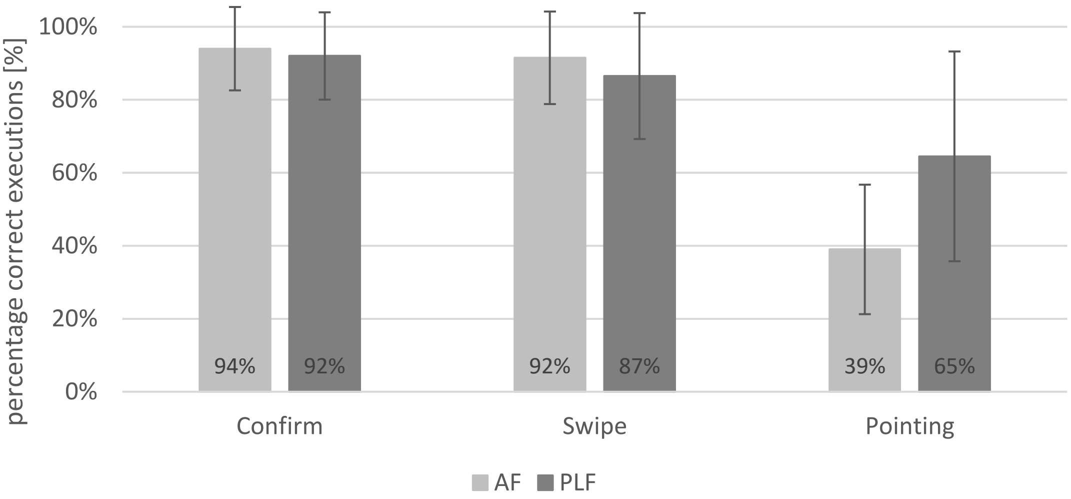


Figure 5.10: The efficacy is the percentage of correct executions for confirm, swipe, and pointing gestures.

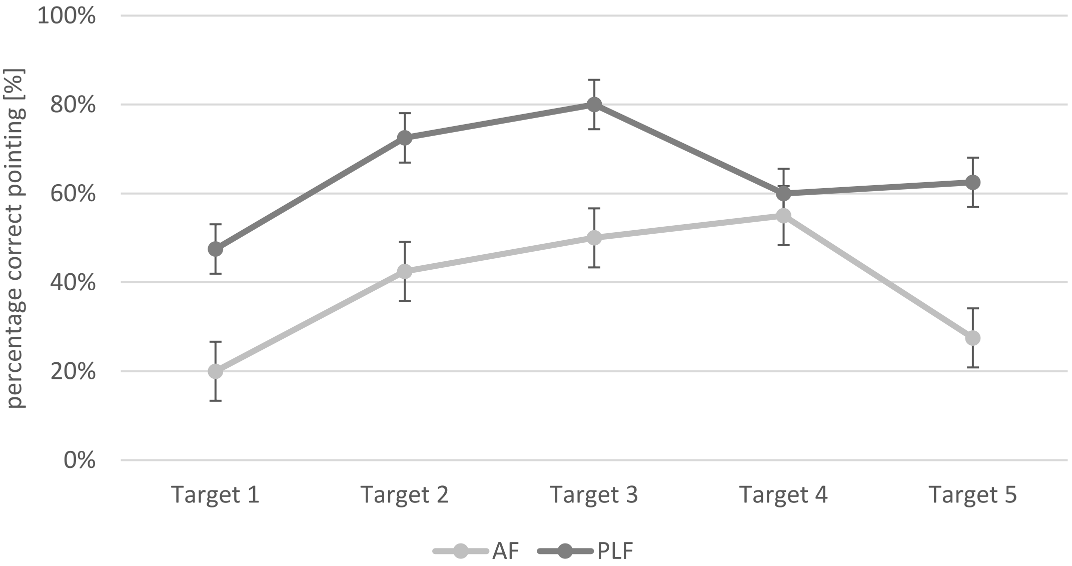


Figure 5.11: The percentage of correct pointing gestures was influenced by the target position. Targets further away from the center had a lower accuracy. This effect could be partially compensated by PLF. Error bars indicate the standard error.

Figure 5.10 shows that the efficacy of confirm (AF: *M=94.0, SD=11.4*, PLF: *M=92.0, SD=12.0*) and swipe (AF: *M=91.0, SD=12.7*, PLF: *M=86.5, SD=17.3*) did not differ between feedback types. In contrast, the feedback type did make a difference for pointing gestures (AF: *M=39.0*, *SD=17.7*, PLF: *M=64.5, SD=28.7*). A Wilcoxon signed-rank test showed that pointing with PLF lead to significant more correct executions than AF ($Z = -2.666, p < .01, r = .59$).

Moreover, our data suggests that accuracy of the pointing gesture depended on the requested pointing target (see Figure 5.11). We observe a lower accuracy for targets further away from the center (target 3). Friedman tests show that target position had a significant influence on the accuracy with AF ($X^2(4) = 10.498, p = .033$), but not for PLF ($X^2$ $(4) = 9.196, \text{ns.}$). This shows that PLF could compensate for the reduced accuracy for pointing targets on the outside.

Task Completion Time

Task times were calculated from the moment of instruction until the successful execution of the gesture. Figure 5.12 illustrates that the difference between AF and PLF for pointing gestures is much more pronounced than for confirm and swipe gestures. In fact, the task times for confirm (AF: *M=2.65, SD=0.99*, PLF*: M=3.11, SD=1.34*) and swipe (AF: *M=3.12, SD=0.89*, PLF: *M=3.10, SD=0.91*) did not differ significantly, but participants needed significantly more time to complete the pointing gesture with PLF (*M=4.47, SD=0.93*) than with AF (M=3.05, SD=0.49, $Wilcoxon$: $Z = -3.362, p < .01, r = 0.83$).

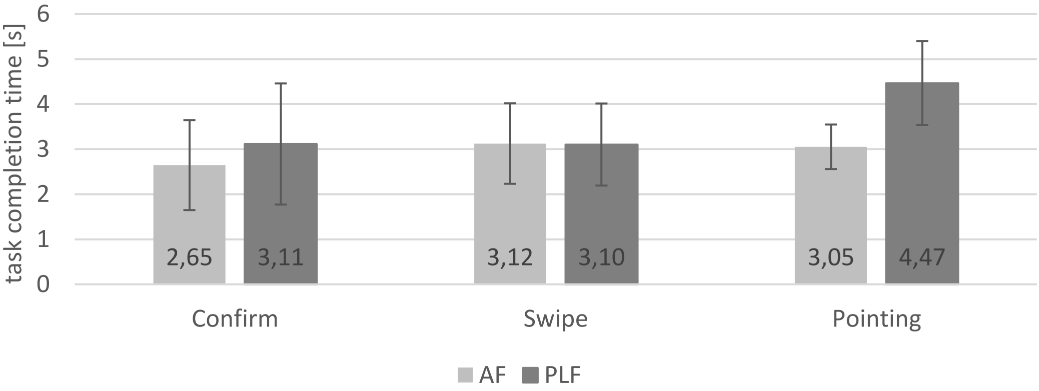


Figure 5.12: The task completion times for confirm and swipe gestures were not affected by the type of feedback. In contrast, pointing gestures took significantly longer when participants had additional PLF.

Subjective Demand

Figure 5.13 illustrates the subjective workload of gesture interaction with both feedback types (over all three gestures). Visual demand was rated noticeable higher for PLF (*M=3.20, SD=1.15*) than for AF (*M=1.90, SD=1.52*). In contrast, auditory demand was lower for PLF (*M=1.35, SD=0.81*) than for AF (*M=2.05, SD=1.19*). Wilcoxon tests show that the differences for both, visual demand ($Z = -2.460, p = .014, r = 0.55$) and auditory demand ($Z = -2.658, p = .008, r = 0.59$) are significant. The average ratings for effort of mental attention, stress, temporal demand, and interference with the driving task are very similar. In accordance with that, the global rating, which is calculated by the mean of all other dimensions does not differ significantly between both feedback types ($Z\ =\ -0.741, ns.$).

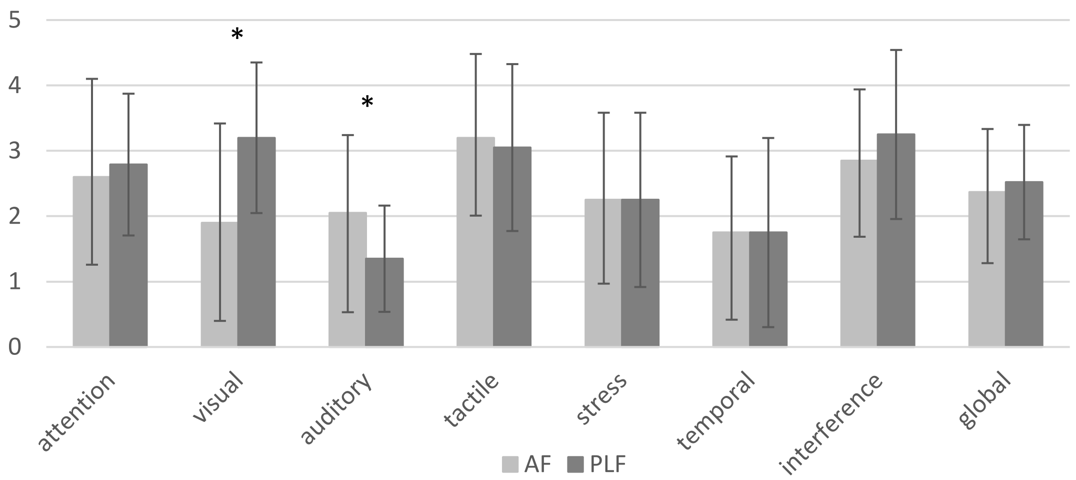


Figure 5.13: The DALI ratings show that the visual demand of PLF was significantly higher than AF. The auditory demand was significantly lower although the acoustic component was the same in PLF. Error bars indicate the standard deviation.

Finally, the participants clearly favored PLF (M=3.65, SD=1.04) over AF (M=2.85, SD=1.14) regarding the overall impressions of the gesture interaction and the support of the feedback type. These results are shown in Figure 5.14. A Wilcoxon signed-rank test showed that PLF was rated significantly better than AF ($Z = -3.36, p < .01$). More specifically, participants felt significantly more supported by PLF (*M=3.85, SD=1.23*) than by AF (*M=2.55, SD=1.67*, $Wilcoxon: Z = -2.41, p = .016$).

5.2.4. Discussion

Our results show that the overall rating of gesture interaction is positively influenced using PLF and participants felt significantly more supported by the additional feedback. While the feedback did not make a difference for the efficacy for confirm and swipe gestures, the percentage of correct selections for pointing was significantly increased by adding PLF. Since the pointing gesture had the greatest room for making errors, a comparison between gestures

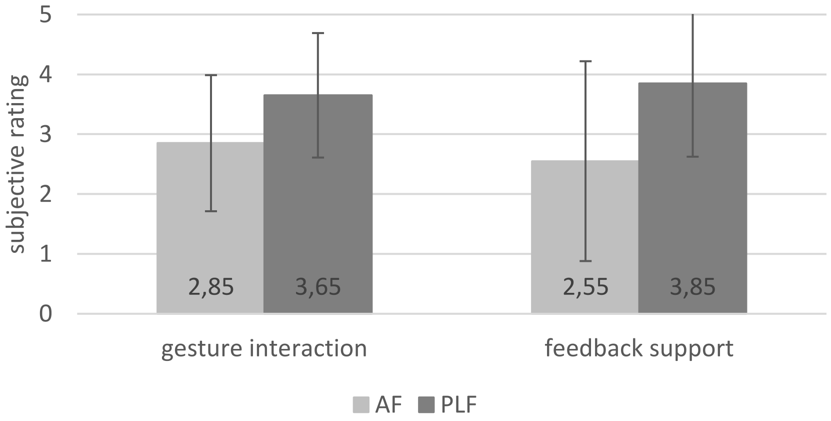


Figure 5.14: The overall rating of the gesture interaction was significantly better with PLF. Participants felt that PLF is significantly more supporting than just acoustic feedback.

would not be fair. The more important insight is that PLF increased the accuracy for the pointing gesture by almost 25%.

However, the tradeoff for higher accuracy are significantly increased task times. In the AF condition, participants did not have the possibility to verify their pointing direction before confirming it. They had to assume that the system correctly recognized where they are pointing at. By enabling the driver to see the current pointing target (in the PLF condition), the responsibility for an exact pointing direction is transferred to the user. This leads to an increased effort for checking and adapting the pointing direction (reflected in the significantly higher visual demand) and finally results in longer task times.

Still, the efficacy of the pointing gesture was relatively low even with PLF. This could be connected to the size of the pointing areas. Earlier research examined finger pointing while driving and identified 4° as the optimal size for target areas in terms of activation times (Cairnie et al., 2000). Compared to this, the target areas in this study used larger angles. They ranged from 6.9° (target 1 and 5) to 10.4° (target 3) (see Figure 5.15). Given the low accuracy of pointing gestures in our results, especially for the outer areas, we propose to use larger target areas of at least 10° in order to receive an accuracy of about 80% (with PLF).

Position	1	2	3	4	5	6	7	8	9
Width	6.9°	8.0°	9.2°	10.2°	10.4°	10.2°	9.2°	8.0°	6.9°

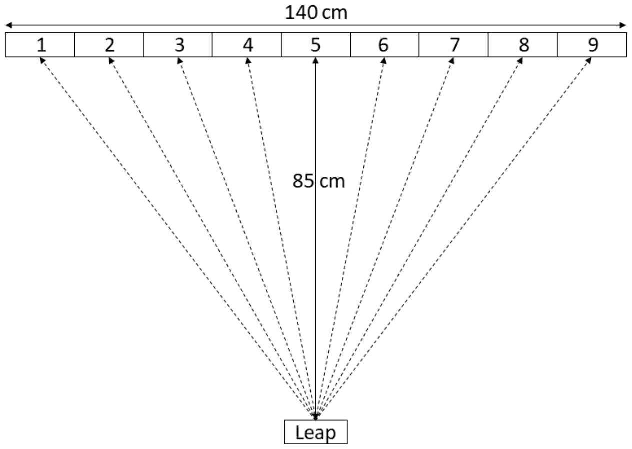


Figure 5.15: The nine target areas on the LED strip were 13 cm wide which corresponds to a width of 6.9° – 10.4° from the position of the Leap Motion controller.

Furthermore, we observe the overall lowest pointing performance for target 1, which was positioned on the left side behind steering wheel. Due to the position of the gesture sensor to the right of the driver's seat, participants had to point using the right hand, which led to unnatural posture of the hand and probably caused the lower accuracy. Automotive applications that enable the user to point towards objects outside the vehicle should therefore enable drivers to use either hand for pointing, whichever is more convenient. This poses new requirements to gesture recognition techniques, which will have to cover a greater interaction space than the triangle between steering wheel, rear mirror, and gearshift described in literature (Riener et al., 2013).

Limitations

The DALI ratings indicate that PLF results in higher visual distraction. However, we could not support this observation with eye tracking measurements. In particular, we planned to track the participants' glances on the LED strip and on the street. However, it showed that the focus area on the street and the LED strip were too close, as the strip was placed directly at the windshield. Therefore, we could not reliably distinguish glances on the street from those on the LEDs. Furthermore, the experiment did not include driving data. A recent study did not find a significant influence of peripheral light feedback behind the steering wheel on driving performance (Shakeri et al., 2017). Nevertheless, the increased visual demand, as well as the trend for higher interference between primary and secondary task in the DALI ratings clearly indicate a higher amount of driver distraction in the PLF condition.

Summary

We presented a user study that compared acoustic feedback (AF) with acoustic plus additional peripheral feedback (PLF) for gestures interaction while driving.

- Gesture interaction benefits from PLF in terms of a better overall rating of gesture interaction and better perceived support by the feedback.
- Simple gestures such as confirming or swiping were not affected in terms of efficacy and task times.
- Pointing gestures profited from PLF by increased accuracy at the cost of longer task times and increased visual demand.

5.3. Summary

This chapter presented two user experiments that demonstrated different techniques on how to support speech and gesture input while driving. In Section 5.1, we explored different visual cues in the user interface to support the usage of speech input. The results show that explicit visual cues are an effective means to increase the usage of speech input and thereby lead to an overall reduction of visual distraction. We concluded that visual cues an unobtrusive way to guide users to make efficient use of multiple alternative input modes by prompting modality switches.

For gesture input, drivers need better support on how to execute gestures correctly. In Section 5.2, we presented a prototype that uses peripheral interior light to provide feedback for gesture interaction. The results of a user experiment show that this form of feedback can increase the accuracy of user inputs and the perceived support of the feedback. Drawbacks are reduced efficiency and higher visual distraction for pointing gestures. Therefore, we recommend using this technique while driving only for very simple gestures, such as confirmation or easy selections.

In summary, the experiments in this chapter showed how to improve the presence of speech and gesture input. Consequently, users get a better understanding of the availability of these input modalities, which is a prerequisite for a flexible use of multimodal interaction in the car.

6. Enhancing Interaction by Combining Input Modalities

The previous chapter focused on supporting speech and gesture input individually without any connection to other input modes. This chapter now demonstrates how existing input modalities can be combined.
The following three experiments present interaction techniques, which combine gaze, speech, gesture input with each other in order to overcome individual weaknesses.

- *Supporting Active Gaze Input* combines gaze input with speech, gestures, and a haptic button on the steering wheel with the aim to create a more efficient and more robust interaction.
- *Enhancing Speech Interaction* integrates gaze input to activate the in-vehicle speech system to increase efficiency and naturalness of speech input.
- *Improving Gesture Accuracy* uses gaze input to enhance the accuracy of pointing gestures while driving.

6.1. Supporting Active Gaze Input

Active gaze input has shown to be a potentially fast but also very distracting input modality (see Section 4.2). However, the efficiency as well as visual distraction of an input modality greatly depends on the specific interaction technique. In other words, how specifically the input modality is used e.g., to make a selection. For example, the last section used a purely gaze-based approach to make a selection. The standard approach for selections in gaze-controlled systems is the use of a dwell time, which has been shown as an effective and convenient method (Huckauf & Urbina, 2011; Jacob & K., 1991).

The major problem of the dwell time approach is that is forces the driver to make glances away from the road, which are longer than naturally occurring glances. This is especially critical since the driving task is mainly visual and therefore the driver's visual resource is overloaded. While this is already a problem for driving on a straight road, the previous experiment has shown, that simple additional visual demands, such as watching out for traffic signs, significantly decreases the efficiency and increases the cognitive workload of gaze input.

Reducing the duration of dwell time could reduce the time the driver looks away from the street, but it comes at the tradeoff of increased potential for wrong selections. Therefore, an alternative to the dwell-time-based approach must be found, in order make gaze input feasible while driving. This section investigates the combination of gaze input with other input modes for the selection of on-screen objects in comparison to a dwell time approach.

6.1.1. Related Work

Fono and Vertegaal (Fono & Vertegaal, 2005) used eye-tracking in combination with a hardware key to select windows in an application with multiple parallel windows. They compared gaze input with automatic selection of the gaze object (without dwell time) to gaze input in combination with a button press. Although they found that the use of a button was bit slower than the automatic selection (about 135ms), they conclude that it was the more effective method overall, as it allowed participants to glance away from the target without triggering other functions.

Therefore, literature in the automotive domain has mainly focused on combining gaze input with hardware buttons on the steering wheel (e.g. (Kern et al., 2010; Poitschke et al., 2011)). Kern et al. also point out the potential of combining gaze input with a speech command to trigger a selection. Please refer to Section 2.2.3 for a more related work about gaze input.

6.1.2. Prototype

Apparatus

The apparatus for the prototype is shown in Figure 6.1. Participants sat in a simple vehicle mock-up on the driver's seat with a steering wheel in front of them. A display to the right of the participants displayed a Unity3D application with an abstract selection task of three rounded icons with symbols (scissors, stone, and paper). An additional laptop with *D-Lab 3.0* with *Dikablis Live Essentials* eye-tracking glasses streamed eye tracking information to the application to capture the participants' gaze on these icons. For speech input, there was an additional head mounted microphone. Speech recognition was handled by the Microsoft Speech Recognition engine, which was integrated in the Unity3D application. Gestures were detected by a Leap Motion controller with was placed below and in front of the secondary display. Another display in front of them displayed the CTT, which was used during the evaluation. It was controlled with two buttons on the left side of the steering wheel.

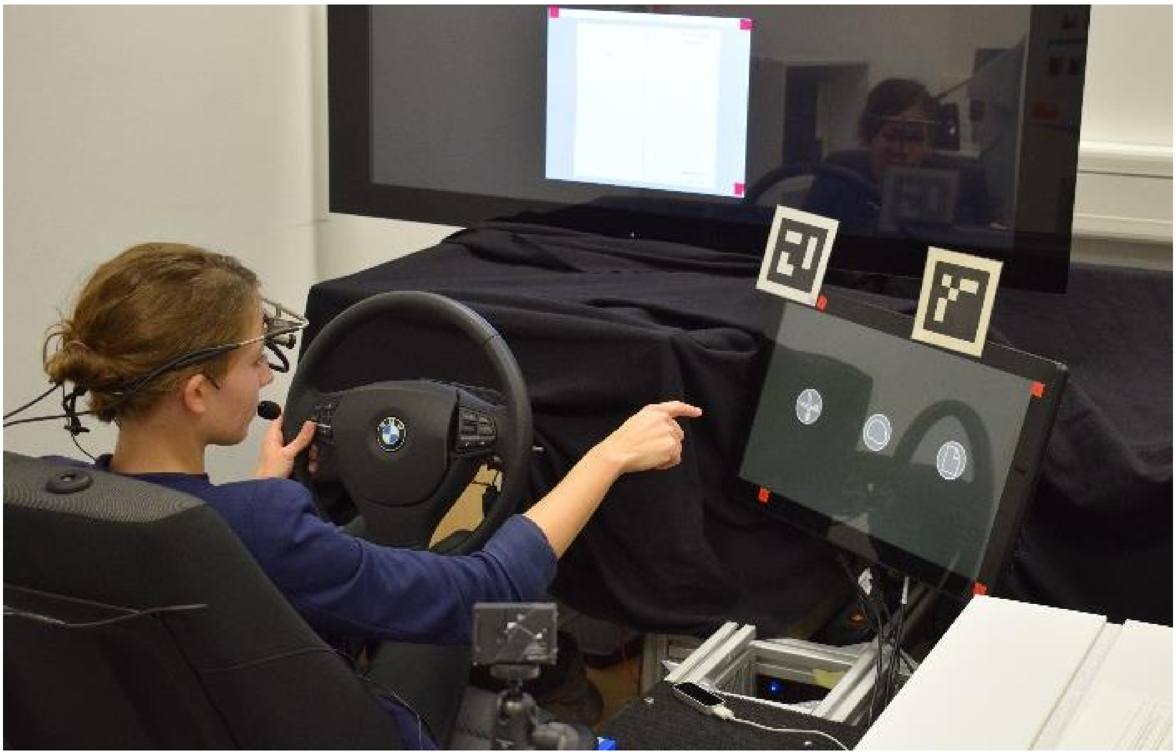


Figure 6.1: The participants focused on of the three elements on the screen with their gaze and made a selection using either a dwell-time, a button on the steering wheel, a voice command, or a gesture.

Selection Techniques

In the *dwell-time* condition, the participants focused the required element with their gaze for a duration of 700 milliseconds to make a selection. A blue circular outline visualized the process of the dwell-time. It started at the top of the icon and filled clockwise until the outline was closed. In a small pre-study, we found that this visualization of the dwell time was preferred by users and resulted in shorter total glance times off the road.

In the next condition a *button* on the right side of the steering wheel was used to make a selection. As soon as participants gazed at an element, a blue outline appeared around it. It could then be selected by pressing the button. The static blue outline to indicate the current gaze element also existed in the following conditions.

Selection via *speech* input could be made with a simple voice command. Participants could say “yes” or “okay” to trigger the selection. As it takes some time to speak and recognize the command, the system cannot give immediate feedback. As a consequence, participants would have to look at the screen until the speech recognition is complete and feedback is given to ensure that they gazed at the correct element. Therefore, we implemented an algorithm that selected the element, which was highlighted in the moment the participants started the voice command. This approach is illustrated in Figure 6.2. It allows participants to look at the requested element, issue a voice command and look back to the primary task without having to wait for the processing of the speech input.

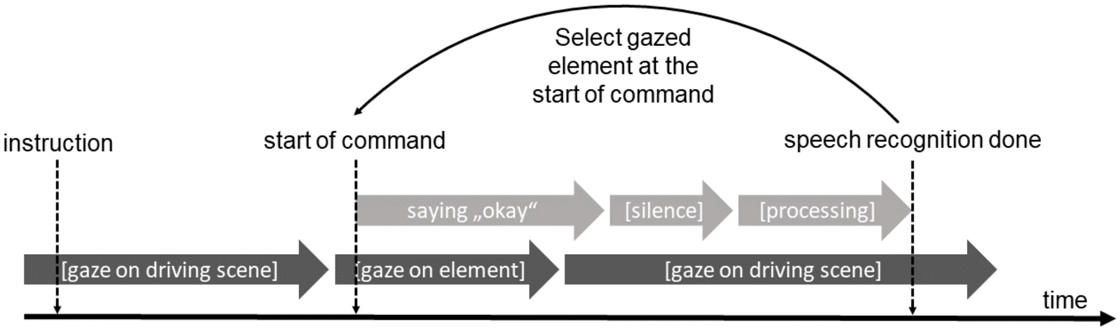


Figure 6.2: For speech selections, the system selected the element that was focused at the beginning of the speech command. Thereby, drivers could look back on the driving scene while waiting for the system to process the command.

Finally, participants could make a selection using a *mid-air gesture*. Since the gesture itself did not have to transmit any information, but its only purpose is to trigger the selection in the right moment, a very simple gesture was used. Analogous to the button and the voice commands, the participants tapped toward with their right index finger toward the screen. A selection was triggered as soon as the distance of the finger to the screen fell below a certain threshold. The pointing direction of the finger was not decisive, i.e., participants did not have to point at the element they wanted to select (since this information was given by the gaze).

6.1.3. Experiment

Participants

A total number of 21 participants (6 female, 15 male) with a mean age of 26.71 years (SD=6.65) took part in this experiment. All but one of them were right-handed.

Design

The experiment used a within-subject design. The independent variable was the selection mode (dwell-time, button, speech, gesture). The order of appearance of each mode was counterbalanced over the participants. The primary task was represented by the CTT. For each selection mode, the participants conducted one trial of 90 seconds.

Procedure

The examiner gave a brief introduction of the study and collected the participants' demographic data. The participants adjusted the position of the seat so that they could comfortably reach into the gesture interaction area and put on the eye-tracking glasses. A short calibration step ensured the functionality of the eye tracker. Then the examiner explained the CTT to the participants and they completed a practice run without any secondary tasks. They also practiced the gaze interaction with each selection technique (first without and later with the CTT at the same time) before starting the measurement trial. Each trial was completed by a DALI questionnaire and suitability ratings on a 7-point Likert-type scale.

6.1.4. Results

CTT Performance

CTT results are illustrated in the left part of Figure 6.3. The best CTT was achieved when the button was used for selection, followed by the dwell-time approach. Selections with gestures and speech commands led to the highest deviations from the optimum. A one-way ANOVA shows that influence of the selection technique on the primary task performance was significant, $F(3,60) = 7.172, p < .001, \eta_p^2 = .264$. In particular, post-hoc pairwise comparisons show that

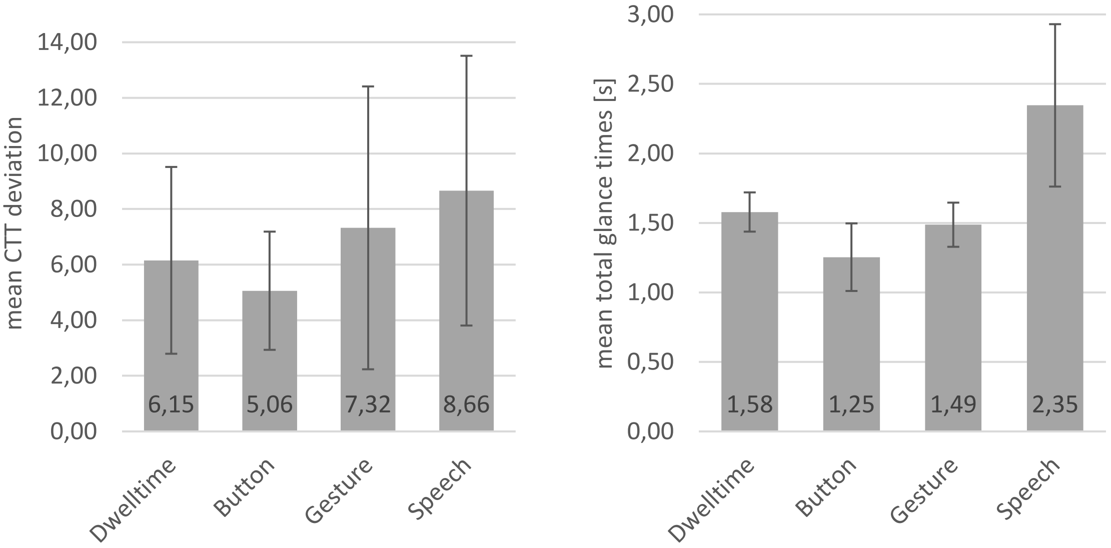


Figure 6.3: *Left*: The best CTT results were achieved when a button was used for selection. Gesture and speech selection lead to relatively highest CTT deviations. *Right*: The total glance time on the screen for the selection of one element was shortest for the selection with the button and longest for the selection with speech. Participants did not look back to the primary task until the selection was confirmed.

selection via speech led to a significantly worse performance than the button ($p < .01$) and the dwell-time approach ($p < .05$).

Glance Behavior

The total glance times on the screen for the selection of one element are illustrated in the right part of Figure 6.3. The results reveal a similar pattern to the CTT results. Selections with the button on the steering wheel resulted in the shortest total glance duration on the screen. Gesture selection and the dwell time are relatively close to that, but speech selection resulted in a very long total glance duration on the screen. A one-way ANOVA shows that the effect is significant $F(3,60) = 50.387, p < .001, \eta_p^2 = .716$. All pairwise comparisons are significant ($p < .01$) except for selections with dwell time and gesture ($p = .078$).

Task Completion Time

The task completion times for the four different selection techniques in Figure 6.4 show a similar picture to the CTT values and total glance times. Again, the selection via button leads to the best results, followed by gesture and dwell time selection. Selection with speech input took more than twice as long as selection with a button or gestures. Accordingly, there is strong effect of the selection technique on TCT, $F(3,60) = 60.697, p < .001, \eta_p^2 = .752$. Post-hoc tests show that the use of button leads to shorter interaction times compared to dwell time and speech (both $p < .01$), but not compared to gestures. However, gestures are not significantly faster than the dwell time approach ($p = .204$). Speech selection leads to significantly longer task times than all other techniques (all $p < .01$).

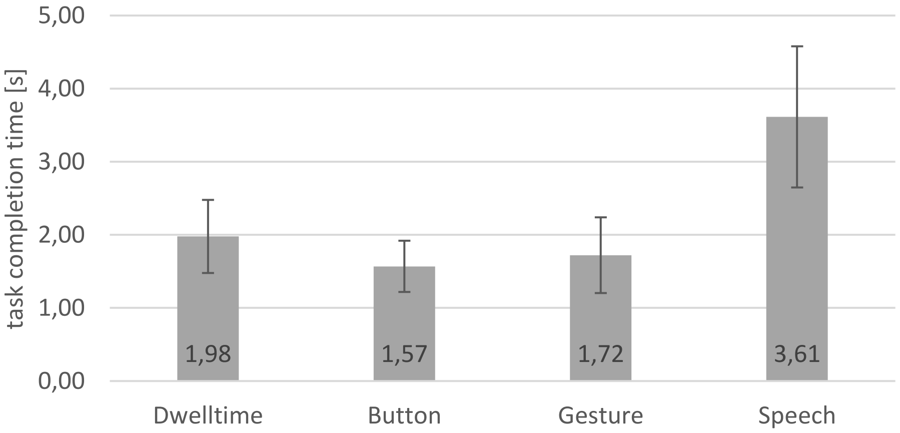


Figure 6.4: The task completion times are shortest for the selection with the button, followed by gestures and dwell time. According to glance times and CTT results, speech selection performed worst.

Error Rates

The conditions differed regarding the ratio of correct selections. Figure 6.5 illustrates the rate of correct selections as well as erroneous selections. Later could be either empty selections or wrong selections. Empty selections describe the case that the eye-gaze did not indicate one of the available icons on the screen in the moment of selection. This type of error did never occur in the dwell time condition, as the gaze could only dwell on the screen icons. The second type of error are wrong selections. They indicate those selections in which the eye-gaze indicated an incorrect icon on the screen in the moment of selection.

The dwell time condition resulted in the highest percentage of correct selections (*97%*) and therefore also had the lowest error rate. The percentage of correct selections for button (85%), gesture (76%) and speech (90%) conditions are noticeably lower. A Friedman test confirms that the effect of the selection technique is significant, $\mathrm{X}^2(3) = 32.33, p < .001, n = 21$. Except for the dwell time and speech conditions, all pairwise comparisons are significant (all $p < .05$).

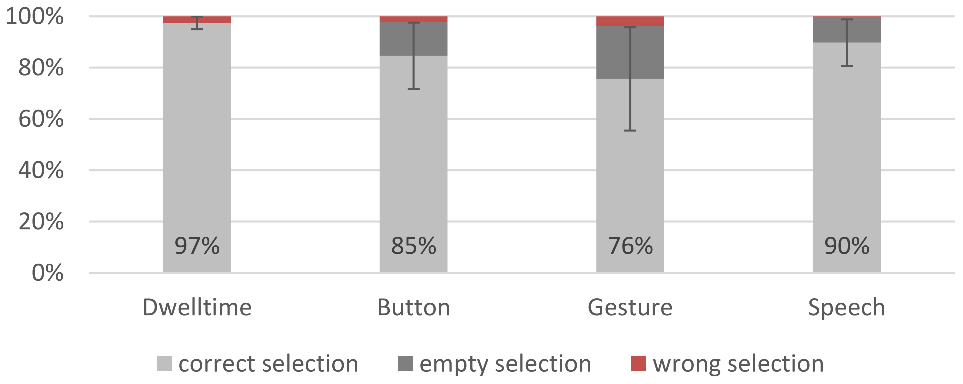


Figure 6.5: The conditions differed regarding the ratio of correct selections.

Subjective Demand

Participants rated the subjective demand after each trial. Figure 6.6 summarizes the results. For the most dimensions of the DALI there is the trend that the button condition leads to the lowest demands while the other three conditions are up to one point higher on the rating scale. The only exception to this is the stress dimension. Speech selection is rated as the least stressful condition, although it is rated highest in other dimensions such as attention, temporal demand, and tactile demand. Besides that, the global dimension provides a good representation of the overall demand. The data is distributed normally. A one-way ANOVA shows a significant effect of the selection technique on the perceived subjective demand, $F(3,60) = 6.630, p <$

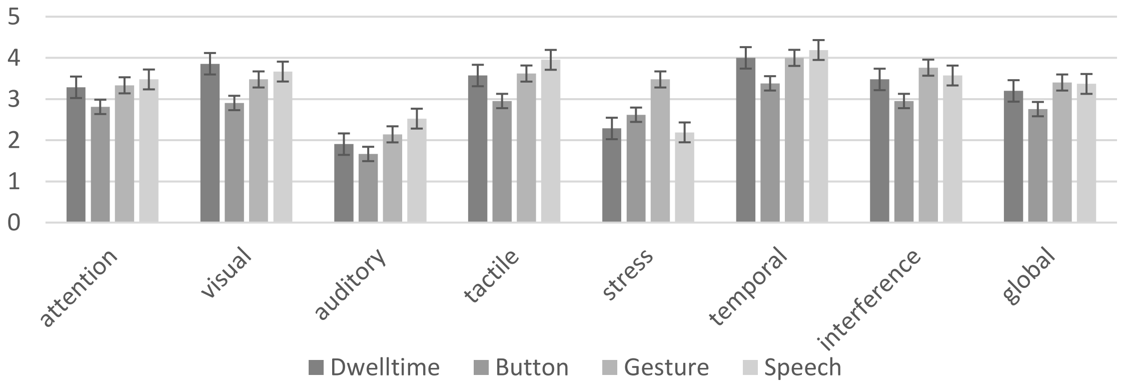


Figure 6.6: The DALI ratings show that the button was perceived as the least demanding selection technique. Error bars indicate the standard error.

$.01, \eta_p^2 = .249$. The button selection is significantly lower than all other conditions (dwell time: p < .05, gestures and speech: p < .01). The differences between dwell time, gesture and speech are not statistically significant.

Suitability Rating

The suitability ratings for each condition represent the previous results, although there was a great variance in the data. Participants rated the selection techniques on a scale from +3 (very high suitability) to -3 (very low suitability). Data was not distributed normally. The button is rated as the most suitable condition (*Mdn = 2*), followed by dwell time (*Mdn = 1*) and speech (*Mdn = 1*). The least suitably condition was the gesture condition (*Mdn = -1*). A Friedman test shows that the perceived suitability of gaze input differs significantly depending on the selection technique, $X^2(3) = 9.34, p < .05, n = 21$. However, none of the post-hoc pairwise comparisons were significant due to large variance. The ratings for the dwell time and gesture condition ranged from a minimum of -3 (dwell time and gesture) and -2 (button and speech) to a maximum of +3 for all conditions.

6.1.5. Discussion

Gaze input can be efficiently used as pointing input, but it struggles to provide a suitable way of confirming the pointing target, which is needed to avoid the "*Midas Touch*" problem. We compared speech commands, gestures, a button on the steering wheel and a dwell time based approach to confirm gaze selections on a screen.

The haptic button on the steering wheel resulted in the shortest TGT and TCT and therefore also allowed the best CTT performance. This supports previous work that combined gaze with haptic input on the steering wheel ((Ecker, 2013; Kern et al., 2010; Poitschke, 2011)). In comparison to the dwell time approach, the explicit button confirmation significantly reduced TGT. However, the difference was surprisingly small (330 ms) given the relatively long dwell time of 700 ms. This leads to the conclusion that the time to make a decision and consciously push the button also takes a considerable amount of time (about 370 ms). This duration is comparable to the times that Schneegaß et al. (Schneegaß et al., 2011) have observed for a button press during the development of a keystroke-level model in the automotive domain (540 ms). One disadvantage of the button is high error rate of 15% in comparison to the dwell time (3%). Most of the errors were empty selections. This indicates a potential problem of the temporal coordination of visually focusing an element and pressing the button at the right time.

The gesture selection allowed to explicitly confirm the selection with a mid-air tap towards the screen. Interestingly, the use of the gesture could not significantly reduce CTT deviation, TGT or TCTs compared to the dwell time approach. Instead, we observe the highest number of erroneous selections (see Figure 6.5). Most of them were empty selections (21%), which again indicates problems with temporal coordination of both gaze and gesture input. Furthermore, we observed that these errors occurred because of participants often tended to look at their own hand performing the gesture. This way the gaze was drawn away from the target icon on the screen and participants instead looked at their right hand in the moment of the selection. In accordance with this observation, seven participants commented on the gesture condition that it was "too much action" and therefore distracting.

The selection via a speech command led to longest glance duration away from the primary task by far. During the explanation of the selection techniques, the participants were told that they could look back to the CTT as soon as the speech command was issued. However, it could be observed that the participants did not make use of this possibility but kept their eyes on the requested icon until the system recognized the command and issued a feedback. The estimated average response time of the system was about 700-900 ms, but even without considering this time, the TGT of speech would still hardly be shorter than the dwell time condition. This observation demonstrates two things. First, long task completion times with speech input cannot exclusively traced back to system response times. Especially when combined with a short-term temporary state, such as gaze focus on the screen while driving, speech input lacks the possibility to give quick and precisely timed input. Second, the use of speech input does not necessarily lead to a reduction of visual distraction. The lack of immediate feedback leaves the driver in a moment of uncertainty while looking at the screen and waiting for system feedback. As a consequence, the secondary task claims much more visual attention than necessary.

Given the problems of the speech and gesture selection, the dwell time approach worked relatively well. Although it resulted in the lowest error rates of all conditions, this experiment did not respect the risk of unwanted selections. Mere inspection of the screen was not possible, as participant always interacted when looking at the screen.

Summary

This experiment investigated different confirmation modalities for gaze pointing input as an alternative to the dwell time approach. Speech commands and gestures are less suited for confirmation of gaze input due to modality-specific problems. A selection button on the steering wheel resulted in the best primary and secondary task performance. Still, the results demonstrate that active gaze input while driving is critical, as it is not meant to be consciously controlled for manipulating objects. For this reason, we don’t see active gaze input as a safe nor a convenient input modality while driving.

- Gaze focus is a short-term temporary state that requires precisely timed input. It benefits from a confirmation mode that allows fast input and provides direct feedback.
- Button confirmation can increase efficiency and reduce glance-times.
- Speech confirmation takes too long to speak and process and is therefore not well suited for such timely critical confirmations.
- Gestures could not provide an advantage over the dwell time approach. There was interference with the gaze selection which resulted in high error rates, as the eye-gaze got easily distracted by the hand movements.
- A relatively high cognitive load persists for all variants and the perceived suitability does not improve compared to the dwell time approach.

6.2. Enhancing Speech Interaction

This section is based on the following publication:

Roider, F., Reisig, L., & Gross, T. (2018). Just Look : The Benefits of Gaze-Activated Voice Input in the Car. In Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI ’18) (pp. 210–214). (Roider, Reisig, & Gross, 2018)

Voice input provides the potential for safe, efficient and natural in-car interaction (e.g. (Alvarez et al., 2011; Maciej & Vollrath, 2009; C. a. C. a. Pickering et al., 2007). Therefore, effort is put into the improvement of automatic speech recognition (ASR) systems, to increase recognition accuracy and to integrate more functionality. Yet, before users can actually make a voice command, all current voice recognition systems require an explicit activation to express that the user's command was directed to the system.

The most widespread option in current vehicles is the use of a push-to-talk (PTT) button on the steering wheel. However, the activation via a haptic button is an artificial action that is in contrast to the intelligence of (future) voice interaction systems and does not live up to drivers’ expectations of natural voice interaction (Ramm, Giacomin, Robertson, & Malizia, 2014). In this regard, the activation via keyword (e.g., "Hey Mercedes ...“[8]) is a much more natural option that also appears in human communication (e.g., "Hey Bob ...“). However, compared to a simple button press, the utterance of a keyword takes relatively long and can easily get annoying for the user for multiple requests. In summary, there is a trade-off between efficiency (PTT) and naturalness (Keyword) for currently used voice activation techniques.

6.2.1. Related Work

In interpersonal communication, people use another mode to express their focus of attention. It has been shown that eye-contact most reliably indicates the target of a users' attention (Vertegaal, Slagter, van der Veer, & Nijholt, 2001). Maglio et al. further investigated the role of gaze patterns during the communication with multiple intelligent devices. They found that users nearly always looked at a device before making a request (Maglio, Matlock, Campbell, Zhai, & Smith, 2000). Based on that, Oh et al. investigated the use of gaze to activate the automatic speech recognition to enable more natural human-computer interaction in a collaborative environment. They compared this look-to-talk (LTT) approach, compared to a keyword-based implementation and PTT and concluded that LTT was a promising approach (under ideal conditions, i.e. good recognition accuracy, short latency) (Oh et al., 2002).

For applications in the driving context, it must be considered that drivers are in a dual-task situation with a visually demanding primary task. Their eye-gaze is mainly focused on the street, even when talking to a passenger next to them. Thus, the results about the potential of LTT in a conversational setting cannot directly be transferred to the automotive domain. Moreover, we have to take additional aspects into account besides the naturalness of interaction, such as distraction and efficiency.

[8] https://www.mercedes-benz.com

This section describes the concept of LTT to activate an in-car speech recognition system while performing a visually demanding primary task. We analyze distraction, efficiency, and user experience and compare LTT with state of currently available activation modes, PTT, and Keyword.

6.2.2. Prototype

We implemented a prototype that supported all three activation modes in an automotive setup. It is displayed in Figure 6.7. A directional microphone was used for recording voice commands. The software prototype ran on a Windows 10 machine and was implemented in Unity3D, which used the integrated system speech engine in Windows 10. A Tobii 4C[9] eye-tracker was mounted behind the steering wheel to track the user's gaze.

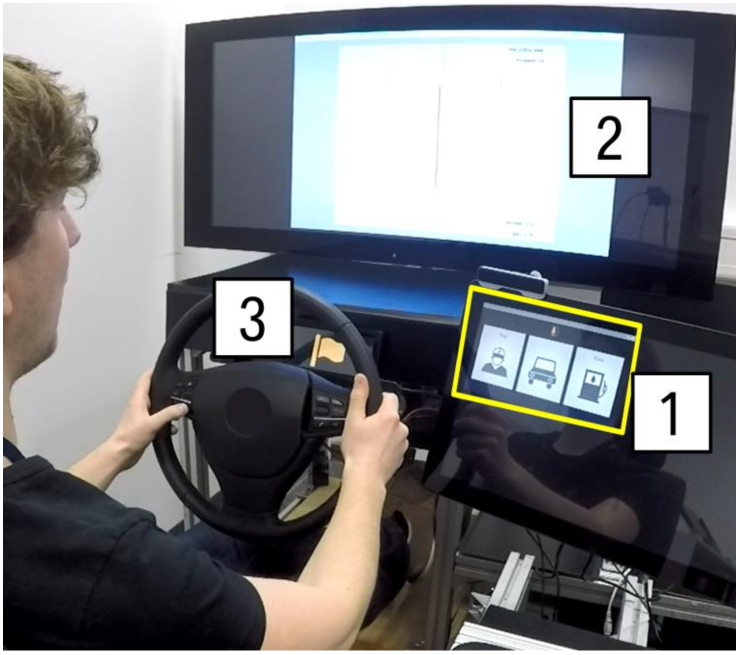


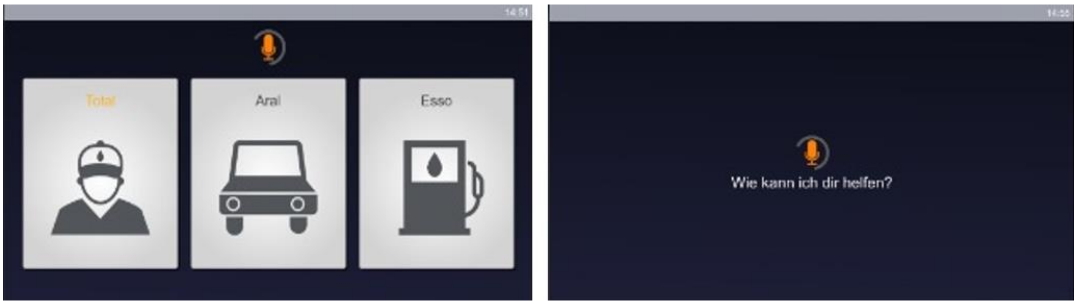
Figure 6.7: The prototype for the Look-to-Talk prototype used a Tobii 4C eye-tracker mounted behind the steering wheel. The speech recognizer was activated when the driver gazed at the CID.

The prototype supported two types of secondary tasks. They are displayed in Figure 6.8. In the *navigation* task, users chose one out of three gas stations (e.g., "Navigate to Esso"), which were displayed on the gaze-aware screen area. The *navigation* task is a *display-related* task, since the driver needs the information on the screen in order to know the options he can choose from. In the *telephone* task, users made a phone call to a person (e.g., "Call Lisa"). In this case, the driver does not need any information from the screen. Therefore, the telephone task is a *non-display-related* task.

Before making a voice request, users have to activate the ASR by looking at the screen area for at least 300 milliseconds, which approximately represents the duration of a short eye fixation

[9] https://tobiigaming.com/product/tobii-eye-tracker-4c/

(Jacob & K., 1991). This short delay ensures that the system would not get active when the driver’s gaze is unintentionally passing over the screen, e.g., when the driver has to look at the right side mirror. Besides that, it can help to reduce activations based on noisy eye tracking data.



Wie kann ich dir helfen?

Figure 6.8: The prototype supported a display-related navigation task (left) and a non-display related telephone task (right). Both tasks show an animated microphone icon as a representation of the ASR.

An animated microphone icon shows the state of the ASR and served as a visual representation to support making eye-contact with the screen (see Figure 6.8). It provides drivers something to look at (like Sam in (Oh et al., 2002)), and makes sure that they receive feedback when eye-contact is established. Upon activation, the icon animation started and the system played an earcon ("pling"). It stays active for at least four seconds after driver’s looked away from the screen, so that they can look back on the street after the activation. Upon recognition of the voice command there was another earcon ("plong") and the task disappeared.

6.2.3. Experiment

Participants

A total of 25 participants (9 female, 16 male) with mean age of 30.24 years (*SD=11.53*) took part in the experiment. Four participants used speech recognition regularly, whereas one participant did not have any experience. The majority used speech recognition at least occasionally.

Study Design

We used a within-subject design with repeated measures. There were two independent variables: three modes of speech activation (PTT, Keyword, and LTT) and two task-types (navigation, phone). Participants completed one trial for each combination of speech activation and task-type. The order of appearance of task-types and activation modes was counterbalanced over all participants. We used the critical tracking task (CTT) as a primary task to create a constant visual demand on the participants. It was displayed on a screen in front of the participants (see Figure 6.7). Another small display was placed behind the steering wheel. It was used to instruct the name of the person to call for the phone task, so that participants did not have to look to the screen area. In the navigation task, the instruction was made directly on the screen area to the right. Users were asked to pick the gas station whose name was written in orange.

Procedure

The experiment started with the adjustment of the seat position and the calibration of the eye-tracker. The examiner explained the CTT and the two task types with the different activation techniques. All participants practiced the CTT and the secondary tasks and quickly familiarized with it. Before each trial, they were informed about the upcoming task and activation technique. One trial consisted of ten voice commands with ten second breaks between a successful command and the instruction of the next task. We recorded the participants' total glance time on the screen, the task completion times to assess the efficiency of each mode and participants completed the User Experience Questionnaire (UEQ).

6.2.4. Results

Task Completion Time

In both tasks, the shortest task completion times (TCT) were achieved with LTT, followed by PTT, and keyword. They are illustrated in Figure 6.9. A repeated measures ANOVA showed that the activation mode had a significant effect on TCT ($F(2,48) = 32.82, p < .01,\ \eta_p^2 = .58$). TCT with the keyword was significantly slower than LTT ($p < .01$) or PTT ($p < .01$), which is probably due to the time that is needed to speak and process the keyword. PTT and LTT allow a much faster and more direct activation.

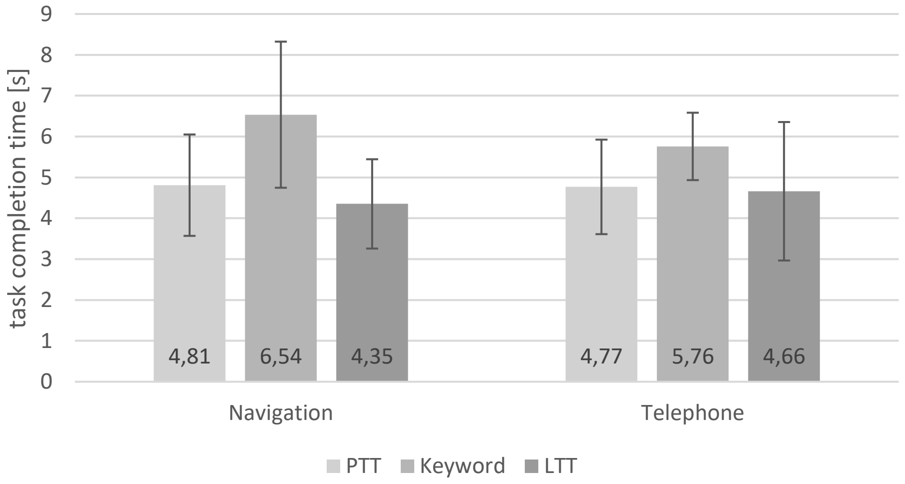


Figure 6.9: Visualization of the Critical Tracking Task: The user's task is to keep the black vertical line in the middle of the screen (dotted line).

Glance Behavior

The total glance time (TGT) is the aggregated duration of glances on the gaze-aware screen area while a task was active. In the *navigation* task, TGT hardly differed between PTT, Keyword, and LTT (see Figure 6.10). In all three conditions the participants had to glance at least once at the display to see the selectable elements. The fact that LTT did not result in a longer TGT away from the primary task shows that the naturally occurring glances (that also appear in the PTT or Keyword condition) can be efficiently used for activation.

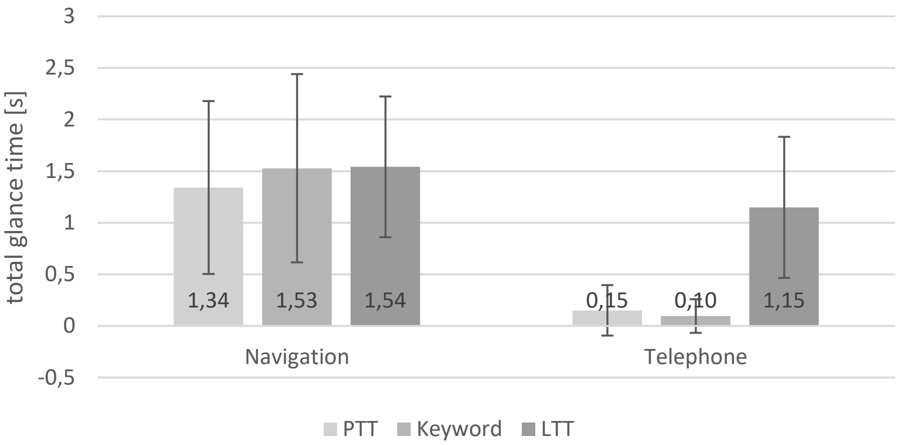


Figure 6.10: The total glance times indicate that the visual distraction of the three activation modes is depending on the type of task.

In contrast, in the phone task the activation mode had a significant influence on TGT (Friedman test: $F(2) = 39.56, p < .01$). LTT was significantly longer than PTT ($p < .01$) and Keyword ($p < .01$). PTT and Keyword did not need glances on the screen, because the task does not refer to on-screen information. LTT however needs a short glance at the screen to activate system. Although TGT for LTT is relatively short (*M=1.15, SD=0.68*) these glances are not naturally occurring, but rather an additional, artificial movement with the eyes. Despite the effects on glance behavior, the mean CTT scores for both tasks did not significantly differ across activation techniques.

Subjective Demand

The participants' ratings on the DALI score are illustrated in Figure 6.11. The global dimension ranges between 1.18 and 2.04, which is relatively low, as there was a break of approximately ten seconds between two tasks. In the navigation task, LTT was rated as the least demanding

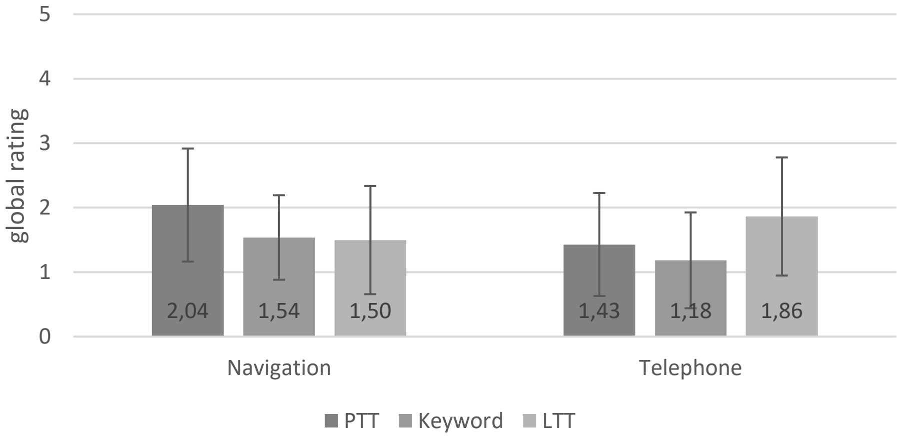


Figure 6.11: The global DALI scores show the influence of the task type on the subjective demands. LTT is rated as the least demanding activation mode in the navigation task, but the most demanding for the telephone. Error bars indicate the standard deviation.

activation mode (*M=1.50, SD=0.84*). A one-way ANOVA shows that the activation mode has a significant effect, $F(2{,}48) = 7.375, p < .01, \eta_p^2 = .235$. LTT (p< .0) and Keyword (p <.01) are both significantly less demanding than PTT. In contrast, LTT was the most demanding mode for the telephone task (*M=1.86, SD=0.92*). The effect of the activation mode was significant, $F(2{,}48) = 12.386, p < .001, \eta_p^2 = .340.$ LTT was significantly more demanding than PTT (p < .05) and Keyword (p < .01). Finally, a two-way repeated measures ANOVA shows that there are significant effects of the activation mode ($F(2{,}48) = 6.292, p < .01, \eta_p^2 = .208$) and the task type ($F(2{,}48) = 4.622, p < .05, \eta_p^2$=.161). Keyword was significantly less demanding than PTT (*p < .01*), or LTT (*p < .05*) in both tasks, while PTT and LTT are strongly depending on the task type. This is also shown by a strong interaction effect between both factors, $F(2{,}48) = 8.294, p < .01, \eta_p^2 = .257.$ In other words, the effect of the activation modes on the subjective demand does strongly depend on the type of the task.

User Experience

The dimensions of the UEQ can be summarized regarding the pragmatic and the hedonic quality of the interaction. In the left part of Figure 6.12, we see that the pragmatic quality was depending on the task type. There is a significant interaction between activation mode and task type, $F(2{,}48) = 8.499, p < .01, \eta_p^2 = .262$. For the telephone task, LTT had a lower pragmatic value than the Keyword (*p < .05*) and PTT (*p < .01*), but for the navigation task the mean LTT rating was higher than the other modes (*ns.*). The pragmatic value of the keyword activation was between PTT and LTT for both tasks.

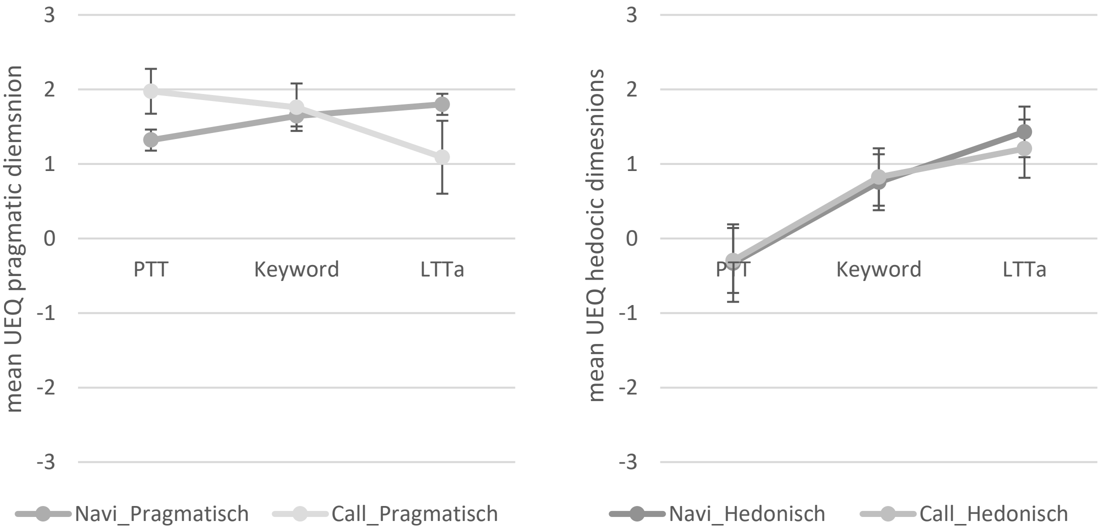


Figure 6.12: There is a strong interaction effect for the pragmatic quality (left). It shows that the pragmatic quality of the activation mode depends on the task. The hedonic quality is greatest for LTT, independent of the task type (right). Error bars indicate 2SE.

In contrast to the pragmatic qualities, the hedonic quality (right part of Figure 6.12) was not influenced by the task type ($F(1{,}24) = .210, ns.$), but the activation mode strongly effects both tasks in the same way ($F(2{,}48) = 43.80, p < .01,\ \eta_p^2 = .65$). LTT was rated significantly higher than Keyword (*p < .01*) and PTT (*p < .01*). Despite some pragmatic disadvantages for the telephone task, it is noticeable that this did not impair the high perceived hedonic quality of the LTT approach.

6.2.5. Discussion

The results of the user experiment confirm the advantages and limitations of current voice activation techniques.

The Push-to-Talk button is an efficient input modality with low demands that is a pragmatic solution with little visual distraction for non-display-related tasks, such as the telephone task, or entering an address for route guidance via voice. When it comes to display-related tasks, PTT is still efficient, but it is connected with higher demands and thus seen as less pragmatic. From a hedonic perspective, PTT was rated worst for all tasks.

The activation via keyword is fundamentally different. Regarding efficiency the keyword performed worst in both tasks. On average keyword was 1.64 seconds slower than LTT and 1.36 seconds slower than PTT. This duration is composed of the time to speak the keyword by the user ("Hey BMW") and the time to recognize the keyword and open the ASR by the system. Future voice recognition systems will reduce the latter factor, but a certain amount of time to utter a voice command will always persist. In fact, activation keywords are explicitly chosen to have multiple syllables ("Hey Siri", "Ok, Google", "Alexa"). Activation words with only one syllable could speed up the entire interaction, yet they are easily overheard by current voice recognition systems and have a higher risk for unwanted activations. Therefore, very short utterances are not suited as activation keywords. Apart from the limited efficiency of the keyword, it causes little visual distraction, and results in the lowest demands on the drivers. Moreover, it is seen as a pragmatic activation mode for both tasks, while PTT and LTT strongly depend on the type of the task.

The LTT approach aimed to overcome the limitations of PTT (poor user experience) and Keyword (low efficiency). As the task completion times show, LTT was the most efficient solution. TCT even shorter than PTT with an average delta of 283 milliseconds (*ns.*). Regarding user experience, the hedonic quality was rated highest for both tasks. The Hedonic quality represents attributes like stimulation and originality. While the haptic button press for PTT has been well known in cars for years and keyword activation can be found in many current consumer electronic devices (though relatively new in cars), LTT was a new experience for almost all participants. Pragmatic qualities however, are not throughout positive. There is an opposite trend to the PTT. LTT is rated as the most pragmatic solution for *display-related* tasks, but the least pragmatic mode in *non-display-related* tasks. In the display-related navigation task, TGT with LTT did not differ from Keyword. This indicates that the participants did not change their gaze behavior. Instead, the system made use of the naturally occurring glances to activate ASR without additional effort of the user. These glances also occur in the PTT and Keyword condition, but in contrast to LTT these activation techniques require an additional activation step. Thus, LTT is perceived as a very pragmatic solution. In contrast, in the non-display-related telephone task, LTT required users to make one extra glance at the screen area, while PTT and Keyword allow to do the task without any glances on the screen. Total glance times of LTT for non-display-related tasks (1.15 seconds) are on average one second longer than PTT and Keyword. This short duration is relatively small and not to be considered as safety critical according to NHTSA guidelines (NHTSA et al., 2016). However, it is an extra effort and thus less pragmatic.

Summary

This experiment compared the gaze-based look-to-talk (LTT) activation mode for speech input with conventional activation modes. Current activation modes for in-vehicle speech interfaces are either slow (keyword) or lacking a user experience (push-to-talk). For two different task types, the look-to-talk concept resulted in highest efficiency and good user experience. It contributes to the hedonic quality of the system, while a pragmatic improvement only showed for display-related tasks. Therefore, LTT cannot replace other activation modes, but for some tasks it can be a valuable alternative, e.g., to the keyword approach, which was perceived as good solution for both task types.

- LTT is a more efficient interaction technique for voice activation than a keyword and push-to-talk.
- LTT leads to a high perceived hedonic quality for both, display-related and non-display-related tasks.
- For display-related tasks, LTT results in lowest workload and highest pragmatic quality.
- For non-display related tasks LTT results in highest workload and low pragmatic quality.
- There is an increase of visual distraction is increased for non-display-related tasks, but not for display-related tasks.

6.3. Improving Gesture Accuracy

This section is partly based on the following publication:

Roider, F., & Gross, T. (2018). I See Your Point : Integrating Gaze to Enhance Pointing Gesture Accuracy While Driving. In Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI ’18) (pp. 351–358). https://doi.org/10.1145/3239060.3239084 (Roider & Gross, 2018)

Mid-air gestures have been repeatedly shown as a promising method for interaction with secondary functions in the vehicle while driving. In this context, it is very important to distinguish, which form of gesture interaction is used, because they vary in regards of usability and demands on the driver. Section 2.2.3 gives a more detailed description of different gesture types. Unlike symbolic gestures, pointing gestures do not have to be learned by the user. They allow to create simple deictic references to all kinds of real and on-screen objects. Users are enabled to interact with a wide range of vehicle functions without having to learn new gestures, which is particularly helpful for novice users (Ahmad & Langdon, 2018). During the execution of a pointing gesture, users must localize a pointing target and make a coordinated pointing movement with their hands. Compared to symbolic gestures, this requires a greater amount of the users' visual attention. However, regarding the advances in autonomous driving and the increasing number of driver assisting functions in modern vehicles, increased visual attention is acceptable, when user experience, effectiveness, and ease-of-use for operating secondary functions are increased in return.

Experiments have shown that drivers make relatively large pointing errors while driving (Brand et al., 2016; Roider & Raab, 2018). This is in line with findings from earlier experiments, which found that pointing errors especially occurred, if users cannot move their head towards the target (Biguer, Jeannerod, & Prablanc, 1982). The authors conclude that the data of eye-gaze fixations provides information about the pointing target before the arm movement has event started. In fact, it has been found that the user's gaze is actually anchored to the pointing target during a pointing movement (Neggers & Bekkering, 2001). This has also been shown in an experiment in the automotive domain. The drivers' eye-gaze is also fixed on pointing targets during freehand pointing while driving (Ahmad & Langdon, 2018). This knowledge suggests that the close relationship between eye-gaze behavior and pointing gestures might be used to improve pointing performance while driving, without creating additional (visual and cognitive) load on the user. At the same time, Section 6.1 shows that the driver's gaze behavior is easily distracted even by simple gestures and may therefore not always provide helpful input.

Considering all these factors, this section describes a practical implementation of a simple algorithm that passively integrates gaze information with the aim to increase the accuracy of pointing gestures while driving. Moreover, we report the results of a user experiment that reveals benefits and downfalls of the approach.

6.3.1. Related Work

The combination of gaze information with gestures was examined in a number of HCI experiments that explicitly use gaze information to select objects on a screen and gestures to manipulate selected objects. Chatterjee et al. showed how the combination of gaze and gesture input can overcome gaze-only or gesture-only systems (Chatterjee, Xiao, & Harrison, 2015). Zhang et al. presented a similar approach that enhanced the interaction efficiency compared to a gesture-only interface, but they also encountered problems regarding the participants eye-hand coordination (Zhang et al., 2015). Similar approaches have been presented in the automotive domain. Nesselrath et al. used gaze information to select real objects of the vehicle, such as side mirrors or windows. Gestures on the steering wheel or speech commands could then be used to control these objects (Neßelrath, Moniri, & Feld, 2016). Kern et al. showed the application of gaze information to select objects on a screen in combination with a haptic button on the steering wheel to confirm the selection (Kern et al., 2010).

All these prototypes have in common that they share the same approach for the integration of gaze information. It is used as a conscious, active selection tool, combined with a second modality for modification. Gaze input replaces the function of another input mode (e.g., touch or mouse input for selection). Salvucci et al. point out two problems that emerge from such gaze-based interfaces: the noise and limited availability of eye tracking information, and the dissociation between the user's glance behavior and the actual visual attention (Salvucci & Anderson, 2000). Especially the latter one, is very relevant in the automotive context. Driving is typically a dual-task situation with driving the car as the primary task and operating non-driving related functions as the secondary task. Since steering a vehicle is visually very demanding, the driver's glances may be directed towards the street, although the mental attention is on the completion of a secondary task. Therefore, the authors propose the usage of gaze-added interfaces, which provide the same basic functionality as non-gaze interfaces, but add the ability to incorporate gaze information, if available (Salvucci & Anderson, 2000).

Zhai et al. presented such a gaze-added prototype that combines passive gaze information with active mouse input. It passively tracked users' eye movements to predict the pointing target of the mouse and uses this information to enhance the movement of the mouse cursor (Zhai, Morimoto, & Ihde, 1999). Oviatt et al. describes this form of multimodal integration as *blended* multimodal interaction. The passive input mode is used to improve the multimodal system's prediction and interpretation of the active mode (Sharon Oviatt et al., 2009).

A number of studies over the last years presented other promising approaches on how to increase the accuracy of mid-air pointing gestures. Mayer et al. demonstrated how systematic displacement of different ray-casting approaches can be compensated using two-dimensional polynomials (Mayer, Wolf, Schneegass, & Henze, 2015). Plaumann et al. showed the influence of ocular dominance and handedness on pointing gestures. They present a selection algorithm, which uses this information, to increase the users' pointing accuracy (Plaumann, Weing, Winkler, Müller, & Rukzio, 2017). Providing additional visual feedback, such as a cursor, further increases accuracy, but at the same time it increases interaction times as well as subjective demands on the user (Mayer, Schwind, Schweigert, & Henze, 2018; Roider & Raab, 2018). For pointing in the automotive domain, the visual demand of the driving task, noisy sensor data, or unintended movement due to driving and road conditions further limit the effectiveness of pointing gestures. Ahmad et al. presented a Bayesian framework that takes additional sensory data from the vehicle (e.g., such as suspension travel data) into account in order to predict freehand pointing targets. Though not evaluated, they also propose that eye-gaze data could offer valuable information on areas of interest on the display (Ahmad, Murphy, Godsill, Langdon, & Hardy, 2017).

In summary, pointing gestures enable drivers to interact with a wide range of vehicle functions, without the high learning effort of symbolic gestures. While technical challenges to detect user's pointing direction exist, a more fundamental problem is that users pointing movements often lack sufficient accuracy to identify user intentions. Existing approaches make use of mathematical functions, users' ocular dominance or vehicle sensor data to increase pointing performance. Eye-gaze data has been shown to provide meaningful information about the users' attention, especially while performing finger pointing movements. In the automotive domain, gaze input has been used as an active selection tool, but it has not been used as a passive input modality to improve pointing gestures. We propose a first practical approach, how gesture data could be enhanced with passive eye-gaze data, to compensate for the lack of pointing accuracy.

6.3.2. Prototype

Apparatus

The prototype is displayed in Figure 6.13. It consisted of a driver seat and a *Thrustmaster* force-feedback steering wheel including foot pedals for throttle and brake. In front of the driver was a LG 34-inch curved monitor, which was recessed in a wooden construction, so that the visible height of the display was only 18 cm. The left side of the display showed only the vehicles speed and vehicle status. On the right side there were four horizontally aligned elements. In Figure 6.13 the elements show a selection of different background colors. In total, there were three different types of contents (Colors, Media, and Routes). Elements were squared with a side length of 6.75 cm x 6.75 cm.

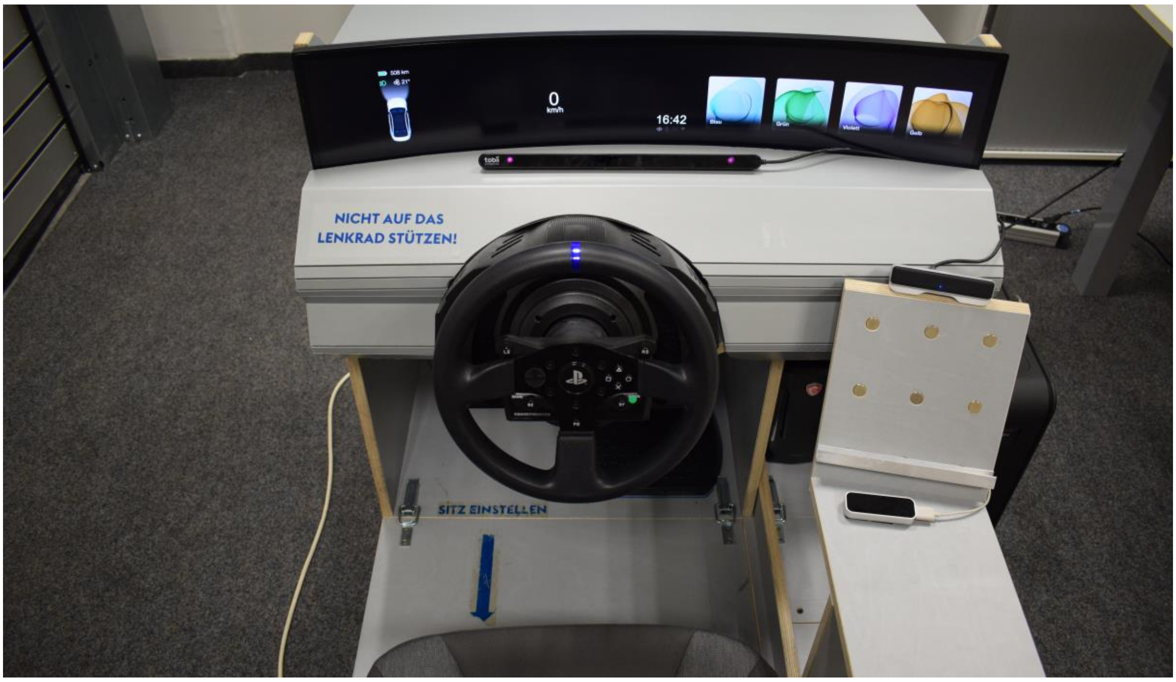


Figure 6.13: The prototype consisted of a gesture sensor to right of the participants to recognize pointing movements towards the four elements on the right side of the display. The eye-tracker was mounted below the display.

For the recognition of the pointing gestures, a *Leap Motion* gesture camera was placed to the right side of the driver's seat to cover the typical in car interaction space in the center of the car (Riener et al., 2013). The distance between the *Leap Motion* and the screen was 33 cm, which results in an angle of 11.6° for each element on the screen, in accordance with Section 5.2.4. Eye-tracking was achieved using a *Tobii 4C* eye-tracker. It was placed in front of the driver and below the curved display. The placement of the eye-tracker was critical to make the combination of gaze and gestures work. Originally, the *Tobii 4C* tracker is developed to work in a Desktop setting, where it is placed directly under the interactive part of the display in a central position. In combination with pointing gestures in this the prototype, the tracker had to be placed to the left of the elements. Otherwise, the driver would occlude the eye tracker when reaching out towards the elements during the pointing movement.

Fusion Algorithm

This approach aims to integrate gaze data in a passive way. This means that users do not consciously use their gaze as an input modality. They might not even know that their gaze is taken into account to modify the gesture selection. As a consequence, their gaze does not necessarily focus on the pointing target element. The challenge for combining passive gaze with pointing gestures is to determine whether the gaze provides relevant input that refers to the current interaction with the system. Based on literature, we know that the gaze point is anchored to the pointing target while pointing (Ahmad & Langdon, 2018). Conversely, we assume that gaze information does not refer to the target of the pointing gesture when the targets points indicated by both modalities differ by more than a certain amount.

Based on this assumption, the algorithm uses three heuristic rules for the fusion of gesture and gaze information. According to the four elements on the screen, the virtual interaction pane was horizontally split into four equally spaced parts. The horizontal coordinate of the point of intersection of the fingertip with the interaction pane in the moment of selection determined the gesture target element ($elem_{Gesture}$). Gaze information was incorporated in the moment of the selection. The user's gaze point on the screen determined, which of the four elements the user is gazing at ($elem_{Gaze}$).

We calculate $diff$ as the absolute difference of the indices of both elements:

$$diff = |\ elem_{Gesture} - \ elem_{Gaze}\ |$$

- $Case\ A: if\ (diff == 0)\ select\ elem_{Gesture}$
  Both input modes indicate the same element. No correction is needed.

- $Case\ B: if\ (diff\ ==\ 1)\ select\ elem_{Gaze}$
  Gesture and gaze indicated neighboring elements. We assume that the gaze refers to the selection, but $elem_{Gesture}$ is wrong due to inaccurate pointing.

- $Case\ C: if\ (diff\ >\ 1)\ select\ elem_{Gesture}$
  Both elements differ by more than one position. We assume that the gaze does not refer to the target element.

These three cases are also illustrated in Figure 6.14.

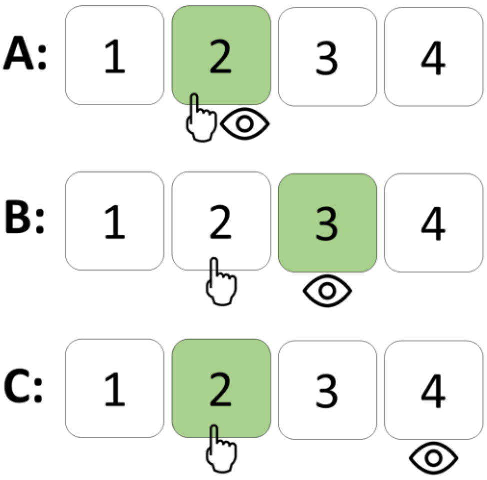


Figure 6.14: The selection algorithm determines one element based on the pointing gesture and one based on the user's gaze. Both elements are merged into one fused element (green) based on three heuristic rules.

Selection Modality

In order to select one of the four elements on the right side of the screen, participants made a pointing movement with their right index finger by moving the outstretched finger towards the screen. The selection was triggered as soon as the finger entered the virtual mid-air interaction pane, with was placed approximately 26 cm in front of the display with the four selectable elements.

Alternatively, selections could also be triggered with a speech command ("Yes, "okay, or "this one"). In this case, the moment of selection is determined by the *beginning* of the speech command. The algorithm takes the horizontal pointing position and the gaze focus on the screen at the moment the user starts to utter the speech command and not from the moment when the recognition is completed. This is implemented in the same manner as described in Section 6.1.2 for the confirmation of gaze input.

6.3.3. Experiment

We conducted a user experiment that examined the benefits of the gaze-added selection algorithm in a driving context. While the main focus of the experiment was gesture-based triggering of the selection, we also investigated the use of speech for triggering the selections.

Participants

A total number of 22 (6 female, 16 male) participants with a mean age of 32.82 years (*SD=8.59*) took part in this experiment. The sample was generally inexperienced with the use of mid-air gesture interaction. 12 participants have never used any form of mid-air gesture interaction before, 7 had very little experience, and 3 participants reported to use it from time to time. None of participants used gesture interaction on a regular basis. All of them had a valid driver's license. Their annual mileage was an average between 10.000 and 20.000 kilometers a year.

Design

We used a within-subject design for this experiment. Each participant completed one run in the driving simulator with each trigger modality. The order of runs was counterbalanced over all participants. During the runs, the participants were texturally instructed to select one out of the four elements on the screen (e.g., "Select green*"*). The instruction was additionally announced by a short sound. In total there was a number of 16 selections with 10 seconds between a selection and the next introduction. The participants were not told that their gaze information was integrated for the selection, because we wanted to avoid an artificial gaze behavior and instead focus on investigating the benefit naturally occurring glances while pointing. Upon selection, the participants received visual feedback, which element was selected based on the fusion rules. It was highlighted for one second and an earcon was played, but the feedback did not explicitly differentiate whether the selection was correct. The system logged the percentage of correct selection based on gaze (*gaze accuracy*), gesture (*gesture accuracy*) and the fused result (*fused accuracy*). By comparing gesture accuracy and fused accuracy, we can determine the effect of the gaze-added system instead of a unimodal approach.

Procedure

At the beginning, the participants adapted the position of the seat (height and distance to the steering wheel) according to their size and length of their arms, so that they could comfortably reach the virtual interaction pane as shown in Figure 6.15. They were informed about the procedure of the experiment and completed a questionnaire covering demographic data. This was followed by short calibration of the gaze system. The experimenter told the participant that the gaze is used to measure visual distraction from the driving task.

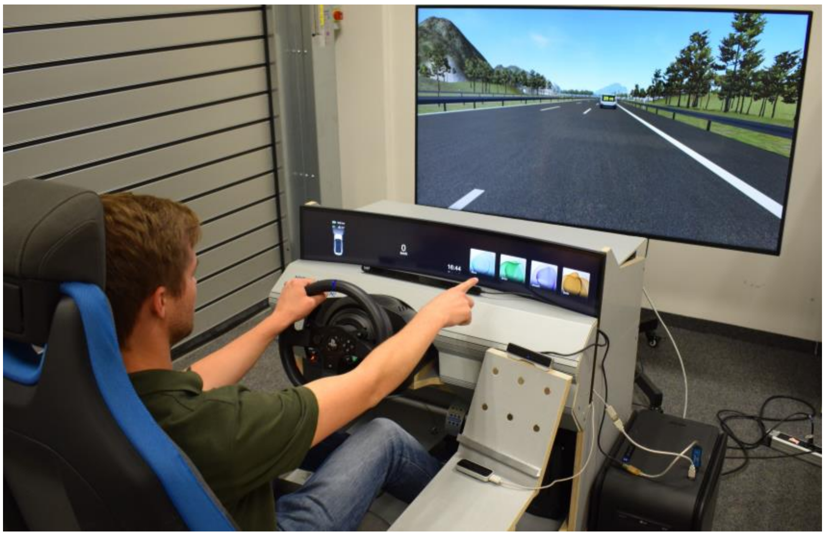


Figure 6.15: Participants repeatedly selected elements on the right part of the screen by pointing at them while following a leading vehicle.

After successful calibration the participants had a few minutes for practicing the gesture selection of the four elements without a driving task. Additionally, there was a practice run with the driving simulator of at least 2 minutes (or more based on the assessment of the experimenter). The driving task was to follow a leading vehicle on a highway with three lanes at a speed of 100km/h. The road was slightly winding and the there was only very little traffic, so that the following vehicle always stayed on the rightmost lane. Finally, the participants started the measurement run. They were instructed to prioritize the driving task over the selection of the elements.

6.3.4. Results

The main part of this section covers only runs that used gesture-triggered selections. A comparison with speech-triggered selections is described in a subsection at the end of this results section.

	All	Element 1	Element 2	Element 3	Element 4
Case A	60.23%	82.22%	60.47%	46.59%	51.14%
Case B	22.73%	1.11%	29.07%	39.77%	21.59%
Case C	2.56%	0.00%	2.33%	4.55%	3.41%
No Gaze	14.49%	16.67%	8.14%	9.09%	23.86%

Table 6.1: *This table illustrates the occurrence of the three cases. The columns summarize selections over all elements (first column) and depending on the target element. The last row shows the percentage of selection for which participants did not gaze at either of the four elements.*

Table 6.1 summarizes the occurrence of the three cases describe in Figure 6.14. Pointing gestures and gaze indicated the same element in an average of 60.23% of the selections (case

A). In these cases, no correction was applied. Both inputs differed by one element in 22.73% of the selections (case B). For these selections the element indicated by the user's gaze was selected. Deviations of more than one element occurred only in 2.56% of the cases (case C). In the remaining 14.49% the participants' gaze point was not detected on any of the elements in the moment of selection.

The data in Table 6.1 further indicates that the occurrence of the different cases is depending on the requested element. For example, case B occurred only in 1.11% of the selections for *Elem1*. Gaze matched with the gesture information in 82.22%. This means that the algorithm did hardly do anything for *Elem1*. In contrast, it changed 39.77% of the selections for *Elem3*. These numbers demonstrate how many of the selections were changed, but they do not show if those adaptions actually led to an increase of the selection accuracy.

Fused Accuracy

The *gesture accuracy*, the *gaze accuracy*, and the combined *fused accuracy* of both inputs are illustrated in Figure 6.16. Over all elements, the average *gesture accuracy* was 72.73% (*Median=81.25%*) and *gaze accuracy* was 78.69% (*Median=87.50%*). There was a great variance over all participants for both individual input channels, which demonstrates that both input modalities differed in how well they worked for individual participants. Gesture accuracy ranged between a minimum of 37.50% and a maximum of 93.75%. Gaze accuracy varied between 25.00% and 100%. Fusion of information from both sources led to a fused accuracy of 89.20% (*Median=93.75*. Wilcoxon signed rank tests show that the fused accuracy is significantly higher than gesture accuracy ($Z = -2.93, p < .01, r = .63$) and gaze accuracy ($Z = -3.44, p < .01, r = .73$).

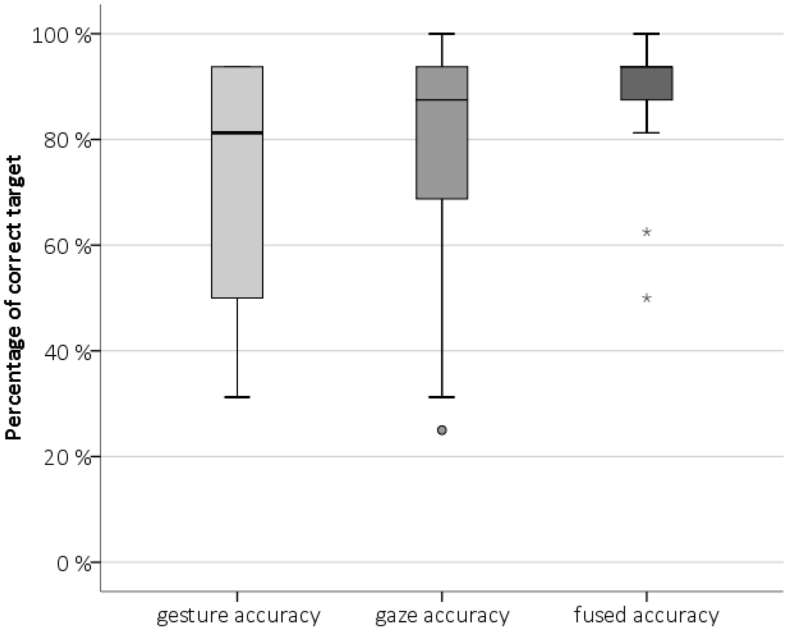


Figure 6.16: By combining the gesture data with gaze information the resulting fused accuracy could be significantly increased in comparison to either the input modes alone.

There is an average increase of accuracy of 16.48%. However, the application of the fusion algorithm did not lead to an increased performance for all participants. Four of 22 participants had a higher gestures accuracy and the gaze input falsified some selections. Three participants were neither improved nor worsened. Overall, the improvement of the fusion algorithm was negatively correlated to the initial gesture accuracy ($r = .-850, p < .01$). This is also illustrated in Figure 6.17. The room for improvement was obviously limited for participants with good gesture accuracy. The other way around, participants with low gestures accuracy profited from a significantly increased accuracy.

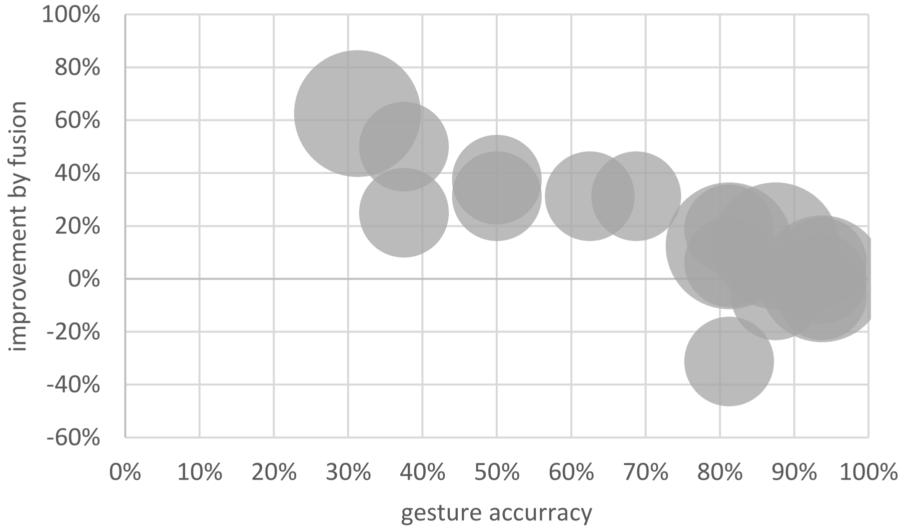


Figure 6.17: There is a negative correlation between the gesture accuracy and the improvement made by the fusion algorithm. The size of the bubbles illustrates the number of occurrences.

Influence of Element Position

The accuracy of the pointing gesture was highest for the leftmost element (M=96.82%, SD=8.24%, Median=100%). Figure 6.18 shows that the gesture accuracy was noticeably lower for elements further right. Element 3 was correctly indicated by the pointing gestures in only 52.27% (SD=39.27%, Median=62.50%) of the selections. The accuracy rates were not normally distributed across participants. A Friedman test shows that the element position had a significant effect on the pointing gesture accuracy, $\mathrm{X}^2 = 20.05, \mathrm{p} < .01, \mathrm{n} = 22$. Due to high variances, a statistically significant difference could only be found between gesture accuracies for elements 1 and 3 ($Z = 1.500, p < .01, r = 0.32$).

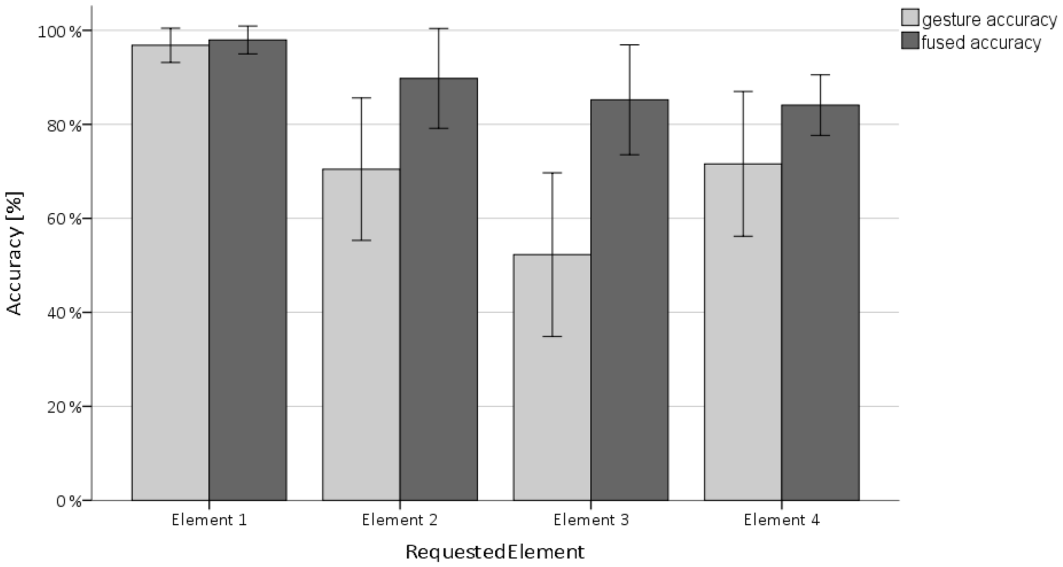


Figure 6.18: The improvement of the algorithm did strongly depend on the requested element. There was only little increase for Element1 while elements further right profited more from additional gaze information.

Speech Trigger

The experiment also investigated the use of a speech selection besides the gesture selection technique. The *gesture* selection was triggered as soon as participants entered the interactive pane as described in Section 5.4.2. For the *speech* selection, the participants pointed at the target object and confirmed the selection by saying “yes”, “okay”, or “this one”. The algorithm selected the element that was targeted at the beginning of the speech command.

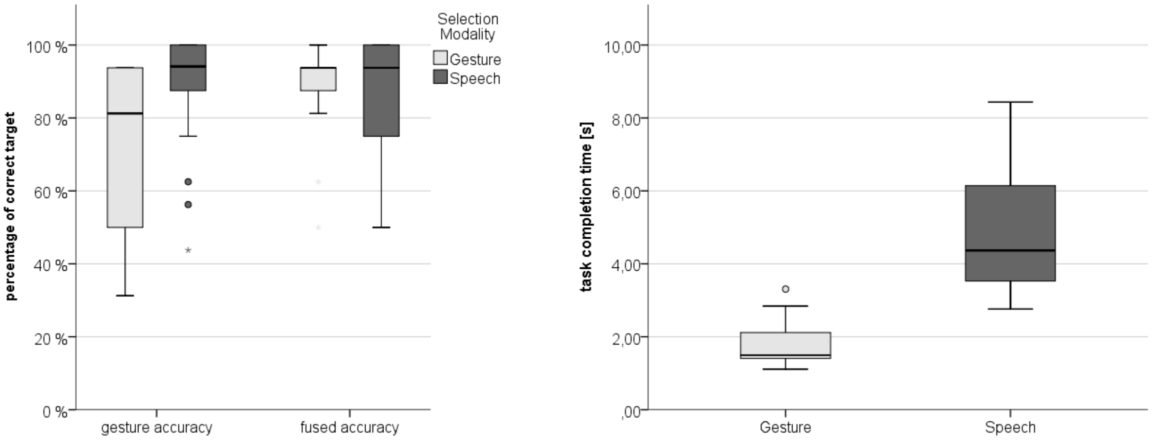


Figure 6.19: Accuracy (left) and task completion times (right) for the confirmation with speech and gestures.

The left graph in Figure 6.19 illustrates the effects of the selection modality on the gesture accuracy (without gaze information) and the fused accuracy (with gaze information). Speech selection (*Median=94.09%)* led to a significant improvement of the gesture accuracy compared to gesture selection (*Median=81.25%*), $Z = -3.709, p < .001, r = .791$. However, when incorporating gaze information, both input modalities performed equally well (both *Median=93.75%*). The right graph in Figure 6.19 shows the task completion times of gesture and speech. On average, speech selections (*M=4.76, SD=1.56*) took more than three seconds longer than the selection with gestures (*M=1.74, SD=0.55*), $t(21) = -10.218, p < .001, r = .912$.

Besides these objective observations, participants rated the selection with gestures and speech on the DALI questionnaire. The results are shown in Figure 6.20. Except for the temporal dimension, speech confirmation led to significantly higher subjective demands. The summary of the statistical tests is given in Table 6.2.

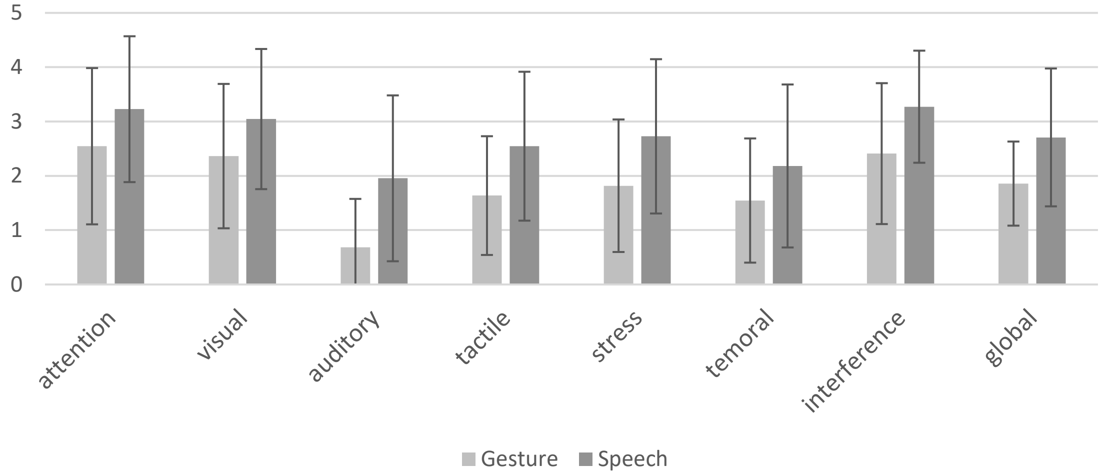


Figure 6.20: DALI ratings for confirming pointing gestures with speech.

DALI-dimension	Gesture	Speech	Wilcoxon
Attention	2.55 (1.44)	3.23 (1.34)	Z=2.839, p < .01,
Visual	2.36 (1.33)	3.05 (1.29)	Z=2.579, p < .05
Auditory	.68 (0.89)	1.95 (1.53)	Z=2.999, p < .01
Tactile	1.64 (1.09)	2.55 (1.37)	Z=2.793, p < .01
Stress	1.82 (1.22)	2.73 (1.42)	Z=2.752, p < .01
Temporal	1.55 (1.43)	2.18 (1.50)	Z=1.806, p = .07, ns.
Interference	2.41 (1.30)	3.27 (1.03)	Z=2.582, p < .05
Global	1.86 (1.00)	2.71 (1.16)	Z=3.340, p < .01

***Table 6.2**. Statistical overview and comparison of DALI ratings for gesture and speech confirmation.*

6.3.5. Discussion

The results show that the gaze-added algorithm led to an overall improvement of selection accuracy for pointing gestures while driving. Users did not know that their gaze behavior influenced the selection algorithm. Therefore, we claim that the approach does result in increased visual demand compared to a normal pointing gesture selection. The increased accuracy of the system might instead lead to general reduction of driver demands, since the correction of wrong selections would draw more of the user's attention. In this context, it has been shown that the enhancement of mid-air selection based on additional data results in reduced driver demands (Ahmad, Langdon, Godsill, Donkor, & Wilde, 2016).

The participants' gaze indicated the correct element in 78.69% of all selections. This supports the findings in literature that observed that the driver's eye gaze is fixed on the target in pointing selection tasks (Ahmad & Langdon, 2018). However, there are also 21.31%, in which the gaze did not indicate the correct element. There are two main possible reasons. First, this could be due to noisy eye-tracking data. Second, the driver might not look at the screen in the moment of the selection, because he made a control glance at the driving simulation, or reached out to the elements before looking at the screen (this was also observed by (Ahmad & Langdon, 2018)) and accidentally triggered a selection.

The results show that the benefit of the approach was depending on the position of the target elements. The maximum difference in accuracy (between elements 1 and 3) was 44.55% for gesture results. These results are in line with findings in literature (Biguer, Prablanc, & Jeannerod, 1984) as well as those from Section 5.2: the pointing accuracy decreases with pointing targets moving further away from the user's line of sight. The fusion algorithm can partly counteract this observation. It reduced the maximum difference between elements to 13.86% (between elements 1 and 4), but the statistical test indicates that there is still a weak effect ($r = .22$) of the target position.

Moreover, the benefit of the approach is influenced by the trigger modality. Speech confirmation led to a higher gesture accuracy. The participants could precisely target the element on the screen before triggering the selection with a voice command. However, this advantage was equalized by the incorporation of gaze for the gesture selection. In contrast, speech selections did not profit from the incorporation of gaze, because participants did hardly look at the target element in the moment of selection. Instead, a common behavior pattern was to target the element, look back on the driving scene and then say the speech command. This step-by-step procedure gave the participants more control over targeting and selection, however it also resulted in a large increase of task completion times and greater subjective demands. A possible explanation is the lack of immediate feedback for speech commands. Participants pointed at the element and uttered the command and then held their hand still until they received feedback. However, it takes a certain time for the user to speak and for the system to process a command before feedback can be given. During this duration, even if it is very short, participants are left in a short moment on uncertainty which causes them to uphold the pointing movement for a longer time and make control glances on the screen, leading to a higher subjective demand.

Finally, we can summarize that the benefit of gaze-added pointing is greater when the initial gesture accuracy is low. This can be due to a variety of reasons, such as target element position or the trigger modality, but also inexperienced users (Roider & Gross, 2018). Fitts describes the difficult to target a requested element as a function of the size and the distance of the target item (Fitts & Peterson, 1964). In this regard, the algorithm is likely to provide greater improvements when selectable elements are more difficult to target with pointing gestures, e.g., when they are smaller, or if the position of the display in future vehicle concepts moves further away from the user. Speech-triggered selections give the users more control over the timing of the selections and therefore allow a higher gestures accuracy, but they do not profit from adding gaze information. At the same time speech-triggered selections lead to longer interaction times and greater subjective demands.

Limitations

Inaccuracies of the used gestures sensor might limit the pointing performance, a problem that also exists in current vehicles on the market that use gestures recognition. Even more, additional gaze data could help to reduce drawback from gestures sensor noise, or inaccurate pointing. Although eye-tracking technology faces similar accuracy problems, especially in the automotive domain, this experiment indicates how the fusion of both input channels can result in an increased selection accuracy. Still, this is only a first step on how to fuse gestures and gaze information while driving. The presented algorithm is based on simple heuristic rules. This aimed to support gesture selection by integrating gaze information in a clear a comprehensible way. The results indicate that this approach works and leads to a significant improvement of selection accuracy, but they also reveal limitations of the algorithm.

Future work should focus on the development of a more elaborate fusion algorithm, in order to integrate gaze information based on a probabilistic model. Such approaches have been presented for the incorporation of vehicle data to optimize pointing and touch performance while driving (Ahmad et al., 2016; European Statement of Principles (ESoP), 2008; Mayer, Le, et al., 2018). The results are derived from a relatively specific setup, namely four large elements on a large screen. The driving simulator did not support any movements, which is likely to further worsen pointing performance (Ahmad et al., 2015). Therefore, further development and evaluation will be needed to investigate the generalizability of the results for a greater variety of tasks, setups, and more realistic driving scenarios.

Summary

This experiment examined a prototype that integrates gaze input to improve the accuracy of pointing gestures while driving. Gaze input is integrated in a passive way. Participants did not consciously use gaze input, but we only took naturally occurring eye-gaze data into account. The results show that the benefit of this approach is depending on the drivers' initial gesture accuracy, which is influenced by various factors, such as the experience with gesture interaction, or the size and position of target elements. The benefit of the presented approach grows with the difficulty to make an accurate gesture pointing selection. This led to major improvements for those elements that are more difficult to select. On the other hand, there was also a lack of support, or even a decline of accuracy, for those people that made very accurate pointing gestures. Despite these limitations, this prototype shows that the gaze-added pointing approach can lead to an increase of selection accuracy, without posing additional demands on the driver.

- Gaze-added pointing can lead to an increase of selection accuracy.
- The benefit of gaze-added pointing is depending on the drivers' initial gesture accuracy.
- The approach does not pose additional demands on the driver, as drivers not consciously control their gaze.
- Speech-based triggers for the pointing gesture did not increase the fused selection accuracy and instead led to longer interaction times and greater subjective demands.

6.4. Summary

This chapter presented three interaction techniques that combine speech, gesture, and gaze input to overcome weaknesses of individual modalities. In Section 6.1, we explored the combination of gaze input with gestures, speech, and a haptic button. The aim was to provide a natural and more efficient alternative to the dwell time approach for confirming gaze selections. However, it showed that both speech and gestures confirmation are not well suited to confirm gaze input while driving due to modality specific issues. Overall, we observed that none of the selection techniques could compensate the high visual distraction and high cognitive load of explicit gaze input to make it feasible input modality while driving. Based on this, we discarded the use of gaze input as an active pointing modality. Instead, we concentrated on the passive integration of gaze information to support other input modes. We present a prototype that uses gaze information to activate the speech recognition system in Section 6.2. The evaluation in a user experiment showed that this approach can boost efficiency and user experience of speech interaction while driving. Since these benefits are limited to display-related tasks, we conclude that gaze-based speech activation cannot replace existing activation techniques, such as an activation keyword, but provide a valuable alternative. Another use of passive gaze information is demonstrated in Section 6.3. The prototype combines pointing gestures with passive gaze input based on heuristic rules. The results of the user experiments showed that this approach can significantly increase the accuracy of gesture selections, while posing no additional demands on the driver.

7. Design Patters for Multimodal In-Vehicle Interaction

The previous three chapters presented at total number of seven user experiments, which investigated the potentials of flexible multimodal interaction, showed ways to support input with alternative input modalities, and explored interaction techniques that combine natural input modalities. As a next step, the knowledge gained in these experiments will be summarized and structured so that other researchers and developers who are facing similar challenges can make use of it. In this regard, Chapter 3 already discussed different forms of how to provide design support and pointed out design patterns as a well-established form for structuring insights from user experiments in HCI. In the following chapter, we propose the first collection of design patterns for multimodal in-vehicle interaction. It focuses on the application and combination of natural input modalities, such as speech, gestures and gaze while driving. The pattern collection contains 15 interaction patterns, which address the specific requirements of the automotive user interfaces. The patterns are mostly derived from the presented user experiments, but also respect interaction design patterns from non-automotive HCI literature that have been applied to the automotive domain.

7.1. Structure and Organization

Design patterns have two major advantages over simple guidelines. First, design patterns are written in a well-defined form. Second, design patterns do not exist individually, but they are usually part of a larger pattern collection or pattern language that aggregates multiple patterns in a consistent way. The following sections describe the form and organization of the pattern collection in this thesis.

7.1.1. Pattern Form

The pattern form used in this pattern collection bases on the typical content elements of the *minimum HCI pattern form* described in Section 3.3.2 (Kruschitz & Hitz, 2010a). A short *name* gives a first hint on the purpose of the pattern and thus makes it easier to remember its content. The *context* describes the situations in which the pattern can be applied. The *problem* is formulated as a question and addresses the main problem that the pattern solves. The *forces* section describes forces and requirements that determine the problem and that have to be resolved by the solution. The *solution* gives a brief description of the proposed solution. This is added by the *consequences* element, which describes how the forces are addressed by the solution, but also points out new potential problems. The *rationale* content element grounds the pattern based on literature and findings from the experiments. The *examples* element describes concrete applications of this pattern. Since it is often difficult to find concrete examples that represent the solution of the pattern (Mirnig et al., 2016), the examples include not only references to existing automotive products, but also to other domains, concepts in literature, and the prototypes in this thesis. Finally, the *related patterns* element defines the relation of the pattern with other patterns in this collection. Table 7.1 summarizes the pattern form used in this thesis.

Name	Name of the pattern indicating its content
Context	The context in which the pattern can be applied
Problem	The problem formulated as a question

Forces	Forces and requirements that determine the problem and influence its solution
Solution	The solution of the problem
Consequences	A description of how the solution resolves the forces but also of resulting new ones
Rationale	Explanation why the solution works based on literature and experiments
Examples	Concrete examples in which the pattern is practically applied
Related Patterns	Related patterns within this pattern collection

Table 7.1. *Pattern form for multimodal in-vehicle applications.*

7.1.2. Pattern Organization

The general organization of patterns is derived from the pattern collection for multimodal interaction published by Andreas Ratzka (Ratzka, 2013). It is described in Section 3.3.5. Patterns within the collection are grouped by the benefits of multimodal interaction they address. The patterns aim at an increased flexibility, efficiency, and robustness of the interaction. Although those factors are often closely interconnected, we assigned each pattern to one group according to its primary benefit. The resulting pattern groups are:

- *patterns for increased efficiency*
- *patterns for greater robustness*
- *patterns for enhanced flexibility*

Furthermore, the patterns are organized in different layers of abstraction that go from high-level patterns to low-level patterns. High-level patterns describe the general interplay of multiple modalities, independent of concrete tasks, contexts, or modalities. Medium-level patterns are focused on a more specific problem context, but not bound to specific input modalities. Low-level patterns describe the most concrete problems and solutions that relate to the use of specific input modalities.

Finally, in the presented pattern collection we use *specializations (“is-a”)* and *associations (“related-to”)* to describe relationships between patterns. Please refer to Section 3.3.3 for a more detailed information about relationships between patterns.

7.2. Pattern Overview

After describing the form and organization of patterns in the previous section, this section presents a compact overview of the pattern collection for multimodal in-vehicle interaction. Figure 7.1 provides a pattern map that illustrates the organizational structure of all patterns and shows relationships between them. It also includes patterns from literature, which have been adopted to the automotive domain and connected to the novel patterns in of this work. Table 7.2 summarizes the contents of the presented patterns by giving a brief description of the problem and the proposed solution followed by a detailed description of each pattern in the following sections.

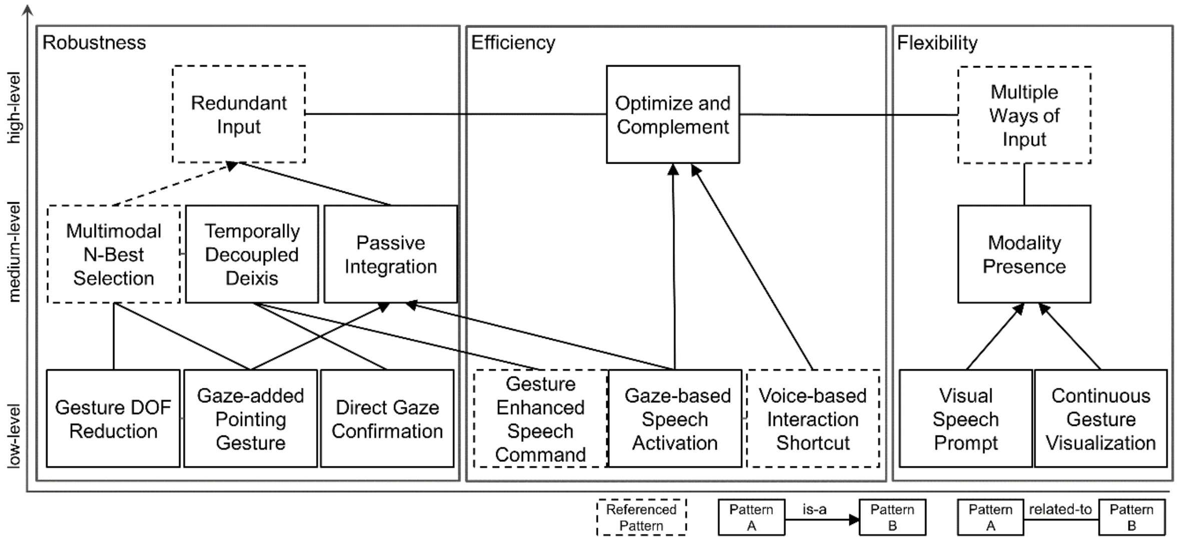


Figure 7.1: A Pattern Map for Multimodal In-Vehicle Interaction Patterns.

#	Name	Problem	Solution
1	Optimize and Complement	How can the driver interact with the vehicle in safe and efficient manner for a variety of secondary tasks?	Optimize for a speech-based system. Incorporate additional modalities to support speech, or to complement it, if spoken input is not feasible.
2	Gesture-Enhanced Speech command	How to enable the driver to input composed commands with parameters of different data types?	Enable the user to combine speech input with pointing gestures to specify locations or interactive objects.
3	Gaze-based Speech Activation	How can speech interaction with the system be made more efficient to use?	Allow gaze information to determine the driver's intention to interact with the system and activate the speech recognition system automatically in addition to explicit activation techniques.
4	Voice-based Interaction Shortcut	Which interaction style allows the user to quickly select the desired item without having to perform tedious navigation actions?	Support speech input to speed up the interaction and enable users to select items by naming them. This should not replace lists and menus but coexist with them to support flexible input.
5	Redundant Input	How to assure input when communication channels are distorted in an unforeseeable way?	Combine several interaction channels to make use of redundancy. Input coming from several channels should be interpreted in combination to reduce liability to errors.
6	Multimodal N-best Selection	Spoken input phrases may result in a set of several recognition hypotheses	Provide the user a means of selecting the correct result from a set of

		returned by the recognizer. The user must specify the correct result.	recognition hypotheses via pointing or key presses.
7	Temporally Decoupled Deixis	How can the driver be enabled to make effective use of deictic input to complement speech input?	Enable the driver to provide deictic and speech information a temporally decoupled manner.
8	Passive Integration	How to enable the driver to profit from multiple interaction channels without increasing the driver's workload?	Use redundant modalities in a blended style so that drivers use only one active input mode at a time.
9	Gesture Dimension Reduction	How can drivers effectively use pointing gestures for the selection of on-screen items?	Reduce the complexity of pointing gestures by using them for one-dimensional selections from small sets of horizontally arranged items.
10	Gaze-added Pointing Gesture	How to enable the driver to make quick and accurate pointing selections without creating additional workload?	Combine pointing gestures with passive gaze information. Do not visualize gaze input to the driver while pointing.
11	Direct Gaze Confirmation	How to enable the driver to confirm a gaze pointing target?	Use an input modality that allows to mark an exact point in time, allows immediate feedback, and can be operated blindly, such as a haptic push-button.
12	Multiple Ways of Input	How can input modalities be adapted to the context of use without burdening the user with additional configuration tasks?	Enable the user to trigger system functions by using one of several alternative interaction modalities, be it speech, typing or pointing.
13	Modality Presence	How can a better awareness of the existence and availability of alternative input modalities be promoted?	Create a visual representation for alternative input modalities. Use it to indicate whenever the according input modality is available for interaction.
14	Visual Speech Prompt	How can the driver be influenced to make greater use of speech input without restricting his freedom to choose a modality himself?	Use visual cues that explicitly prompt the driver to use speech and makes it easier for the driver to find the correct wording for the command.
15	Continuous Gesture Visualization	How can the system provide adequate feedback to support the driver during gesture interaction?	Visualize the recognition status of the driver's hands when gesture input is available and the progress of the gesture during execution. Use interior lighting to create ambient, visual gesture interaction.

Table 7.2. *This table summarizes interaction patterns for multimodal in-vehicle interaction. The collection also incorporates existing patterns from HCI and puts them in relation with each other.*

7.3. Patterns for Increased Interaction Efficiency

This group contains four patterns that can be applied to increase the efficiency of multimodal in-vehicle interaction. Two of these patterns are based on related patterns in literature: *Voice-based Interaction Shortcut*, and *Gesture-Enhanced Speech Command* (Ratzka, 2013). They have been adapted to the meet the specific requirements of automotive interaction. On top of this, we present two novel patterns that describe concrete modality-specific solutions. While these patterns mainly contribute to interaction efficiency, they might contribute to other potentials of multimodal interaction such as robustness and flexibility.

- *Optimize and Complement* uses one primary input modality (i.e., speech input while driving) combined with additional modalities to overcome its weaknesses, or to serve as complementary or alternative input if the primary input modality is not an option.
- *Gesture-enhanced Speech Command* describes the combination of speech input with gestures or pointing actions to enhance descriptive input with spatial information.
- *Gaze-based Activation* combines speech input with gaze information to determine the driver's intention to issue a speech command to the vehicle. Thereby, drivers do not have to explicitly activate the voice recognition system with a keyword, resulting in increased speech interaction efficiency.
- *Voice-based Interaction Shortcut* uses speech input to selection items from a large set or to directly jump to specific points in the menu hierarchy to avoid navigation overhead.

7.3.1. Optimize and Complement

Context

Drivers find themselves in a dual-task situation. They have to maneuver the vehicle safely and at the same time they want to interact with secondary functions. Current IVISs provide several alternative and largely independent paths to interact with the system, such as touch and speech input. However, many modern vehicles focus on touch-based interaction, where speech input is considered as a widely isolated, alternative path of interaction.

Problem

How can the driver interact with the vehicle in safe and efficient manner for a variety of secondary tasks?

Forces

- The primary task of the drivers is driving, which is highly visually and manually demanding task. Visual and manual interference should be avoided to enable safe interaction.
- High cognitive workload must be avoided. Besides visual and manual distraction, drivers can be cognitively distracted, which may result in decreased driving performance, e.g., due to inattentional blindness.
- Speech input allows hands-free and often eyes-free interaction, but it may be limited in certain situations (e.g., personal, social, or environmental aspects).

Solution

Optimize in-vehicle interaction system for speech interaction, but also incorporate additional modalities to support speech interaction or to serve as alternatives when speech input is not feasible e.g., due to task requirements, or social contexts. Accept reduced flexibility for the sake of an optimized interaction with one modality. Not every interaction step must be achievable with every available input modality.

Consequences

- Speech input enables drivers to keep their hands on the wheel and eyes on the road for a majority of tasks.
- If speech input is not feasible for a specific task or subtask, drivers can use alternative modalities that allow to complete the task in an efficient and convenient way.
- Drivers switch between speech input and complementary modalities depending on the task and context.

Rationale

Sections 4.1, 4.2., and 5.1 demonstrate that speech input requires least cognitive demand and resulted in the lowest visual distraction for typical in-vehicle tasks. Environmental influences can affect the suitability of speech input (see Section 4.2); however, it is often still preferred over alternative input modes. Furthermore, Section 5.1 demonstrates that the usage of speech input can be effectively promoted by using visual cues in the UI, and Section 5.3 shows that efficiency and user experience of speech input can be increased by providing alternative activation techniques. Still, social aspects may prevent drivers from speaking or certain task types may put a heavy load on human working memory (Bradford & H., 1995). In these cases, drivers can switch to another modality (e.g., touch) without a significant loss of efficiency (see Section 4.1).

Examples

More and more automotive manufacturers lay a strong focus on speech-based interaction in the car. Many functions can be directly controlled using natural speech input. Moreover, many functions allow a combination of speech input with touch. For example, the IVISs *Mercedes MBUX* and *BMW iDrive 7* allow drivers to use speech input for complex queries, such as finding the nearest Italian restaurant within five kilometers, and then switch to touch input to select one of the results. It is simply easier and faster to select the restaurant with a tap on the screen, instead of speaking the name. Additionally, in this example, it might be hard for drivers to pronounce an Italian name correctly.

The *Amazon Echo Show* is a speech-based interactive speaker. It features an additional touchscreen that can be used to display additional information and to interact with the device via touch. Still, as an evolution of a voice-only device, the echo's primary input modality is speech input, which is enhanced by touch input for certain tasks. Although not directly related to driving, the Echo Show serves as good example here. Similar to the automotive domain, users are enabled to interact with the devices in multitasking scenarios, in which their hands and eyes are busy.

Related Patterns

This pattern is related to *Multiple Ways of Input*. Although in-vehicle interaction should be optimized for speech input, complementary input modes must exist in case speech input cannot be used. It is also related to *Redundant Input*. Redundant information from other modalities can be used to optimize speech input for a wider range of tasks and environments, such as the incorporation of facial expression to enhance the recognition in loud environments (Sezgin et al., 2009).

Two patterns in the collection are specifications of this pattern. *Voice-based Interaction Shortcut* allows to directly select functions without navigating through a menu, *Gaze-based Speech Activation* helps to optimize speech interaction by providing a very efficient activation technique, and *Multimodal-n-best Selection* allows to switch to another modality (e.g., pointing) to disambiguate results from speech input.

7.3.2. Gesture-Enhanced Speech Command

Context

Some use cases require the driver to input several parameters of different data types such as spatial and descriptive information. For example, for getting the opening hours of a shop next to the car, drivers need to create a spatial reference to objects outside of the vehicle (i.e., the shop) and combine this with descriptive request for information (i.e., getting the opening hours).

Problem

How to enable the driver to input composed commands with parameters of different data types?

Forces

- Spatial descriptions are inefficient using only verbal descriptions with speech (see Section 4.1).
- Drivers often cannot correctly name objects they want to interact with, or there is a great variance of how objects are called (Pfleging et al., 2012)
- Pointing gestures are well suited to determine spatial relations, but not to determine functional parameters due to a lack of semantic information.
- Unlike virtual objects on a GUI, referenced objects in the real world (e.g., a shop outside the car) cannot provide any representation of selectable actions.

Solution

Enable the user to combine speech input with pointing gestures. The pointing gesture allows to create spatial references to all kinds of objects, whereas speech input can be used in parallel to describe functional parameters.

Consequences

- Pointing gestures allow drivers to create spatial references in a simple and an efficient way.
- Functional descriptions can be efficiently given using speech input, without the need for a GUI. Drivers can simply utter the action or functions they want to perform on the referenced location.

- Pointing gestures and speech input can be performed in parallel as they make use of different communication channels, which can contribute to a greater efficiency and higher naturalness of interaction.
- Free hand pointing gestures have to deal with a limited accuracy (see Section 5.2). This might result in erroneous selections and the need for error correction.

Rationale

Users prefer pointing devices for inputting spatial information, and speech input for descriptive data (Grasso, Ebert, & Finin, 1998). One particular benefit of combining pointing with speech input is a significantly improved efficiency over speech only approaches (Sharon Oviatt et al., 1997). It has also been shown that pointing gestures are feasible to interact with distant objects while driving (Rümelin et al., 2013).

Examples

There are several research publications presenting interaction techniques that make use of this pattern outside the automotive domain. A popular example is Richard Bolt's *"Put-that-there"*, which combines voice and pointing gestures to control a large interactive display surface (Bolt, 1980). So far, this approach has not been directly applied to interaction with the car, but there are similar approaches that serve as examples. Researchers demonstrated the use of pointing towards targets inside (Ahmad et al., 2016), as well as outside of the car (Rümelin et al., 2013) as a feasible approach. The *SIAM* prototype uses gestures on the steering wheel to provide a reference to front windows, or outside mirrors and combines them with speech input to determine the function (Mitrevska et al., 2015).

Related Patterns

This pattern is a specification of *Optimize and Complement.* It describes how speech input can be complemented with the pointing input to increase efficiency. Moreover, it is related to the pattern *Temporally Decoupled Deixis*, which allows drivers to temporally separate the pointing gestures from the speech command, in order to exploit the benefits of both input modes.

7.3.3. Gaze-based Speech Activation

Context

The driver wants to select and element that is currently displayed on an in-vehicle display via speech input. The speech recognition system is currently not active and has to be activated before the driver can utter a speech command.

Problem

How can speech interaction with the system be made more efficient and more natural to use?

Forces

- Drivers must explicitly activate the speech recognition system via a push-to-talk button (not natural), or a keyword (not efficient) (see Section 5.4).
- The activation of the system via keyword takes time and limits the efficiency.
- Always-on speech recognition systems do not require a manual activation of the system via button or keyword but react to voice input whenever the system can interpret it. This

approach has the risk of leading to many false positives, especially regarding the growing capabilities of voice-based agents to interpret any voice request.

- The push-to-talk button is typically placed on the steering wheel. Higher automation levels might not continuously have a steering wheel available.

Solution

Enable the driver to activate the speech recognition system based on his gaze behavior. Use gaze information to determine the driver's intention to interact with the system. Provide a virtual or physical representation of the speech system that allows the driver to establish "eye contact". Make the state of the system easy perceivable in an ambient way so that the driver does not have to do an explicit control glance to check whether the system was activated. At the same time, the feedback must be unobtrusive so that control glances do not result in annoying feedback. Thus, use acoustic feedback for the activation carefully. It can help to make the system state transparent to the user at the risk of being annoying.

Consequences

- Speech commands for display-related tasks can be used without explicit activation.
- Speech input gets more suitable for short commands.
- The total duration of speech interaction is decreased.
- Visual distraction for non-display-related tasks can be increased.
- For display-related tasks, the visual distraction is not increased compared to push-to-talk, or keyboard activation.

Rationale

Gaze tracking has been shown to be a natural and effective means to disambiguate the target of verbal conversation. In human conversations, gaze is a good indicator of people's focus of attention (Vertegaal et al., 2001). This is also valid for interaction with technical devices, as people tend to look at the devices they want to interact with (Maglio et al., 2000). Similarly, it was observed that drivers also tend to look at the CID when using speech for address input (Reimer et al., 2013).

The task time of speech input is aggregated of four steps: (1) the driver activates the speech system, (2) the driver speaks a command, (3) the system waits for a short moment of silence to recognize that the voice command is complete, and (4) the system interprets the recognized phrase and returns a result. Step two can hardly be improved, since speech is sequential modality and always takes a certain amount of time to speak, even for very brief commands. Shortening step three could speed up the interaction, at the tradeoff of recognizing incomplete commands. Step four will get faster and more robust with the evolution of ASR and NLU systems. Consequently, the largest room for increasing efficiency can be made in the first step. Activation via keyword is the state-of-the art activation technique for speech-based system, but it reduces the efficiency of the interaction, especially for short commands.

Examples

Oh et al. demonstrated the feasibility of the look-to-talk approach in a conversational setting (Oh et al., 2002). The participants in this study collaborated with a real person and an additional virtual character via voice recognition. Whenever they wanted to address the virtual character, they had to look at it to activate the voice recognition system. The authors pronounce the importance of gaze tracking accuracy and speed to exploit the full potential of the approach. Compared to a conversational setting, the in-vehicle setting provides a much better controllable environment. The driver is seated a defined position so that eye trackers can be installed and optimized to provide accurate and fast eye tracking.

Section 5.4 describes a prototype that successfully applies gaze-based speech activation in an automotive context. Participants gazed at the CID to activate the voice recognition system and select items on the screen via speech. This resulted in a very efficient interaction style and did not increase gaze durations compared to activation via PTT or keyword.

Related Patterns

This pattern is one specification of *Passive Integration*. It shows how gaze information can be passively used in combination with speech input to increase the efficiency of speech-based interaction. Therefore, is also a specification of *Optimize and Complement*, as it provides a specific solution how to optimize speech by using additional input.

7.3.4. Voice-based Interaction Shortcut

Context

The driver wants to select an item out of a large set of elements. This could be a contact name in a list of all contacts, or a specific song within a long playlist, but also one specific function that may be nested within a more complex menu hierarchy. Especially those functions and settings that are not frequently used are often hidden deep in the menu structure, and thus require several interaction steps to get there.

Problem

Which interaction style enables the driver to quickly select the desired item without tedious navigation actions?

Forces

- Pointing selections are easy and efficient, but they are limited to elements that are visually displayed to the driver.
- The number of elements that can be displayed to the driver is limited by different factors, such as a high number of possible items, large sizing of items (to enable good touch, pointing accuracy), or a limited screen size.
- Scrolling requires the driver to repeatedly check the screen. While driving this is a potential visual distraction and cognitively demanding.
- The navigation through hierarchical menus is often the costliest step in a selection process. It often takes longer to get to a specific function, than to perform the desired manipulation (Müller et al., 2011).
- Some vehicles provide shortcut keys that can be arbitrarily assigned to vehicle functions. However, the number of shortcuts is limited by the number of available buttons and even with a small number of buttons it is often hard to remember the assignment of functions.
- Operating systems for personal computers and mobile devices typically enable users to directly search and select file without navigating to the location in the file system. E.g., Windows 10 allows users press the Windows-key and type the name of a file or an

application to directly open it. However, typing letters on a keyboard while driving is highly distracting while driving (Kujala, 2017; Tsimhoni et al., 2004).

Solution

Enable drivers to select desired functions or items by naming them via speech input. Allow alternative wordings, as it is not guaranteed that users will use the same expressions.

Consequences

- Drivers can quickly select items from lists or menu structures independent of the size of the list or depth of the menu structure.
- Menu navigations can be minimized to avoid screen clutter.
- In comparison to hardware shortcut buttons, the number of possible speech shortcuts is unlimited and easier to remember.
- The use of speech input shortens the interaction compared to touch- or controller-based menus. Moreover, the driver's hands can stay on the steering wheel and eyes on the road during the interaction.
- Drivers might not always be able to name specific menu items, for example when browsing through a list of unknown items.
- Some list elements may be difficult to name, especially when they are long or in a foreign language (e.g., song titles or street names). In this case, numbering list items can help to create an easy, alternative naming of the items.

Rationale

Direct speech input is more efficient than touch input for descriptive data (see Section 4.1), and thus it is preferred for selecting objects among large sets (Grasso et al., 1998).

Examples

The *Mercedes MBUX* (Daimler, 2019) or *BMW iDrive 7* (BMW, 2019) infotainment systems have a menu-based hierarchy, but they also enable the driver to change many functions with a one-shot voice command. Additionally, drivers can name any menu entry to directly jump into a specific point in the menu hierarchy. Similarly, the *SpeeT* concept uses speech shortcuts to reference a variety of in-vehicle functions before manipulating them with touch gestures (Pfleging et al., 2011).

Moreover, many people use their smartphones in car as a navigation system. Voice-based smartphone assistants, such as the *Google Assistant* or *Apple Siri* since they allow to control many functions with speech. Users can start any application with a voice command, without searching and selecting the app icon on the screen.

Related Patterns

This pattern describes how speech can be optimized in the context of *Optimize and Complement.* It is further related to *Gaze-based Speech Activation*. Activation of the speech system with the driver's gaze can further increase the efficiency of voice-based shortcuts.

7.4. Patterns for Greater Robustness

This group contains seven patterns to support robustness for natural multimodal in-vehicle interaction. The patterns *Redundant Input* and *Multimodal N-best Selection* are based on patterns in literature and adopted to the automotive domain (Ratzka, 2013). Moreover, we present five novel patterns that describe specific problems and solutions for the automotive domain:

- *Redundant Input* integrates several input channels that provide redundant information to overcome limitations from individual channels.
- *Multimodal N-best Selection* combines speech input with pointing input to disambiguate dialogs.
- *Temporally Decoupled Deixis* describes how to enable to the driver to use deictic information in combination with speech input by removing temporal dependencies and thus allows drivers to exploit the advantages of both modalities.
- *Passive Integration* specifies how multiple redundant input modalities can be used without creating workload on the driver. Redundant input modes should only be used in a blended style, meaning that the driver only uses one active input modality at a time, while additional modalities should only be passively integrated.
- *Gesture Dimension Reduction* provides a solution to prevent errors during pointing gestures. Specific in-vehicle limitations mainly affect the vertical control of pointing gestures. Focusing on pointing gestures in the horizontal dimension is therefore a simple solution that contributes to interaction robustness.
- *Gaze-added Pointing Gestures* addresses the problem of inaccurate pointing while driving by using the drivers gaze behavior as an additional passive input channel.
- *Direct Gaze Confirmation* addresses timing problems to confirm deictic gaze information, which has been shown to provoke erroneous selections with gesture or speech confirmations. The best solution is to combine gaze input with haptic input elements that can be operated blindly.

7.4.1. Redundant Input

Context

In-vehicle interaction has to work in a highly mobile context that involves a variety of changing conditions. For example, lighting condition change drastically during day and night drives, background noise rises with increases velocity, or when driving through a loud environment. As a result, individual input channels may be unpredictably distorted.

Problem

How can the system ensure robust driver input when error-prone channels are unpredictably distorted?

Forces

- All possible individual channels can be distorted to some degree. It is hardly possible to find one channel that is working well during all conditions, especially in the mobile context.

- Distortions do not only apply to individual interaction channels, but several channels can be distorted at the same time.
- Some input modalities are limited by inaccuracies or ambiguities in general, such as gaze input, or pointing gestures.

Solution

Combine several input modalities to make use of redundancy of information. It is important to make use of different channels, to avoid that situational influences have a negative impact on all integrated modalities.

Consequences

- The use of multiple input modalities increases the probability that the system will be able to recognize and interpret user input in the desired way.
- Although distortions may affect multiple channels, they rarely affect the same pieces of information from different channels. Fusion of information may allow to reconstruct the entire input by combining parts from both input modalities.
- Input modalities with limited accuracy can be combined to disambiguate recognition errors.

Rationale

Independent distortions rarely affect the same aspects of content (Ratzka, 2013). For example, audio-visual speech recognition is often used to illustrate this. The combination of acoustic and visual signals leads to better recognition performance in loud environment (Benoit, Martin, & Pelachaud, 2000). In this example, the distortion – loud environment noise – mainly affects the auditory channel, while the visual channel is unaffected.

Examples

Sezgin et. al demonstrate how speech and facial expressions can be combined to enable a high accuracy for driver intention recognition. Thereby it is possible to provide speech input in the presence of loud engine- and road noise (Sezgin et al., 2009).

Related Patterns

This pattern is related to *Optimize and Complement* as it allows to enhance an individual input modality (e.g., speech input) to make it applicable in a wider range of different contexts so that the UI can be optimized consistently.

7.4.2. Multimodal N-best Selection

Context

Speech interactions in the vehicle may be imprecise. One reason is the inherent ambiguity of speech recognition technologies. For example, when the driver wants to input an address into the vehicle's navigation system via speech input, the speech recognizer might return n-best similar sounding results, instead of only one result. The driver then has to select the desired result. Another reason is the imprecision of the drivers' requests. Since speech inputs are not restrained by the user interface, drivers might utter incomplete or imprecise queries. For example, a driver might ask for a charging station without further specifications. In such cases, the system presents a number of n-best recommendations for the driver to make a selection.

Problem

What is the best way for the driver to make follow-up selections after imprecise speech requests?

Forces

- Speech input allows to transmit a large information bandwidth at the cost of a certain interaction time. It is therefore suitable to make an initial request to the system. However, making follow-up selections out of n-best results is a much simpler form of interaction for which speech input might be oversized.
- A follow-up selection is an additional interaction step that slows down the interaction. It should be kept as short as possible.
- Similar sounding words in speech queries are one reason for the use of n-best results. Using speech input again to disambiguate one out these results might again return an n-best list. Moreover, results might be difficult to name, for example when asking for Chinese restaurants on the route.
- Providing (spoken) acoustic feedback for multiple possible options is inefficient and quickly gets annoying.

Solution

Allow the driver to select the desired result with a pointing modality, such as direct touch input or a pointing gesture. The number of possible results should be restricted to a manageable size so that they can be easily overlooked and selected with a sufficient accuracy and high efficiency.

Consequences

- Touch input or pointing gestures do not convey semantic information, but they are efficient solutions to select one element out of a small set.
- Drivers do not have to read and speak the names of the desired results but can simply point towards it.
- Switching to another modality avoids that repeated spoken input is ambiguous (again).
- This quick confirmation step can help to avoid costly error correction due to false recognition results, or false assumptions of the user’s preferences.
- Results of the n-best list should be presented on the visual channel to allow drivers to touch or point at items. Moreover, it is more efficient than reading out each item and causes less cognitive load, as drivers do not have to remember all possible options.

Rationale

The selection out of n-best results is a simple selection task. Our experiments show that pointing gestures outperform speech input regarding efficiency with only little visual distraction while driving (see Section 4.2). Our results also indicate that modality switches (to a more efficient input modality) that occur during a task switch do not lead to a loss of efficiency but increase the overall efficiency (see Section 4.1).

Examples

This pattern is applied in *BMW iDrive 7* and *Mercedes MBUX* systems. Drivers can use speech input for route guidance destination input. If the speech recognizer does not get a result with a sufficient confidence, the system returns a list of n-best results. Drivers can then switch the input modality and touch the desired item or use the central controller to select it. The same pattern is used when the driver asks for points-of-interest and further specification is needed.

Related Patterns

This pattern is a specification of *Redundant Input*, as both inputs, speech and pointing, aim to provide the same piece of information.

Moreover, is related to *Gaze-added pointing Gestures* and *Gesture Dimension Reduction*. Both patterns enable the driver to make fast follow-up selections using mid-air pointing gestures. However, they are only feasible for a very limited set of items, which can be achieved in combination with a previous speech command.

7.4.3. Temporally Decoupled Deixis

Context

The driver wants to get information about an object inside or outside of the vehicle, e.g., asking for opening hours of a restaurant he is passing by. Combination of modalities allows to provide a spatial reference to the restaurant using a deictic modality such as pointing gestures, or gaze in combination with speech input to specify and trigger the function (e.g., “When does this restaurant open?”).

Problem

How can the driver be enabled to make effective use of deictic input to complement speech input?

Forces

- The temporal relation between a speech utterance the deictic information is unclear (Wagner, Malisz, & Kopp, 2014). This has the following consequences:
    - Drivers uphold the pointing gesture, or their gaze for a long duration until their speech command has been processed.
    - False selections occur due to an incorrect timing of pointing and confirmation.
- Holding the pointing gestures leads to a longer time with only one hand on the steering wheel. This affects the driver in:
    - Increased visual and manual distraction (pointing by gaze or eye-hand coordination for gesture pointing)
    - Increased physical effort for holding the arm.
    - Reduced tolerance of pointing gestures. The tolerance is negatively correlated with hold duration (Rümelin et al., 2013)
- The interaction is hardly interruptible without losing deictic information (especially when using gaze). Instead, it should avoid uninterruptible sequences of manual/visual interaction (Alliance of Automobile Manufacturers (AAM), 2006)
- The driver has little control over the timing of the interaction. The system should not force the driver to make time-critical inputs (Alliance of Automobile Manufacturers (AAM), 2006).

Solution

Enable the driver to provide deictic and speech information in a temporally decoupled manner. In a first step, provide a fast means to specify a reference to spatial information individually, so that the hands-off the wheel and eyes off the road time is kept as short as possible. In a second step, drivers can use speech to specify the function for the referenced object.

Consequences

- The interaction can be split into two temporally decoupled steps, which makes it easier to interrupt the entire interaction.
- Drivers can pace the interaction themselves and are not forced to confirm within a specific time window.
- Pointing gestures and gaze deixis can be kept as short as possible, resulting in reduced physical effort and visual distraction.
- Speech is not used to confirm the moment when to incorporate deixis, but only to specify the function. This requires one additional step to confirm the deictic information.

Rationale

The combination of deictic information, especially pointing, in combination with speech is a well-known example in multimodal interaction (e.g. “Put-that-there” (Bolt, 1980), or *Gesture-Enhanced Speech Command* (Ratzka, 2013)). Pointing provides a fast and easy way to reference objects in the environment. Speech input allows to transmit verbal information, but it takes a certain amount of time to speak a command. By coupling both modalities in a tight manner, the advantage of quickly referencing objects by pointing is limited, because it is slowed down by the speech confirmation. Drivers tend to wait for this feedback before they consider the pointing action completed (see Sections 5.2 and 6.3). On the other hand, the benefit of eyes-free speech input is limited since eye-hand coordination is needed to provide correct deictic information while speaking. The decoupling of both modalities allows to exploit the potential efficiency of pointing, as well as hands-free and eyes-free speech interaction.

Moreover, fast confirmation of deictic references avoids misinterpretation of temporally sensitive deictic input (see Section 6.1). Some experiments in literature suggest that stressed syllables are synchronized with deictic pointing gestures, however the correct temporal interplay between speech and deictic gestures is far from clear (Wagner et al., 2014). Deictic terms (such as *“this”*) are frequently not spoken at all. Often speech and deictic gestures input do not overlap at all (Sharon Oviatt, 1999). The more transparent the moment of confirmation is for the driver, the easier it is to time deixis and confirmation correctly.

Examples

Kern et al. presented a gaze-based interaction technique, which allows the driver to select an item by gaze in first step, while a different modality (e.g., a button or speech) can then be used to execute an action on the selected item. This temporal decoupling allows the driver to shift his visual attention back on the street before continuing with the interaction.

Related patterns

This pattern is related to *Multimodal N-best Selection*, which enables users to disambiguate the results of a previous speech query by using a pointing modality. While driving, gestures and gaze pointing for the selection of n-best results should be decoupled from speech input.

Furthermore, the pattern is related to *Direct Gaze Confirmation*. It describes how deictic gaze information can be effectively confirmed using direct input on the steering wheel.

7.4.4. Passive Integration

Context

Multimodal concepts in desktop settings allow users to make inputs using two or more active input modalities at the same time. Both inputs can be combined to disambiguate distorted input channels. For example, speech recognition quality could be enhanced by speaking a name and writing the letters at the same time (Müller et al., 2011). In the driving context, driving the car and controlling secondary functions already puts the driver in a multitasking situation.

Problem

How to enable the driver to profit from multiple interaction channels without increasing the driver's workload?

Forces

- Drivers are already in a dual task situation. Providing active input with two different modalities puts them in a triple task situation.
- Parallel input of two or more active modalities bears a higher risk for cognitive distraction.

Solution

Use parallel modalities in a blended style. Drivers should consciously use only one active interaction mode at a time. Additional input should be passively integrated, without the need for drivers to consciously control them.

Consequences

- Drivers consciously use only one input modality at a time.
- Cognitive workload is kept low compared to parallel input with two active modalities.
- Additional passive input contributes to the performance of the active input modality.

Rationale

Neuss concludes that parallel redundant input with different input modalities is difficult to handle for drivers in the car (Neuss, 2001). In contrast, passive input modalities refer to unintentional actions or behavior the driver that can be monitored by the system (e.g., facial expressions or lip movements). They are not actively controlled by the drivers (e.g. by issuing explicit commands) and thus no extra cognitive effort is necessary (Sharon Oviatt, 2012; Vilimek, Hempel, & Otto, 2007).

Examples

Ecker integrates gaze input as an implicit input modality to change the focus between the HUD and the CID. The driver uses controls on the steering wheel to interact with the display he is currently looking at (Ecker, 2013). Sezgin et al. shows how speech recognition accuracy can be improved by passively integrating facial expressions (Sezgin et al., 2009).

Related patterns

This is a specification of the pattern *Redundant Input,* which describes the fusion of input from several channels to reduce recognition and interpretation results (Ratzka, 2013). In comparison to that, *Passive Integration* describes the fusion only with passively integrated channels, in order to keep cognitive workload of the driver low.

7.4.5. Gesture Dimension Reduction

Context

Pointing gestures are conducted in the three-dimensional space. They can be used to specify horizontal and vertical coordinates on a display e.g., for selecting items. They have the potential to transmit multidimensional information, as pointing gestures are executed in mid-air and not restricted by a physical device. On the other side, this freedom also increases the risk for making errors.

Problem

How can drivers effectively use pointing gestures for selections on the screen while driving?

Forces

- While driving, people do not point accurately. Interaction in three dimensions increases the possibility for wrong selections.
- People have problems to find the gesture interaction area (May et al., 2014).
- A typical pointing method is eye-finger ray-casting (EFRC). The pointing error for this method mainly manifests in the y-axis (Mayer et al., 2015).
- Pointing gestures can be physically exhaustive for longer use. Vertical control requires the driver to move the arm up and down, which is more physically exhaustive than moving the arm left and right.
- Road conditions cause vehicle movements that can significantly impair the performance of pointing movements while driving, mainly on the y-axis (Ahmad et al., 2015; Mayer, Le, et al., 2018).
- Selection gestures cause additional movement of the hand that increases the risk of falsifying the pointing direction. Typical selection gestures, such as tapping in the air with the index finger (e.g., Microsoft HoloLens) especially affect the vertical axis.

Solution

Reduce the complexity of pointing gestures by using them for selections from small sets of items. Display selectable elements on the GUI in the horizontal dimension only. Allow movement in the vertical axis to give drivers more flexibility of movement or to support selection gestures.

Consequences

- Pointing input in the horizontal dimension only is more resistant to errors caused by inaccurate pointing and vehicle movements while driving.
- Horizontal placement of UI elements diminishes the problem of vertical deviation in typical pointing methods, such as EFRC.
- Considering only the horizontal pointing allows the driver to rest the elbow on the middle console, or on the armrest and use it as a pivot point. This reduces physical effort and creates additional stability.
- Established direct selection gestures, such as the air-tap can be better supported.

- The amount of selectable UI elements is reduced compared to a two-dimensional visualization.

Rationale

Many experiments have investigated human pointing performance in static environments (Chakraborty et al., 2012; Wong & Gutwin, 2010; Wu et al., 2011). There are only few experiments that investigate pointing specifically in the automotive domain (Brand et al., 2016; Rümelin et al., 2013), but it has been shown that pointing accuracy is affected when people are not allowed to move their eyes or head toward the target (Biguer et al., 1984). Additional studies suggest that pointing performance while driving is especially affected on the vertical axis (Ahmad et al., 2015; Mayer, Le, et al., 2018). However, in the context of multimodal in-vehicle interaction, pointing gesture input should only be used to make selections in small sets or disambiguate between few options (e.g., *Multimodal N-best Selection*). For any interaction that requires a greater amount of descriptive information, i.e. selection out of larger sets of elements, speech input should be used as suggested in *Voice-based Interaction Shortcut* (Ratzka, 2013). Furthermore, many large in car displays have a landscape orientation, which promotes a horizontal presentation of UI elements. Consequently, the drawback of a limited number of selectable elements on the horizontal is less critical. Instead, the reduction of interactive dimensions supports pointing gestures as a fast and direct selection modality.

Examples

Not many manufacturers that have integrated gesture interaction in their vehicles. The 2018 BMW 7 series and 2017 VW Golf support horizontal swipe gestures. The BMW uses left and right gestures with the outstretched thumb to skip through songs. Moreover, there is a pinch gesture that allows to rotate the parking camera view around the virtual vehicle, which is limited to rotations in the horizontal plane by moving the hand left and right.

Related Patterns

This pattern is related to *Multimodal N-best Selection*. N-Best results are best displayed horizontally in order make them easily selectable with pointing gestures. Combination with *Gaze-added Pointing Gestures* further increases the accuracy of selection.

7.4.6. Gaze-added Pointing Gestures

Context

Pointing gestures enable drivers to create deictic references to their environment, such as objects outside the vehicle, but also elements on a screen. Thus, they are a valuable complementary addition to speech input, which is less suited to create spatial references. A critical aspect for the benefit of pointing gestures is therefore to provide high accuracy and good efficiency for the selections.

Problem

How to enable the driver to make quick and accurate pointing selections without creating additional workload?

Forces

- Accurate pointing gestures require eye-hand coordination, but the driver's visual attention is primarily claimed by the driving task. The reduced availability of the visual resource affects the accuracy of pointing gestures (Biguer et al., 1984; Brand et al., 2016).
- Technical factors limit the accuracy for the detection of pointing directions. Small errors in the detection of a pointing vector can result in large deviations on the target depending on the distance of the selection pane.
- User-specific aspects limit pointing accuracy:
    - users use different pointing strategies, e.g., index-finger vector or head-fingertip vector. Pointing targets can greatly differ based on the used strategy.
    - users are not pointing very accurately, even when different pointing strategies are respected (Mayer, Schwind, et al., 2018).
- Providing visual feedback for pointing gestures (e.g. a cursor) helps to increase accuracy, but slows down the interaction and puts additional workload on the user (Mayer, Schwind, et al., 2018).

Solution

Combine pointing gestures with passive gaze information to determine the pointing target. Ensure functionality also when gaze information is not available. Do not provide a cursor in order to avoid additional workload and to prevent visual distraction of the driver's gaze from the target element.

Consequences

- The overall selections accuracy can be increased.
- Drivers do not know that their gaze is passively incorporated. Therefore, glance behavior is not influenced and no additional workload is created.
- The passive integration might also bear some risks, such as the risk of distorting correct selections and a less transparent system behavior.
- Drivers should be supported in directly referencing target objects. By leaving out a cursor, the driver is not tempted to control the cursor, but focuses directly on the object.
- The system must assess whether the user's gaze refers to the pointing selection (and fuse this information), or if gaze information does not refer to the interaction and must therefore be ignored.

Rationale

It has been shown that people anchor their gaze in the pointing target during mid-air pointing while driving (Ahmad & Langdon, 2018). This shows that gaze provides meaningful information about the pointing target in the moment of gesture selection. Furthermore, leaving out the cursor leads to shorter interaction times and ensures that the gaze is not distracted. Thus, for fast mid-air interaction a cursor should not be displayed (Mayer, Schwind, et al., 2018).

Examples

Section 6.3 presents an implementation that illustrates how gaze can be combined with pointing gestures in an automotive context. The drivers make selection out of four items on the CID. A gesture target element is given by the pointing direction of the index finger. In the moment of selection, the system determines the driver's gaze target element. Both elements are fused based on heuristic rules, which led to an overall increased accuracy without additional workload for the driver.

Related Patterns

This pattern is a specification of *Passive Integration* based on active mid-air gestures and passive gaze input. Furthermore, it can be used in combination with *Multimodal N-best selection* (Ratzka, 2013), because of the focus on fast and direct selection.

7.4.7. Direct Gaze Confirmation

Context

Drivers frequently have to localize interactive elements, such as virtual objects on screens (e.g., buttons), or objects in the vehicle interior, or the environment. Eye-tracking technology allows to capture the driver's gaze and to determine those targeted objects. The Midas-touch problem requires to provide an explicit confirmation for gaze-based input.

Problem

How to enable the driver to confirm a gaze pointing target?

Forces

- Gaze is well suited to derive deictic information, however it does not transport any semantic meaning. As a consequence, gaze-based confirmation techniques, such as gaze-gestures are artificial and inefficient (Huckauf & Urbina, 2011).
- Some selection techniques, such as blinking for confirmation are not applicable in the automotive domain. The driver should never be forced to close his eyes, especially since confirmation blinks would have to be quite long to distinguish them from naturally occurring blinks.
- Dwell times are an effective method to make selections in desktop settings. While driving, however, they promote long, uninterruptable glances away from the road and are therefore not feasible.
- Speech confirmation is simple and little distracting, but it lacks the possibility to give precisely timed input and immediate feedback to the driver. This results in longer glances on the target than necessary while drivers wait for feedback.
- Simple mid-air gestures allow to provide direct feedback. But the movement of the hand near the driver's line of sight can distract the gaze from its target. This problem is increased due to the unfamiliarity of most drivers with mid-air gestures, which leads to control glances to ensure that the gestures is performed correctly.

Solution

Use an input modality that allows to mark an exact point in time and to give immediate feedback. It must be operable blindly and not require or tempt drivers to make control glances to ensure correct execution. A simple haptic input element, such as a push-button can be a suitable solution. Place the button directly at the driver's hand position, so he does not have to reach for it.

Consequences

- Eyes-off the road time for gaze input is minimized.
- Placing the button on the steering wheel enables drivers to press it blindly. This way no eye-hand coordination is needed that might distract the driver's gaze.
- Drivers receive immediate feedback.
- Drivers have full control over the pace of the interaction. The interaction can be easily interrupted without loss of information.
- Pressing a button involves only a very small movement of the finger, thus it is hardly eye-catching.

Rationale

Section 6.1 compares four different selection techniques for gaze-based input in an automotive setting. Speech input results in long task completion times and long glance durations away from the primary task. Gesture confirmation led to many "empty" selections, because participants looked at their hands in the moment of selection. The button on the steering wheel resulted in the best primary task performance, least visual distraction and the shortest interaction times and was therefore clearly preferred by participants.

Examples

Kern et al. use a button on the steering wheel to confirm gaze input on an in-vehicle screen. They assess this interaction technique as a valid alternative to traditional touch input and also propose a combination with speech input (Kern et al., 2010). Similarly, Poitschke combine gaze input with a barrel key on the steering wheel. Drivers press the barrel key for confirmation of the gaze focus before changing values (Poitschke et al., 2011).

Related Patterns

This pattern can be used to enable the pattern *Temporally Decoupled Deixis*. A deictic gaze target can be efficiently selected with *Direct Gaze Confirmation*. This deictic information that then be combined with speech input to specify a function on the selected element.

7.5. Patterns for Enhanced Flexibility

The last pattern group focusses on supporting the potential of enhanced flexibility for in-vehicle multimodal interaction. It contains four patterns, one high-level, one mid-level and two modality-specific low-level patterns. The pattern Multiple Ways of Input is based on literature and adapted to the automotive domain (Ratzka, 2013).

- *Multiple Ways of Input* provides alternative input modes to the driver to cope with a variety of tasks, user preferences and environmental demands.
- *Modality Presence* enables drivers to make more flexible use of alternative input modalities, by creating a better understanding about the existence and availability of alternative input modes.
- *Visual Speech Prompt* uses on-screen notifications to promote the usage of speech input for suitable in-vehicle tasks. Drivers can be effectively influenced in their modality choice without posing restrictions to them.

- *Continuous Gesture Visualization* addresses the lack of presence and transparency for gesture input as an alternative input modality by providing continuous ambient feedback.

7.5.1. Multiple Ways of Input

Context

Drivers must be enabled to interact with a variety of vehicle functions in changing contexts. This is illustrated by the following example: a driver wants to input a navigation destination. He is alone in the car and standing on a parking lot. Another day, he wants to enter a different destination, but this time he is driving on the highway at 130 km/h together with two colleagues who are having a conversation.

Problem

How to provide a suitable interaction for each driver in a variety of tasks, user preferences and environmental demands?

Forces

- The choice of input modality is mainly influenced by the task type. Drivers favor specific input modalities for some task types, such as speech input for text entry (see Section 5.1).
- Driver preferences are highly individual. Some people might refuse to use speech input, while others appreciate speech as easy and safe input while driving (see Section 5.1).
- Situational demands can restrain the perceived suitability of input modalities although there might be no objective drawbacks (see Section 4.2).

Solution

Enable the driver to trigger in-vehicle functions by using one of multiple alternative input modalities. These alternatives should not be provided for the sake of complete equivalence, but only if there they can provide a benefit over the primary input mode in certain situations. In these cases, available alternatives should be active and ready to use without further configuration needs.

Consequences

- The system can be used for a greater variety of tasks, user preferences and environmental conditions.
- Drivers can use their preferred interaction style, which leads to higher user satisfaction.
- However, driver's preferred interaction style might not be the best choice regarding efficiency and distraction (e.g., text input via touch instead of speaking).
- Drivers can choose different modalities for single interaction steps of an interaction sequence.
- Drivers have to know which modalities are available. Especially novice users will require some hints or guidance. The system must provide effective help and prompting strategies that reveal alternative input modes.

Rationale

The environment of the automotive domain already poses specific demands to input modalities for driver-vehicle interaction. Furthermore, user characteristic, preferences, environment and situation have an impact (Sharon Oviatt et al., 2009). For those reasons, it is difficult to always provide the best input option using only one input mode. Offering multiple ways of input allows the driver to choose a suited input option for a greater variety of contexts. Moreover, multiple ways of input enable users to switch input modalities after recognition errors (Sharon Oviatt et al., 2009).

Examples

Many examples in research and industry make use of this pattern. The *Ford Model U Concept* enables drivers to switch between speech and touch input at any point in time of the interaction (Pieraccini et al., 2004). The *Sammie* system provides alternative in a similar way (Becker, Poller, et al., 2006). It provides equivalent input through speech and a central turn-push-controller. The pattern is also applied in many current production vehicles. For example, in the *Mercedes A-Class 2018*, drivers can complete each interaction step using direct touch input on the screen, a central touchpad in the middle console, or touch sensitive areas on the steering wheel. Additionally, suitable functions can be directly accessed using speech commands.

Related Patterns

This pattern is related to *Optimize and Complement*. It provides alternative modalities that can serve as complementary input to a primary modality. Moreover, *Modality Presence* is needed to create an awareness of these alternative modalities. This is especially important for those modalities that do not have a visual representation and are therefore more difficult to discover and explore.

7.5.2. Modality Presence

Context

Some secondary tasks enable the driver to choose from multiple available input modalities. Depending on the task, environment, or user preferences the suitability of these modalities can change. A flexible use of different input modes provides a major potential of multimodal in-vehicle interaction. However, there is usually one input mode that is most prominent (e.g., touch input in many vehicles). Alternative input modes can be valuable alternatives in certain situations, but they may not be available for all tasks. This leads to a lack of awareness when alternative input modalities are available.

Problem

How can the driver get a better awareness about the availability of alternative input modalities?

Forces

- Alternative input modalities are usually not available for all tasks.
- Natural input modalities (e.g., speech, gestures) are usually “invisible” in the interior. They do not have prominent visual representations, which are physically located in the interior, such as a touchscreen or a turn-push controller.
- Drivers have insufficient awareness of the availability of alternative input modes.

- Drivers do not make use of the whole interaction bandwidth which is offered by modern vehicles.

Solution

Create visibility for alternative input modalities with the help of visual representation in the vehicle. These representations can exist virtually on a display, but they can also be physical objects in the cockpit. They should indicate whenever the according input modality is available for interaction.

Consequences

- Drivers can easily see which modalities are available.
- Drivers are enabled to make use of the whole bandwidth of input modalities.
- Additional visual representations must be designed carefully to avoid visual distraction.

Rationale

The pattern *Multiple Ways of Input* suggests making each system function controllable by using one out of several alternative input modalities. This presupposes that drivers are aware of all interaction alternatives. This is especially important since it is often not feasible to offer all available input modalities for each interaction step. For example, it does not make sense to provide pointing gestures for text input, but they can be a valid alternative for simple selections. Consequently, the subset of available input modalities is changing depending on the system status and thus drivers might not always be aware of all currently available input modes. Providing representations for alternative input supports the driver in making flexible use of all modalities.

Examples

The *Nio ES8*[10] uses a small round, head-like, display that represents their in-car speech assistant called Nomi. Nomi represents a permanent physical representation of speech input. It makes the driver aware of the possibility to talk, just as a large touchscreen is a representation for touch input.

The *BMW 7 series* supports a rotation gesture to control the audio volume. When drivers turn the haptic volume knob, a small display overlay pops up that illustrates how to use the rotation gesture for this action. Thereby, users get a better awareness about when gesture actions are available.

Related Patterns

This pattern supports *Multiple Ways of Input* (Ratzka, 2013; Ratzka & Wolff, 2006), which aims to provide the flexibility to select the interaction modality that is most appropriate, by helping drivers to choose out of all available modalities. *Visual Speech Prompt* and *Continuous Gesture Visualization* are two specifications of this pattern.

[10] https://www.nio.io/de_DE/es8, accessed on 7.02.2019

7.5.3. Visual Speech Prompt

Context

Many modern vehicles have large screens to display information to the driver. They also allow drivers to interact with content on the display using different input modalities. Typical state-of-the-art input options are touch and speech input. While speech input has many advantages over touch in regards of distraction and workload, many drivers use touch input in situation where speech would provide a better alternative.

Problem

How can the system influence the driver to make greater use of speech input without restricting a free choice of modalities?

Forces

- Drivers tend to use the input modality with the strongest affordance in the vehicle interior and large in-vehicle screens provide a strong affordance for touch input.
- Drivers do not make use the benefit of flexibly choosing the input modality depending on the task and situation.
- User interfaces give the driver full flexibility in the choice of available modalities. They usually do not promote an individual modality.
- Promoting modalities might cause drivers to feel limited in their freedom to make their own decision.
- Lengthy textual prompts might lead to increased visual distraction that outweighs the benefit of using speech input

Solution

Create a visual representation for speech input whenever it is feasible and available. Display a text notification in combination with a prominent icon on the screen that explicitly prompts the driver to use speech and makes it easier for the driver to find the correct wording for the command. These visual cues should be easily perceivable by the driver, but it should not occlude other important UI elements.

Consequences

- The prompt depicts a visual representation of speech input.
- Explicit prompts increase the likeliness that drivers will decide for speech input.
- Visual prompts are less obtrusive than auditory ones.
- Drivers don’t feel disturbed by the prompts if they are designed in a subtle way.
- Long term effects of this solution might educate drivers to use speech input more often without paying attention to the prompt at all.

Rationale

Section 5.1 shows that explicit visual prompts lead to a significant increase of speech usage and thereby reduce the overall visual distraction compared to touch input. The prompts are designed as a notification overlay that does not restrict the use of touch input. Please refer to Section 5.1.2 for exemplary illustrations.

Examples

The Google search app for Android smartphones offers user the possibility to enter search queries via a touch keyboard or via speech input. Especially for longer queries or in the mobile context (e.g., while driving, or walking) speech input is the preferable choice. The app uses an explicit textual prompt to support the use of speech input ("Say 'Ok Google'…") in combination with a microphone icon.

Related Patterns

This pattern is a specification of *Modality Presence*. It describes how speech input can be supported by providing a visual representation.

7.5.4. Continuous Gesture Visualization

Context

Mid-air gesture input is a relatively novel way of driver-vehicle interaction. Many drivers don't have any experience with this form of interaction. Accordingly, they may struggle to execute gestures correctly. Moreover, it is often not clear which functions are controllable using gestures.

Problem

How can the system provide adequate feedback to support the driver in using gesture interaction?

Forces

- Gesture interaction is invisible in the vehicle cockpit, as there are no physical control elements. Drivers do not know if and when gesture input is available.
- Drivers find it difficult to know the gesture interaction space (May et al., 2014).
- Current system only provide feedback after a gesture was successfully executed.
- There is no transparency about reasons for failure, which can be based on recognition errors of the system, or due to wrong execution by the user.
- Gestures can be performed in larger three-dimensional area in the cockpit. Typical in-car displays are too small to provide feedback that reflects this spatial freedom.

Solution

Visualize the recognition status of the driver's hands when gesture input is available and the progress of the gesture during execution. Interior lighting can be used to create ambient, visual feedback for gesture interaction.

Consequences

- Continuous feedback supports users in learning and understanding the correct execution of gestures.
- Gesture functionality is made more transparent for the driver and thereby support the learning process.
- Ambient lighting can support the visualization of the entire gesture interaction space.
- Many vehicles already have integrated interior lighting that can be used to provide feedback.

- The additional feedback can lead to increased visual distraction and thus affect driving performance.

Rationale

Peripheral light feedback in the car using a light strip behind the steering wheel is a feasible feedback modality for gesture input and less distracting than visual feedback on a display (Shakeri et al., 2017). While acoustic feedback is generally least distracting it is not suited for providing continuous feedback during the execution of gestures without being annoying. Moreover, Section 5.2 presents a prototype that uses a similar interactive LED strip at the bottom of the windshield to give light feedback for confirm, swipe and pointing gestures. Participants reported that they felt significantly more supported by additional light feedback for gesture interaction compared to acoustic feedback alone.

Examples

The *Gesture Thermostat* (Freeman, Brewster, & Lantz, 2014) is a thermostat that can be controlled using hand gestures. The moment the user is recognized by the thermostat, it turns on white light and pulses gentle on a low brightness. This signals the user that the system is aware of the user's movement. As user are performing dynamic rotation gestures the thermostat gives continuous feedback. They highlight the potential of light feedback to increase display size. Additionally, they illustrate that light feedback can also be used to give feedforward, that supports the user to perform the gesture correctly.

The *VW Golf 7* uses left and right swipe gestures in front of the CID to enable the driver to move through radio stations. As soon as the driver's hand is close enough to the screen, it is in the interaction area. The CID then displays a small hand-icon and a blue glow at the bottom of the screen to indicate that the system is ready for gesture input.

Related Patterns

This pattern is a specification of *Modality Presence* for gesture input.

7.6. Summary

This chapter presented the first collection of design patterns for multimodal in-vehicle interaction. The collection contains both, patterns derived from the user experiments and prototypes in the previous chapters, as well as patterns that were derived from the field of multimodal interaction and adapted for in-car interaction. The structure and organization of the patterns is based on HCI patterns in literature. Each pattern is described based on nine content elements: pattern name, context, problem, forces, solution, consequences, rationale, examples, and related patterns. The patterns are organized in three categories that represent the main benefits of multimodal interaction: patterns for increased interaction efficiency, patterns for interaction robustness, and patterns for enhanced flexibility. Each pattern is assigned to one category but might also contribute to other benefits of multimodal interaction. Furthermore, the collection contains patterns of different levels of abstractions. For example, high level patterns describe general solutions that are independent of specific modalities, while low level patterns represent concrete solutions for certain modalities. The entire pattern collection is visualized as a graph that points out the levels of abstraction and the relationships between patterns.

8. Application and Evaluation of Patterns

The previous chapter presented the first pattern collection for multimodal in-vehicle interaction, which puts individual patterns into relation. This suggests that they can act together in a more complex multimodal system that incorporates multiple patterns. However, the design patterns have been derived from experiments and prototypes that each examined one very specific aspect of the interaction, such as the costs of modality switches or the activation of the in-vehicle speech system. The next step is therefore to investigate the application of multiple design patterns in one system. This chapter describes a multimodal in-vehicle interaction prototype, which serves as an example to demonstrate the implementation of the presented design patterns and allows an overall evaluation. The first section describes the prototype and how the proposed patterns from the previous chapter are applied based on a small set of featured use cases. The second section presents an evaluation of the prototype in a driving simulator experiment in comparison to a speech-only approach. We report the results and discuss the impact of the applied patterns.

8.1. Prototype

The software prototype is an interactive IVIS that demonstrates the integration of speech, gesture, gaze, and touch interaction using the presented design patterns from Chapter 7. It supports three use-cases from typical domains of current in-vehicle systems, such as navigation and communication.

8.1.1. Integration of patterns

The system follows a speech-first approach. It is optimized for speech interaction and every interaction can be completed using speech input, but it is complemented with additional gesture input for certain steps in the interaction (**Optimize and Complement**). Figure 8.1 shows the start screen of the prototype.

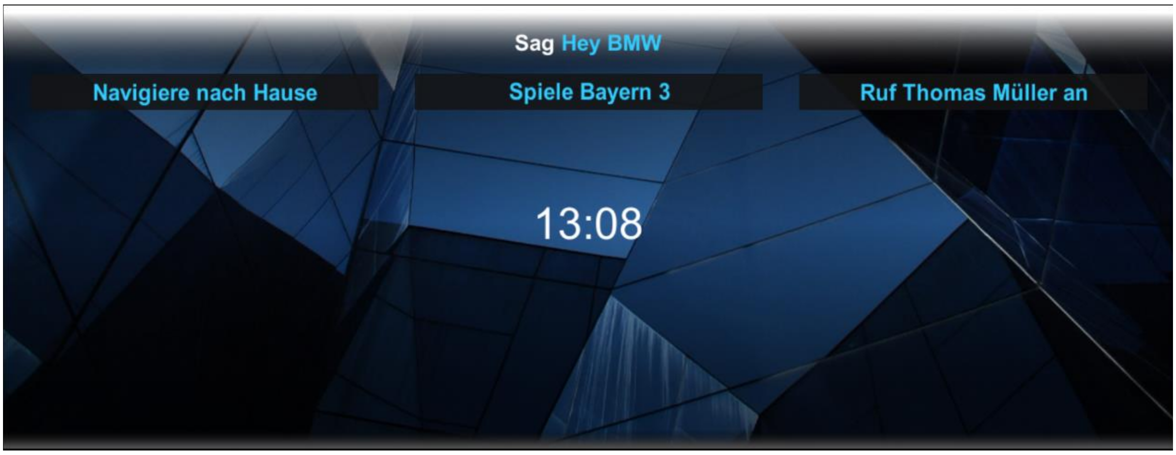


Figure 8.1: The start screen of the prototype provides only few information. A textual hint prompts driver to use speech input. Three function links provide formulation examples and allow direct function calls.

In comparison to typical IVIS it provides only very few elements. Speech input is supported in multiple ways. Drivers can directly call functions with their voice, such as starting a navigation, or calling one of their contacts. Thereby, they can directly jump to any function without any references on the screen (**Voice-based Interaction Shortcut)**.

The speech system is efficiently activated by drivers' gazes on the screen (**Gaze-based Speech Activation)**. Alternatively, they can activate the system by saying the keyword (**Multiple Ways of Input).** The status of the speech recognizer is permanently visualized at the top of the screen. An ambient white glow indicates that the speech system is inactive (as shown in Figure 8.1) and turns blue when the system is activated. Together with the textual suggestions, this creates a constant visual representation of the speech system on the screen (**Modality Presence**). Furthermore, drivers are explicitly prompted to use speech input. The start-screen of the system displays a textual message ("Say Hey BMW") that explicitly prompts the user to use speech input (**Visual Speech Prompts**). Additionally, there are three functional links below this speech prompt. They give an impression of which types of functions are supported by the system and give examples of possible formulations. These three links can also be selected with a pointing gesture and thereby provide a way to directly select the offered functions (**Multiple Ways of Input**). Due to the limited number of functions for this type of interaction, theses links should be provided in an intelligent way to provide the most relevant functions based on contextual knowledge of the IVISs, such as user preferences, time, and location.

Drivers' initial speech requests might result in several n-best results provided by the system. In this case, the prototype allows to switch input modalities and supports the complementary use of gestures for an efficient selection from small sets of elements on the screen (**Multimodal N-best selection**). N-best results have a limited number, and they are displayed in a horizontal layout to facilitate the selection with pointing gestures (**Gesture Dimension Reduction**). Alternatively, the driver can complete the interaction with speech input (**Multiple Ways of Input**). The driver's gaze is passively integrated to improve the accuracy of the pointing gestures to allow a fast and yet accurate selection without providing a cursor (**Passive Integration**, **Gaze-added Pointing Gesture**). In order to further support gesture interaction, the system provides simple visual feedback at the bottom line of the screen whenever the driver's hand is in the interaction area (**Continuous Gesture Visualization**).

8.1.2. Use Cases

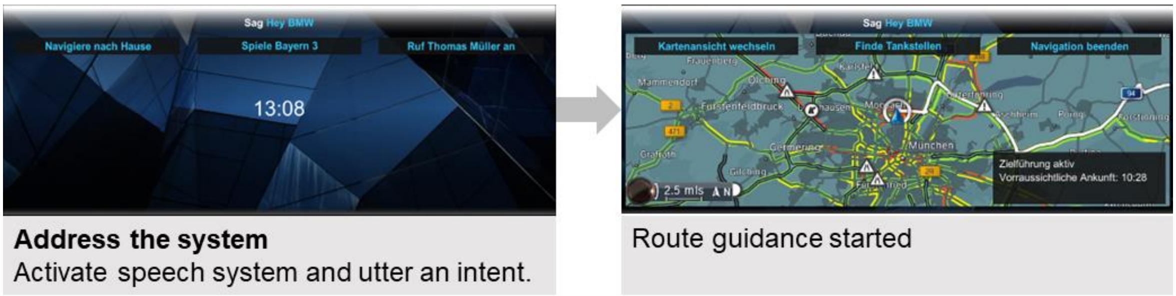
The prototype supports different navigation and communication use cases. They provide a context to demonstrate the implementation of patterns for different interaction steps and are not intended to reflect the full functionality of a system. The interaction always starts from the screen in Figure 8.1.

Use Case 1 (UC1): Direct Function Call

Drivers can perform a direct function call when they have all the necessary information to directly execute a system function. Typically, this can be the name or address of a target destination, the name of a radio station, and the complete name of a contact to call. Drivers can then make a direct function call that does not require any dialogue with the system. In this case, the interaction consists of only one step (see Figure 8.2):

1. Address system

The prototype supports three types of direct function calls: **starting a route guidance**, **playing a radio station**, or **calling a contact**. The driver can for example say, “Call David Schubert please” or “Bring me to Main Street 32” and the system will directly start a call or route guidance. Alternatively, if the desired function is offered in one of the three functional links, users can select it with a pointing gesture.



Address the system
Activate speech system and utter an intent.

Route guidance started

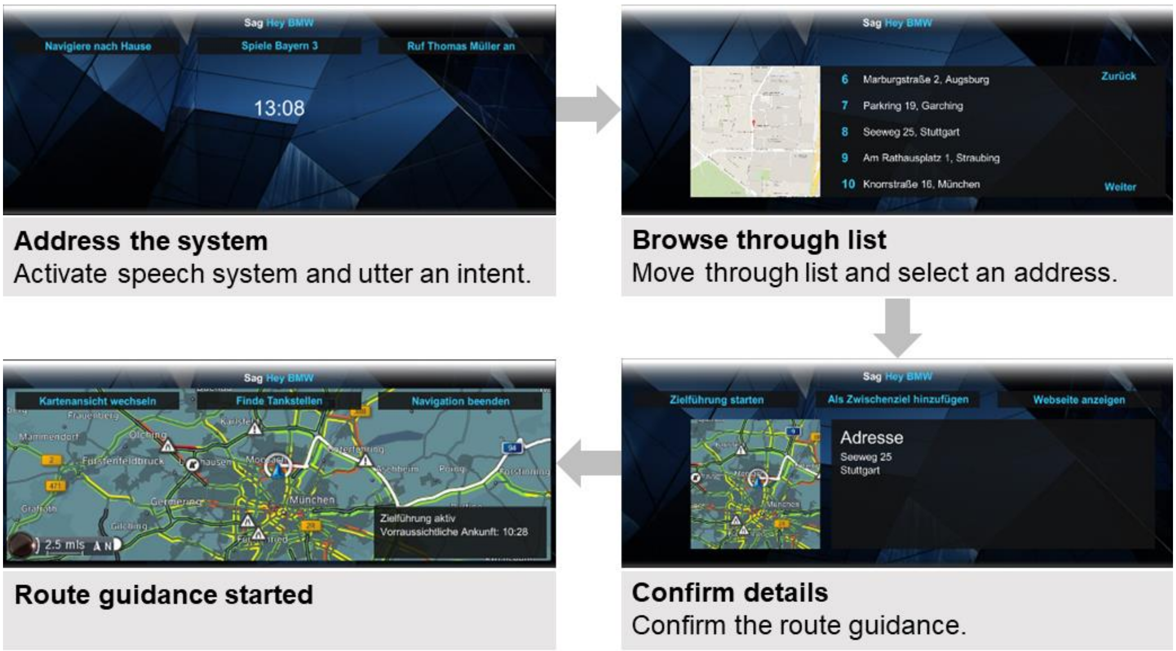
Figure 8.2: Interaction steps for direct function calls (example: starting a route guidance).

Use Case 2 (UC2): Browsing through list

There are other situations in which drivers may not be able to verbalize the concrete function they want to activate in a one-shot command. In these cases, users are usually required to browse for the element in a list, because they cannot directly search for it but need to recognize it when they see it. Examples for this case are actions in which part of the information is unknown, such as playing a song with a hard to remember title or starting a route guidance to an address from one of the last destinations. The interaction consists of four steps (see Figure 8.3):

1. Address system
2. Browse list
3. Confirm details

The prototype lets drivers browse through a **list of contacts** and a **list of recent destinations**. First, they ask for the according list (e.g., “Show all my contacts.” or “Show my last destinations.”), before browsing through the list, finding and selecting the desired element and confirming the call or the route guidance.



Address the system
Activate speech system and utter an intent.

6 Marburgstraße 2, Augsburg
7 Parkring 19, Garching
8 Seeweg 25, Stuttgart
9 Am Rathausplatz 1, Straubing
10 Knorrstraße 16, München

Browse through list
Move through list and select an address.

Adresse
Seeweg 25
Stuttgart

Confirm details
Confirm the route guidance.

Zielführung aktiv
Voraussichtliche Ankunft: 10:28

Route guidance started

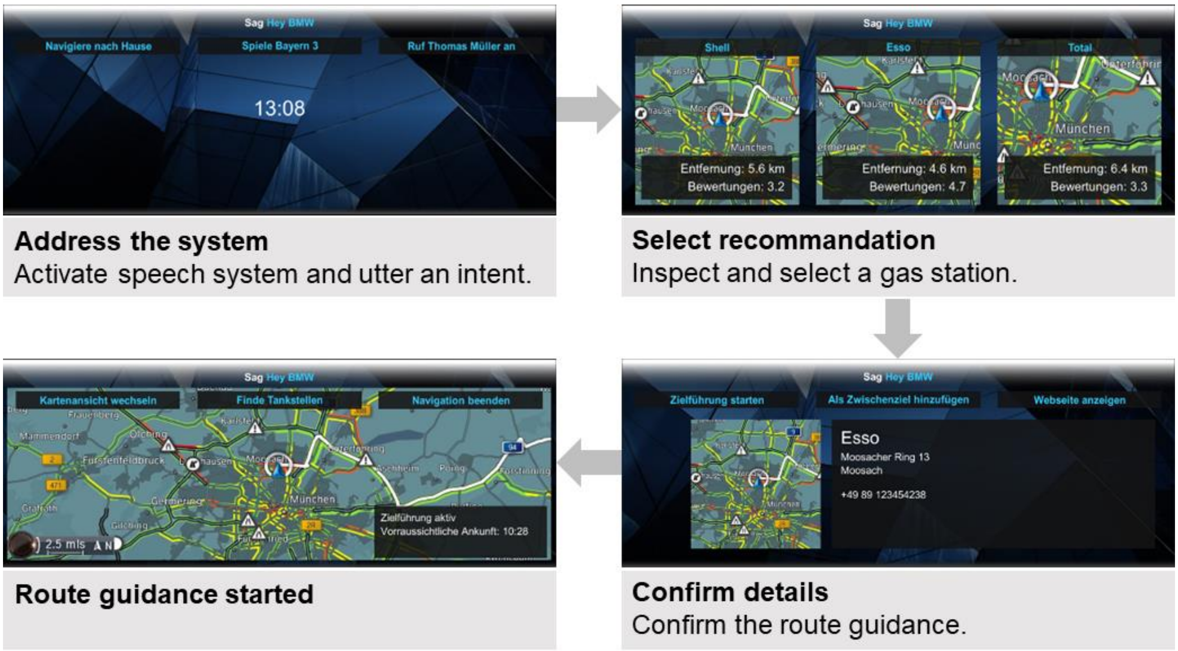
Figure 8.3: Interaction steps for browsing through unknown list (example: list of last destinations).

Use Case 3 (UC3): Asking for recommendations

Finally, drivers may not provide all the information needed to directly start a function. Although, they may not know the entire information, they can often provide some filtering criteria. The system then provides a small set of the n-best results from which the driver can select. For example, a driver might not remember the exact name of the restaurant he wants to navigate to, but he can ask for Italian restaurants nearby. Or the driver wants to call a contact by its first name although there are several contacts with the same first name. The interaction consists of three steps (see Figure 8.4):

1. Address system
2. Select recommendation
3. Confirm details

The prototype allows users to ask for **recommendation for restaurants** or **recommendation for gas stations** depending on the distance or the rating (e.g., “Show me nearby restaurants with a good rating”). The system will then present three recommendations. The driver selects one element and then confirms the start of a route guidance.



Address the system
Activate speech system and utter an intent.

Select recommandation
Inspect and select a gas station.

Confirm details
Confirm the route guidance.

Route guidance started

Figure 8.4: Interaction steps for asking for recommendations (example: gas stations nearby).

8.1.3. Tasks

Each of the use cases from the previous section is composed of multiple interaction steps. These can be summarized in four different tasks: *address system*, *browse list*, *select recommendation*, and *confirm details*. Each task can be completed using only speech input, but also offers an alternative input modality. All tasks and input alternatives are summarized in Table 8.1.

Task	Use Case	Available Input Modalities
Address system Activate speech system and utter an intent	**UC1, UC2, UC3**	• **Speech:** say the keyword followed by the intent. Example: “Hey BMW, start route guidance home”, or “Hey BMW, show me my last destinations”. • **Alternative Gaze:** glance at the screen and utter the intent. Example: [glance at screen] “start route guidance home”.
Browse list Move through a list and select the desired item	**UC2**	• **Speech**: move through the list page by page Example: “Next [Page]”, “Previous [Page]”. • **Alternative Touchpad**: scroll with two fingers to directly move the list as known from laptop touchpads. This only moves the list. A selection is made by saying the number of the list element.
Select recommendation	**UC3**	• **Speech**: say the name of the recommended restaurant/gas station Example: “Take me to the Shell please”.

Inspect and select one out of three suggestions		• **Alternative Pointing Gesture**: pointing movement with the index finger towards the recommended restaurant/gas station.
Confirm details Confirm the route guidance or call	**UC2, UC3**	• **Speech**: say the name of the function Example: "Start route guidance". • **Alternative Pointing Gesture**: pointing movement with the index finger towards the function

Table 8.1*: Speech input and alternative input options for each interaction step.*

Address System

The first interaction step for all use cases is to address the system, which is composed of activating the speech recognition system and uttering a speech command. As the prototype follows a speech first approach, drivers always use a speech command to utter their intents, but they can choose between two ways to activate the speech recognition system. The first option is to use a conventional keyword "Hey BMW". If there is a break after this keyword, the system answers with a short feedback ("Yes?") before drivers can speak their command. Alternatively, the keyword and following command can be uttered in one sentence, e.g., "Hey BMW, show me gas stations nearby." The second option to activate the speech system is to glance at the CID. In contrast to the keyword, this will only cause visual feedback on the CID as illustrated in Figure 8.5, to avoid annoying acoustic feedback whenever drivers make control glances at the CID. Moreover, the feedback was designed in an ambient way at the top of the screen, so that it is easily perceivable when actively looking for it, but at the same time it does not disrupt the user when triggered unintentionally.

When the system is active, it responds to the driver's commands. The prototype integrates a NLU engine, thus commands can be formulated in a natural way. If the driver uses a keyword to explicitly activate the system, it will reply to the user if the command could not be successfully mapped to a valid intent. In contrast, if they system was activated implicitly via the user's gaze, it does not respond to commands that could not be mapped to a valid intent. This is another way to reduce the impact of unintended activation of the system. Driver utterances that cannot be interpreted by the NLU, such as conversations between passengers, will not result in acoustic system feedback if the system was activated implicitly.

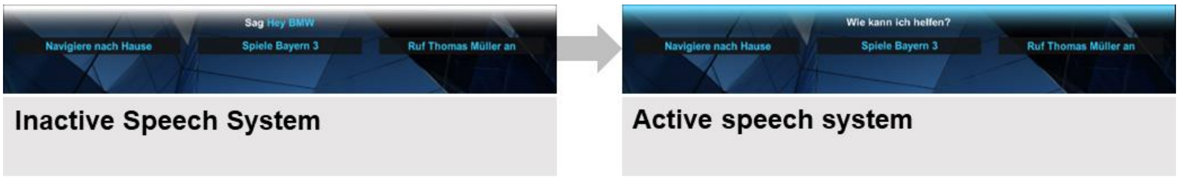


Figure 8.5: The speech recognition system can be activated with a keyword or by looking at the CID. The state of the speech system is permanently visualized at the top of the screen.

Browse List

In this task, the system presents a list of unknown elements. Drivers do not have enough information to name the element they want to select and thus have to browse through the list until they find the desired element. Again, this task is composed of two steps, a) moving through the list and b) selecting the desired element once visible. There are to two alternatives input modes to achieve part a): Drivers can use speech commands, such as “Next Page”, or “Back” to navigate through the list page-by-page. The second option is to use the remote touch pad in the middle console. It allows two-finger scrolling to browse the list, according to scrolling on laptops. Step b), the selection of elements, is always done via speech. As soon as the desired element is visible, drivers can select it by saying the number, the name, or a natural command, e.g., “please select Parkring 19 in Garching”.

Select Recommendation

User can ask the system for points of interest (POIs) such as restaurants or gas stations. The system displays a selection of three possible options from which users can choose. Each selectable POI is described with a name, as well as further details such as distance or average user ratings. Drivers can make their choice based on these details. One driver may prefer the nearest gas station nearby, while another may prefer the cheapest. Again, there are two alternative input options to select an option. First, drivers can say the name of the POI, which is written on top of each option. Second, they can use a pointing gesture with the right index finger towards the desired element. In order to compensate for pointing errors, the driver’s gaze point on the screen in the moment of selection is additionally incorporated as described in Section 5.5. Those two input options allow the driver to adapt to different requirements of the selection task. For example, some POIs, such as gas stations usually have simple names, such as “Esso”, “Shell”, or “Total”. They are relatively short, easy to pronounce and thus easy to speak and recognize. Names of restaurants, however, are often longer, unconventional or in a foreign language. This requires the user to read the name carefully to be able to say it correctly. Moreover, such names are challenging for speech recognizers and often result in erroneous recognition results. A direct selection of available options via a pointing gesture, does not require the user to read and speak the name of the restaurant, but simply point toward the desired option.

Confirm Details

After selecting a list item or system recommendations, the prototype requires users to confirm details of the action. For example, this could be confirming to start a new route guidance (instead of adding it to an existing route) or confirming to start a call (instead of sending a message) for contacts. These details are displayed on the top of the screen in horizontal alignment. Similar to *select recommendation*, users can confirm details by saying the name of the display command (“Start route guidance”), or a similar natural voice command (“Please start a route guidance to this address”). The difference to *select recommendation* is that drivers do not have to search for the best element first nor read the names, because the system will always provide the same details at the same position. For example, for a POI, the system will display the options “start route guidance”, “add to existing route”, and “show website” from left to right (also see Figure 8.4). Speech input profits from a consistent use of short and simple commands that are known to the driver. Gesture input is made easier due to a consistent placement of the available options (e.g., “start rout guidance” is always the leftmost item).

8.1.4. Sensor setup

The prototype was built into a vehicle mock-up of a 2016 BMW 5-series. Figure 8.6 shows the placement of the devices and sensors in the mock-up. The interaction took place on the CID (A). A *Tobii 4C eye-tracker* was placed behind the steering wheel and recognized the driver's glances on the CID (B). A *Logitech touchpad* was mounted on the middle console for scrolling through lists (C) and a *Leap Motion* gesture camera was mounted behind the central rear-view mirror (D) so that it covered an interaction area between steering wheel, CID and gearshift (according to (Riener, 2012)). The microphone was also placed on the rear-view mirror and directed towards the driver's seat. Speech recognition was achieved using the *Microsoft Cognitive Services* cloud platform with *Microsoft Speech Recognition* in combination with *Microsoft LUIS* (Language Understanding Intelligent Service) for NLU. The main software was developed in *Unity3D.*

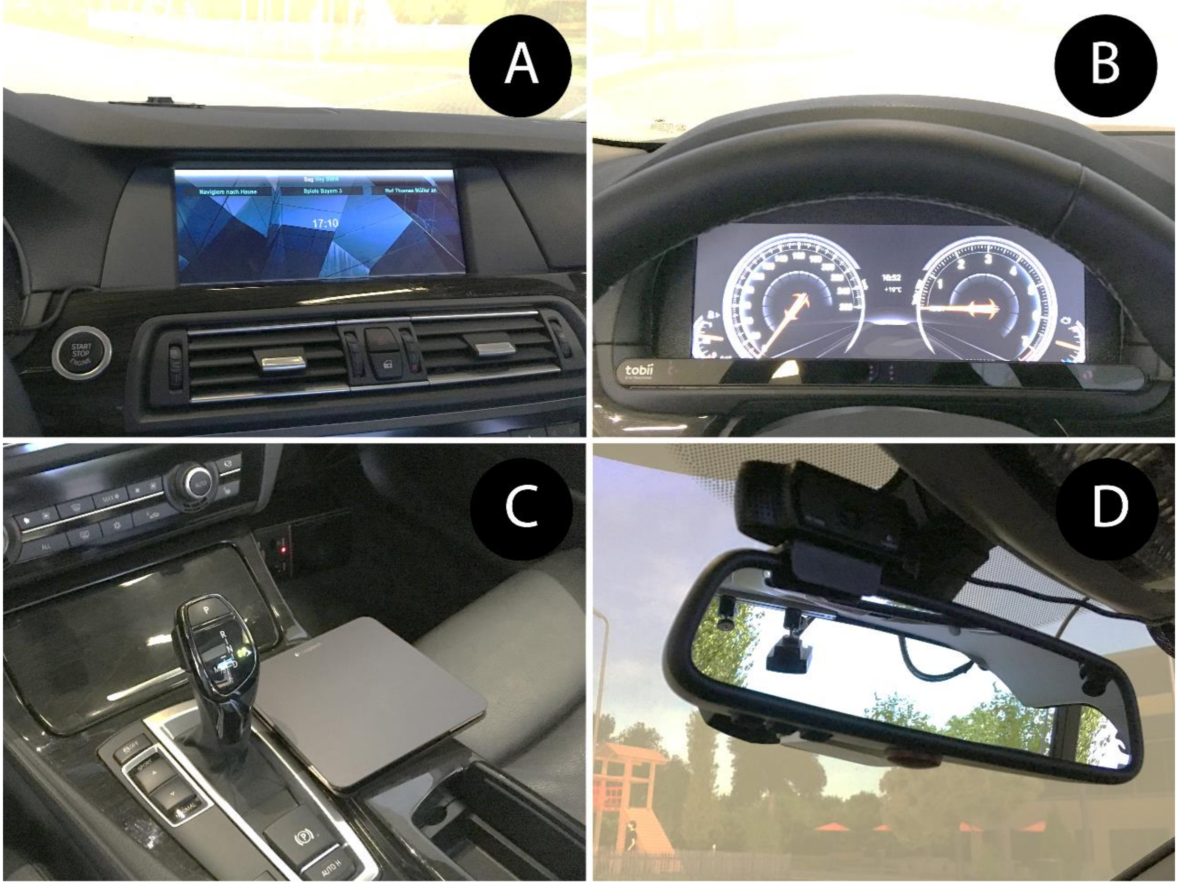


Figure 8.6: Placement of devices in the apparatus for the evaluation study.

8.2. Evaluation

The previous section described a multimodal prototype for operating secondary tasks while driving, which is based on the presented interaction design patterns. In this section, we present an evaluation of this prototype in form of a user experiment in a driving simulator. The evaluation will not focus on the specific effects of individual patterns since this was basically covered in the experiments in the previous chapters. Instead, the goal of this evaluation is to assess the overall potentials for the proposed form of multimodality for IVISs. Section 2.2.5

illustrated the potentials of multimodality in the automotive domain: reducing driver distraction, reducing task completion time, reducing driver workload, and increasing user experience. We investigate the prototype regarding these factors. As a reference we compare the multimodal interaction with a unimodal speech interaction approach.

8.2.1. User Experiment

We conducted an evaluation study of the prototype to determine the effects of our multimodal approach based on the presented design patterns in comparison to a unimodal speech only interaction in the car. The research questions for this evaluation are:

- RQ1: How does multimodal interaction perform in comparison to speech-only interaction in terms of driving performance, task completion time, workload, and user experience?
- RQ2: How do drivers use alternative input with gestures and touchpad when they have free choice of modalities?

Participants

We evaluated the prototype with 40 participants (8 female, 32 male) between 20 and 63 years (M=33.98, SD=12.52). All participants were BMW employees. Except for one left-handed and one ambidextrous participant, all participants were right-handed. They reported their openness to voice-, gesture-, and haptic control in the car on a scale from -3 (not open) to +3 (very open). There was a generally a higher openness to speech input (*Median=2*), compared to gestures (*Median=0*), and haptic input (*Median=1*). This is probably connected to the fact that most participants have experience with the use of voice-controlled systems, such as the BMW speech control (36%), Alexa (16%), Siri (23%), and Google Assistant (18%). 7% did not have any experience with speech interaction.

Design

For RQ1 each participant conducted two trials. One trial with only speech interaction and one trail with multimodal interaction. During the latter, the participants always used the alternative input modalities for the tasks, i.e., gaze for the activation of the speech system, gestures for selecting recommendations and confirming route guidance, and the touchpad for scrolling through lists (see Table 8.1). The order of appearance of both trials was counterbalanced over all participants. Within each trial, the participants conducted 28 use cases (12 x UC1, 8 x UC2, 8 x UC3) that were presented in randomized order. Use cases were instructed with automated speech announcements. There was a break of approximately five seconds after a use case was completed. We measured the participants driving performance, glance behavior, efficiency of interaction, cognitive load, and user experience.

For RQ2 each participant conducted a third trial in which they could decide for each interaction step which input modality they want to use. This trial was always performed at the end to ensure that participants had experienced both other trials before and were familiar with all input modalities. During this trial, the participants conducted 14 use cases (6 x UC1, 4 x UC2, 4 x UC3) in randomized order.

Procedure

After completing a form with demographic information, the participants adjusted the seat and steering wheel position in the mock-up so that they could comfortably reach for the touchpad in the middle console. Participants were instructed to keep their hands on the left and right side of the steering wheel so that the eye tracker was not covered. The driving simulation was started, and the eye-tracking system calibrated. After that the experimenter explained the system and the interaction for the three use cases using speech only and multimodal interaction. Participants completed each use case several times with both input forms while still standing on a parking lot. Next, they drove on the highway without interaction until they familiarized with the driving simulation (based on the experimenter's assessment). Then the trials with speech only and multimodal input conducted, followed by the third trial with free choice of input. During these trials, participants were instructed to follow a leading vehicle at 100 kilometers per hour with a headway of 50 meters. They were also instructed to prioritize the driving task, to keep in the middle of the right lane, and not to overtake the leading vehicle. After each trial the participants completed DALI and UEQ questionnaires. Finally, the participants rated the suitability of speech and the according alternative input modality for the interaction four steps.

8.2.2. Results

Driving Performance

The driving performance was measured based on standard deviation of lateral position (SDLP) and the standard deviation of distance (SDD) of the requested distance to the leading vehicle. The results are shown in Figure 8.7. The data for SDLP and SDD is distributed normally according to Kolmogorov Smirnov tests. Pairwise t-test indicate that the mean SDLP is lower in speech only (M=0.139, SD=0.355) than multimodal (M=0.166, SD=0.397), $t(39) = -5.468, p < .001, r = .467$. In contrast, SDD did not differ significantly between speech (M=13.53, SD=1.57) and multimodal input (M=13.69, SD=1.89), $t(39) = -0.605, ns$.

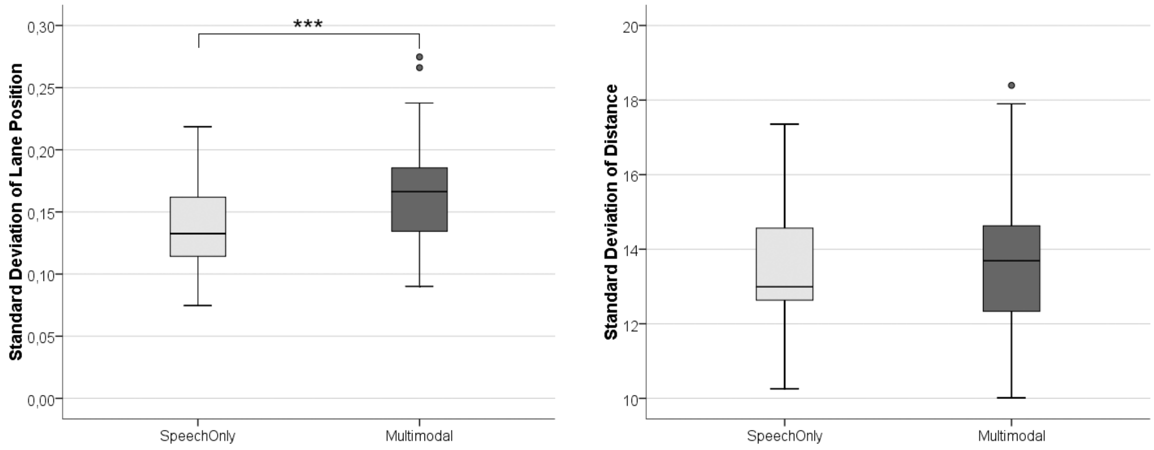


Figure 8.7: The standard deviations of lateral position (left) and requested distance (right) during speech only and multimodal interaction trials.

Glance Behavior

The participants glance behavior was analyzed based on the average total glance time (TGT) during the trials and for the individual tasks. The left graph in Figure 8.8 indicates that multimodal input (M=4.76, SD=1.45) resulted overall in a longer total glance time on the CID than only speech input (M=2.85, SD=1.12). A t-test shows that this difference is statistically significant ($t(39) = -11.408, p < .001, r = .769$). The effect size indicates a strong effect.

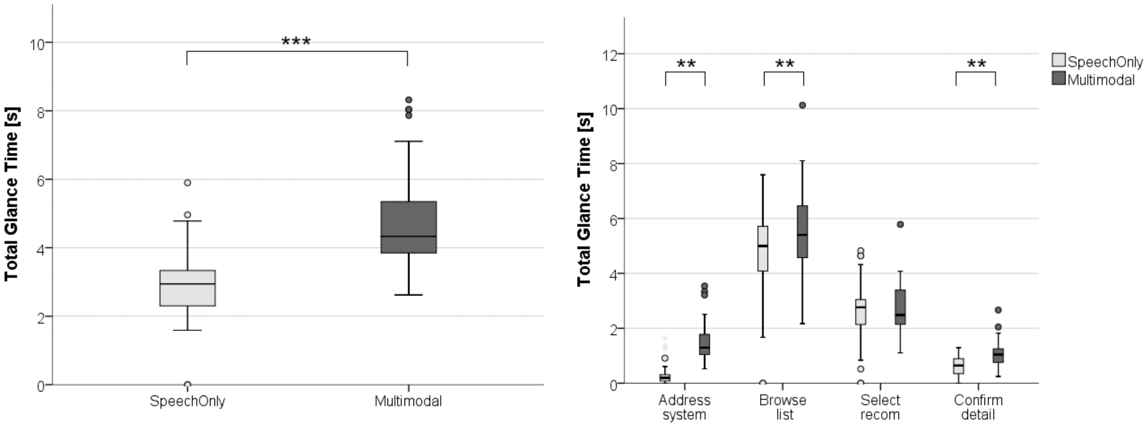


Figure 8.8: Total glance times during speech only and multimodal interaction per trial (left) and per individual task (right).

Furthermore, we can analyze the total glance durations for individual tasks. This is shown in the graph on the right in Figure 8.8. On average all tasks require longer glance times on the screen in the multimodal condition. This shows that the overall increase of visual distraction in the multimodal condition is not only caused by one task (e.g., gazes on the CID for *address system*), but also by using the touchpad for *browse list* or pointing gestures for *confirm details*. However, the use of pointing gestures for s*elect recommendation* did not lead to a significant increase of glance time on the screen compared to speech selections. Table 8.2 gives an overview over the statistics.

Task	Speech Only	Multimodal	t-test / Wilcoxon
Address system	0.33 (0.41)	1.49 (0.71)	Z=5.511, p<.001, r=.871
Browse list	4.71 (1.56)	5.58 (1.45)	T(39)=3.587, p< .01, r=.248
Select recom.	2.60 (1.09)	2.77 (0.97)	T(39)=1.138, p=.26, ns.
Confirm detail	0.62 (0.33)	1.07 (0.48)	T(39)=6.44, p<.001, r=.515

Table 8.2. *Statistical overview of total glance times [s] for individual tasks during speech only and multimodal input.*

Task Completion Time

The overall task completion time was lower for multimodal input (M=11.29, SD=2.11) compared to speech only (M=12.62, SD=1.60). This is shown in left graph of Figure 8.9. The difference is statistically significant ($t(39) = 4.929, p < .001, r = .390$). The effect size can be classified as a medium effect.

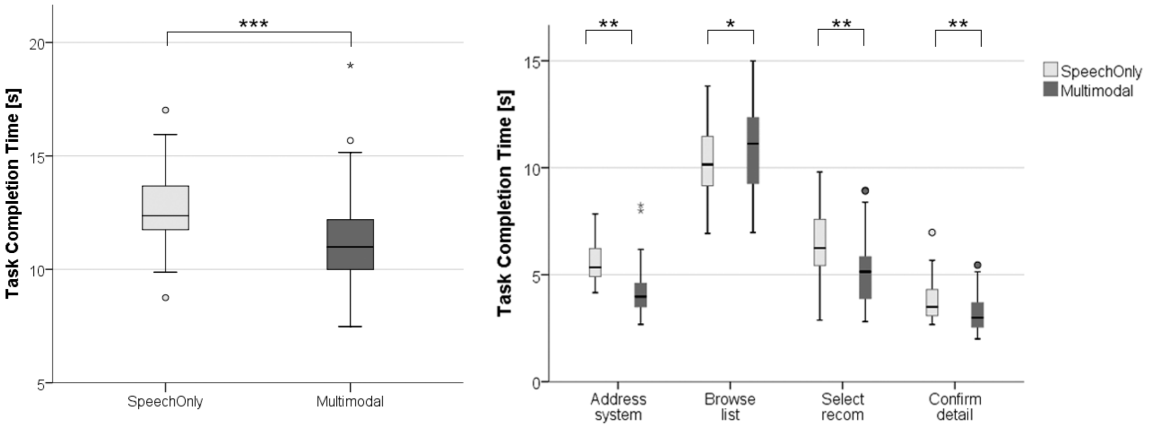


Figure 8.9: Task completion times during speech only and multimodal interaction per trial (left) and per individual task (right).

The right part of Figure 8.9 shows the TCT for the individual tasks. Multimodal interaction led to significant improvement of TCT for *address system*, *select recommendation* and *confirm detail*, while *browse list* was more efficient in the speech only condition. Table 8.3 shows the statistical comparisons.

Task	Speech Only	Multimodal	t-test
Address system	5.61 (0.90)	4.21 (1.17)	T(39)=10.193, p<.001, r=.727
Browse list	10.31 (1.70)	10.89 (2.10)	T(39)=2.212, p<.05, r=.111
Select recom.	6.36 (1.59)	5.19 (1.53)	T(39)=3.664, p<.01, r=.256
Confirm detail	3.73 (0.90)	3.15 (0.80)	T(39)=2.920, p<.01, r=.179

Table 8.3: *Statistical overview of task completion times [s] for individual tasks during speech only and multimodal input.*

User Experience

The ratings of the UEQ questionnaire are summarized in Figure 8.10. Attractiveness and pragmatic quality did not differ significantly between speech only and multimodal conditions. However, the hedonic quality of the multimodal condition was rated significantly higher than speech only ($Z = 3.573, p < .001, r = .565$).

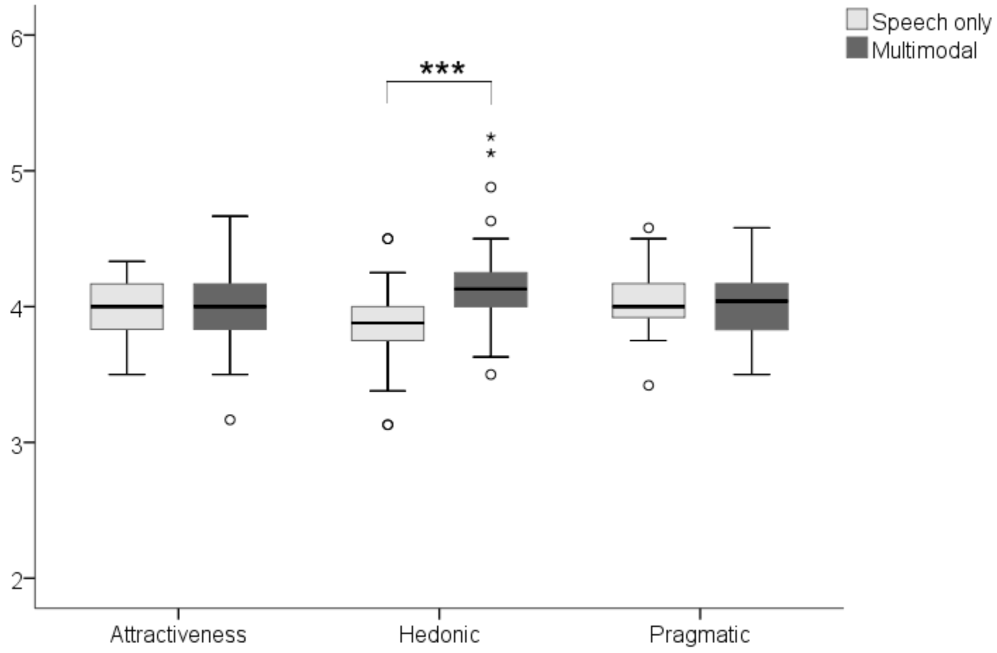


Figure 8.10: User experience ratings based on attractiveness, pragmatic and hedonic quality. Note that the scale has been adapted to a smaller range for better readability. The complete UEQ scale ranges from 1 to 7.

UEQ-dimension	Speech Only	Multimodal	Wilcoxon
Attractiveness	4.01 (0.21) 4	3.95 (0.30) 4	Z=-1.278, ns.
Pragmatic Quality	4.04 (0.23) 4	4.00 (0.25) 4.04	Z=-0.386, ns.
Hedonic Quality	3.84 (0.32) 3.875	4.13 (0.36) 4.125	Z=3.573,p<.001, r=.565

***Table 8.4**. UEQ Ratings on three dimensions for speech only and multimodal input.*

Subjective Demand

Overall, the multimodal condition led to significantly higher ratings on the DALI questionnaire for all dimensions. The only exception were ratings for auditory load, which was higher for speech only (*ns.*). All DALI results and statistical comparisons are summarized in Table 8.5.

DALI-dimension	Speech Only	Multimodal	Wilcoxon
Attention	1.92 (1.10)	3.15 (0.92)	Z=-4.941, p<.01
Visual	1.82 (1.08)	3.20 (0.97)	Z=-4.779, p<.01
Auditory	1.85 (1.17)	1.64 (1.04)	Z=-0.836, ns.
Tactile	0.70 (1.02)	2.70 (1.38)	Z=-5.043, p<.01
Stress	1.25 (1.01)	2.70 (1.40)	Z=-4.599, p<.01
Temporal	1.07 (1.02)	1.93 (1.23)	Z=-3.554, p<.01

Interference	1.77 (0.92)	3.05 (1.04)	Z=-4.646, p<.01
Global	1.62 (0.81)	2.62 (0.81)	Z=-4.977, p<.01

***Table 8.5**: Overall DALI scores for the speech only and multimodal condition.*

Use of Alternative Modalities

After the multimodal and the speech-only runs, participants conducted a third run with the same use cases during which they could freely decide which input modalities to use. The following results do only relate to this third run.

At the beginning of each use case, participants decided whether to address the system with a keyword ("Hey BMW") or by looking at the screen. Latter was used in 57.0% of all use cases. The active use case did not have an influence on the participants' choice (UC1: 56.7%, UC2: 57.0%, UC3: 57.3%). If the instructed function was offered on the start screen, participants could alternatively select functions the start-screen directly with a pointing gesture, which was used in 50% of the available cases.

In UC2, the participants asked for a list of recent contacts or destinations and could then decide between speech input, and a remote touchpad to browse through the list. The touchpad was used in only 18.29% of the cases to move the list. In the following step, gestures were used in 31% to confirm details. Figure 8.11 shows differences depending on which modality was used to move the list in the task before. After moving the list with speech, gesture confirmation was used in 26.62%, but after using the touchpad it was used in only 4.17% of the selections.

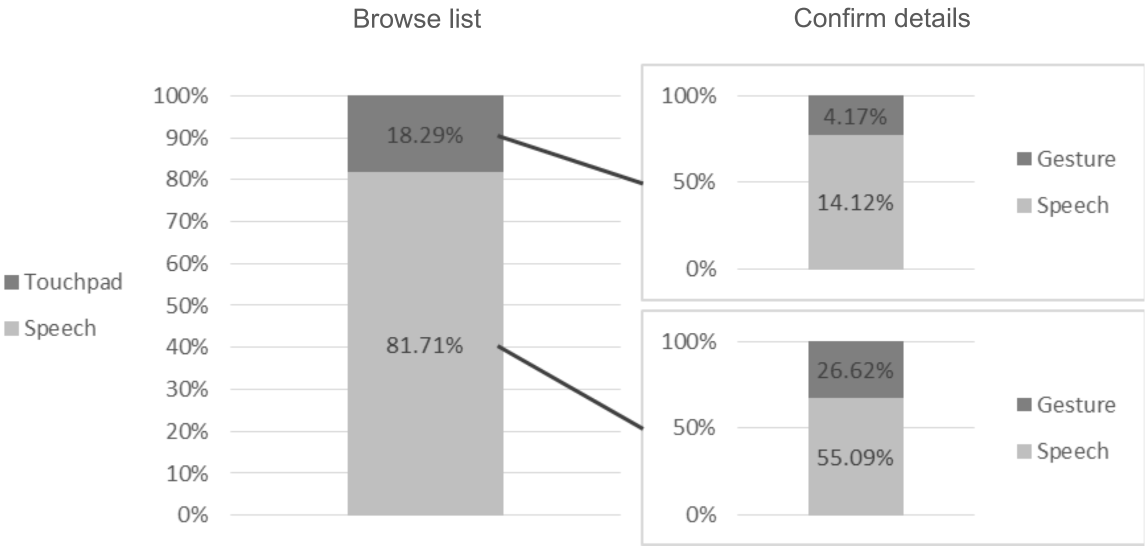


Figure 8.11: Use of modalities in UC2: Browsing through the list followed by confirm details.

In UC3, participants asked for gas stations or restaurants nearby and then selected one out of three recommendations using a pointing gesture or by saying the name of the desired element. Figure 8.12 shows that gestures were used in 47.15% of selections for recommendations. In the following step, they were used in 41% for confirm details. This is split into 31.58% when gestures were previously used for *select recommendation*, and 9.21% gesture usage when speech was used for *select recommendation*.

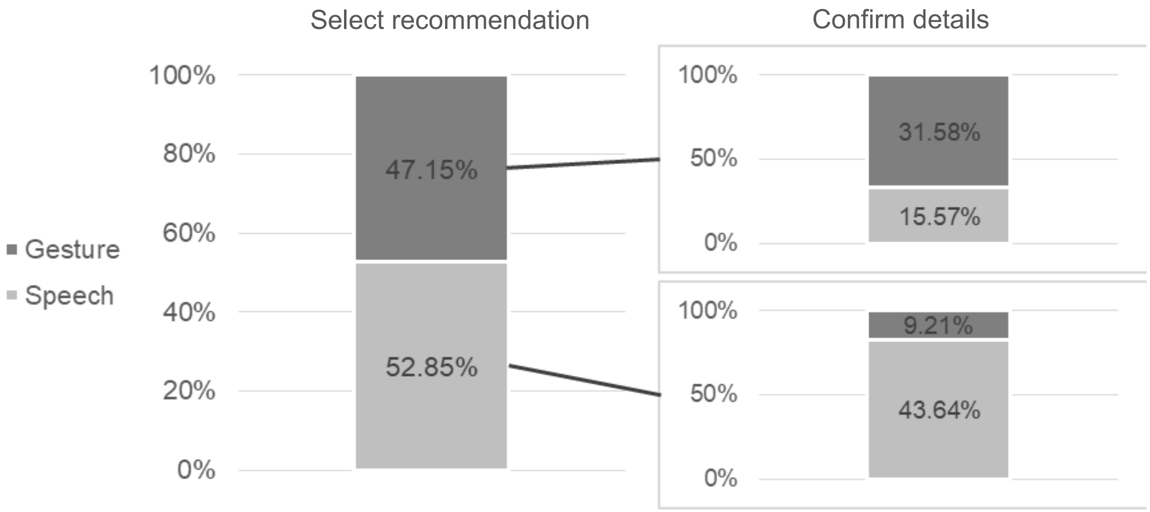


Figure 8.12: Use of modalities in UC3: Select recommendation followed by confirm details.

Suitability of Input Modalities

At the end of the evaluation, the participants rated their overall impression regarding the suitability of the input modalities, gaze for the activation of the speech system, the touchpad for moving lists, and pointing gestures for selections on the screen, in comparison to speech input. Ratings were conducted on a 7-point Likert-type scale ranging from very bad (-3) to very good (3). For each interaction step, speech input was rated significantly more suitable than the according alternative input modality. This is illustrated in Figure 8.13.

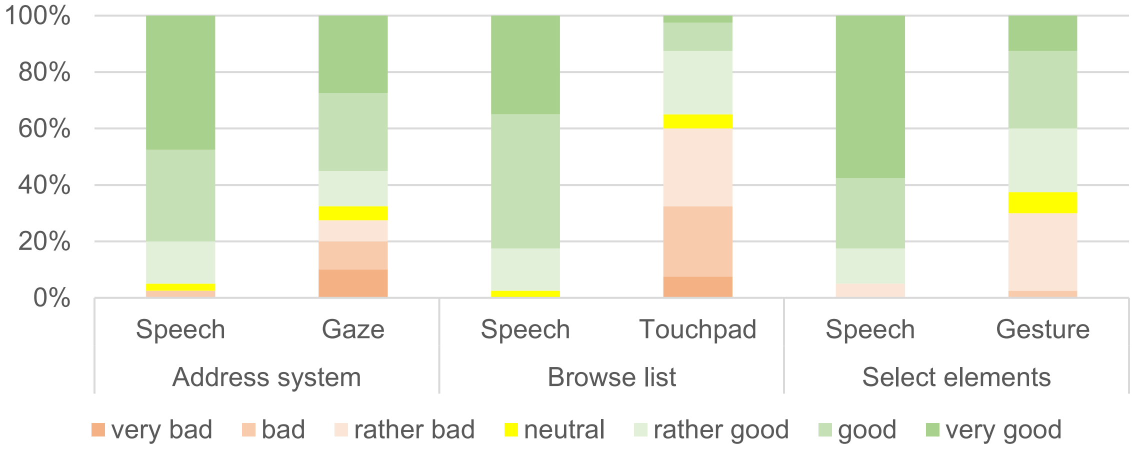


Figure 8.13: Suitability ratings for the interaction with speech and alternative input modalities for individual types of tasks.

For addressing the system, speech (*Median=2*) was rated better than gaze (*Median=2*), $Z = 2.771, p < .01, r = .438$. The difference was greatest for browsing the list. Speech (Median=2) was rated much better than the touchpad (*Median=-1*), $Z = -5.82, p < .01, r = .0920$. Finally, the selection of elements on the screen (select recommendations and confirm details) was also rated better with speech (*Median=3*) than with gestures (*Median=1*), $Z = -3.79, p < .01, r = .600$.

8.2.3. Discussion

Multimodal vs. Speech only

RQ1 focused on the effects of multimodal input in comparison to a speech only approach in terms of driving performance, visual distraction, efficiency, and user experience. Overall, the multimodal condition led to significantly longer glance times on the screen than speech only, which also led to a slightly decreased lane keeping performance. Moreover, glance data shows that increased visual distraction occurred for all alternative modalities, except for the selection of recommendations. This task first required drivers to identify the desired element, by comparing distances (or ratings) of the recommendations. With gesture input they could then directly select this element, whereas speech input required to also look for the name of the element in order to say it. The direct selection of elements could thus be beneficial when speech input would require drivers to name longer and/or more complicated descriptions that they must read first before they can speak them out loud. For all other tasks, however, speech only input is less visually distracting.

The greatest advantage of multimodal input was an overall increased efficiency. It reduced the average task completion time for all use cases by 8% compared to speech only interaction. Gaze-based speech activation and gesture selection for recommendations and confirmation of details mainly contributed to this improvement. However, browsing through the list with the touchpad was significantly slower than speech. It must be respected that there was a different scrolling behavior for both input modes. Speech commands moved the list forward or backward by an entire page, whereas touchpad input scrolled the list smoothly. Moreover, the participants never had to scroll more than two pages during the experiment. For longer lists with many pages, the direct manipulation with the touchpad is likely to provide an input option, which is at least equally efficient as speech input.

Participants rated their cognitive load during the multimodal condition significantly higher than during speech only. We attribute this difference to three aspects. First, almost all participants were experienced with the use of speech input, while the experience with gesture and gaze interaction was very small. Although we included practice runs before the actual trials, the level of workload is very likely to be influenced by the novelty of modalities in the multimodal condition. Second, during the multimodal condition participants were instructed which modality to use for the individual steps. Despite the small number of steps (max. 4 steps) and the practice runs, some participants found it hard to remember the instructed modalities, which potentially resulted in increased workload for recalling the correct input modality for each step. Third, the mean difference of the global dimension is one point of the 5-point scale. The greatest difference of user ratings occurred for the tactile demand dimension (0.70 vs. 2.70). This suggests that especially the use of the remote touchpad contributed to the higher cognitive demand, which is in accordance with the longer interaction times. The results of the free trial further confirm this observation, as the use of the remote touchpad was largely omitted. In regard of these aspects, we expect that more experience with gesture and gaze input, longer training phases to practice the sequencing of input modes, as well as an improved touchpad functionality would lower the cognitive load of the multimodal approach.

UEQ ratings for the multimodal condition indicate a higher hedonic quality than only speech input. Based on observation we assume that especially gestures and gaze led to higher ratings for originality and stimulation of the interaction because of the novelty of these input modes. Participants thought of the multimodal condition as an interesting path, however, without seeing a pragmatic benefit over speech only interaction.

Use of Modalities

RQ2 focused on the use of alternative modalities during the free trial. Gaze-based speech activation was used for a majority of use cases, which indicates a good acceptance for this approach. This could have been additionally influenced by laziness of the participants, as they completed quite many use cases and could get tired of speaking over time. In this context, gaze-based activation can be a convenient alternative, especially when users are frequently addressing the system via speech.

Regarding gesture selections, the greatest use of gestures was made for the selection of available links on the start screen. Half of the selections of these links were made with a pointing gesture. This way participants could sometimes completely avoid speaking for UC1. In UC2 and UC3, participants always had to start with a speech command, which led to a lower use of gestures in the following interaction steps. Moreover, if there was a switch for the second interaction step, participants tended to stay with this mode, as further switches for the last step hardly occurred (see Figure 8.11 and Figure 8.12).

The touchpad generally suffered from poor acceptance, reflected in the low usage during the free trial. Several aspects could have an influence on this behavior. First, the interaction could not be completed using only the touchpad. Thus, using the touchpad to browse the list always demanded for another modality switch to confirm details. As a consequence, participants avoided the touchpad in order to avoid multiple modality switches. Second, speech input was faster and required less glances to control the list and therefore participants did not profit from switching to the touchpad.

8.3. Summary

This chapter described an interactive multimodal prototype that illustrates the implementation of the patterns in an IVIS and evaluates it in a driving simulator experiment. In Section 8.1, we describe the software application and how it integrates the presented design patterns from Chapter 7. It is designed as a speech-first application that uses speech as primary input mode, but also provides complementary input with gestures for confirmations and small selections, and a remote touchpad list interaction. In addition to that, the drivers gaze input is used for speech activation and to improve the accuracy of gesture inputs.

Section 8.2 describes the evaluation of the prototype in a driving simulator experiment. We investigated the performance of the multimodal system compared to a speech-only system, as well as the usage of the different modalities in a while driving. The results show that the multimodal approach based on the presented patterns led to increased efficiency and slightly better user experience. The downside of the multimodal approach is a higher visual distraction which also reflects in a slight decrease in lane keeping performance.

9. Conclusion

This thesis investigated natural multimodal input for in-vehicle interaction. It aims to support developers in combining speech, mid-air gestures, and gaze input for controlling secondary tasks while driving. This chapter summarizes the main contributions of this thesis and presents answers to our research questions. Finally, we point out future research directions before completing this work with some concluding remarks.

9.1. Summary of Contributions

The contribution of this dissertation to the field of multimodal in-vehicle interaction contains three parts: First, we presented the first pattern collection for multimodal in-vehicle interaction based on the findings from our experiments and related literature from the field. Second, we demonstrated interaction techniques, which combine speech, gestures, and gaze input to compensate for individual limitations. Third, we presented empirical results from a series of user experiments that generate insights regarding the benefits of multimodal input while driving. Table 9.1 gives an overview over the contributions and how they answer our research questions. The contributions are discussed in detail in the following sections.

No.	**Research Question**
R1	**Investigating potentials of multimodality while driving**
R1.1	**How do modality switches influence interaction while driving?** Switching to better suited modalities for subtasks increases the efficiency of the interaction. The process of switching input modalities does not induce costs when it occurs during a task switch. However, people tend to avoid modality switches when the benefit over the current modality is not sufficient.
R1.2	**How do situational demands influence input modalities?** Situational demands have individual effects on speech, gesture, and gaze input. However, the examined demands did not change the relation of demanded resources during the primary task of driving. Therefore, speech input was the preferred input modality over a variety of situations and tasks, despite some limitations in auditory demanding situations.
R2	**Supporting the flexible use of alternative input modes**
R2.1	**How can user interfaces promote the use of speech input?** The use of speech input is mainly determined by the type of secondary task and the demands of the driving scene. Additionally, visual prompts have a significant influence on the driver's choice without restricting the freedom of choosing a modality himself and can therefore be used to promote the use of speech input.
R2.2	**How can user interfaces effectively support the execution of gestures?** A lack of feedback and limited accuracy of gesture input can be overcome by providing peripheral light feedback (PLF) that visualizes the execution of gesture commands. Generally, drivers prefer PLF over only acoustic feedback and feel better supported. For simple gestures such as confirmations, PLF is a feasible approach. However, for pointing gestures PLF should be used with care, as a higher accuracy comes at the expense of longer interaction times, and perceived greater distraction

R3	**Enhancing interaction by combining input modalities**
R3.1	**How can multimodal input be used to improve gaze input?** The confirmation of gaze pointing with speech or gestures did not lead to an improvement compared to the dwell time approach. We point out modality-specific problems of combining speech and gestures with active gaze input. Thus, the preferred confirmation mode was a button on the steering wheel, which led to shortest interaction times and least visual distraction. Overall, we observed a high cognitive workload for active gaze pointing in general. Therefore, we do not recommend the use of active gaze input for selections while driving manually.
R3.2	**How can multimodal input enhance speech input?** Gaze input can be used in a passive way to determine the driver's intention to interact with the system and to activate the speech system automatically. Evaluation of the prototype shows that it unites a high efficiency and good user experience. However, the benefits of the approach are limited by two factors. First, it is limited to display-related tasks. Gaze activation for non-display related tasks led to higher visual distraction and increased workload. Second, gaze information is connected to strong personal preferences and resulted in polarizing feedback. Therefore, gaze-based speech activation cannot replace exiting techniques, such as the keyword but provide a valuable alternative.
R.3.3	**How can multimodal input support mid-air gestures?** We believe that the potential of gesture input in the car lies in fast interaction for simple tasks like pointing, which cannot be performed with speech input. Visual feedback can help for compensating the lack of accuracy for pointing, but it leads to longer interactions with higher distraction. Therefore, gaze information can be passively integrated to increase the accuracy of pointing gestures, while maintaining a fast and easy selection.
R4	**Providing design support for developers**
R4.1	**How can design knowledge be made available for designers in a reusable way?** We present a collection of design patterns for multimodal in-vehicle interaction, which combines the insights from the presented experiments with existing design knowledge from the field multimodal interaction. Currently, this collection includes 15 design patterns that describe how to apply natural input modalities while driving. Besides entirely novel patterns, the collection also contains patterns from non-automotive literature, which have been adapted to the automotive domain.

***Table 9.1**. Overview of contributions and research questions.*

9.1.1. Design Patterns

The results of the empirical studies and the evaluation of the prototypes generated a large amount of knowledge for the design of multimodal systems. We use the concept of design patterns to summarize, structure, and organize our findings in a formalized way. Thereby, we provide effective empirically validated design support for designers of multimodal IVISs. At the same time, the structure of the presented pattern collection serves as an example how to apply the concept of design patterns on in-vehicle interaction concepts. The pattern collection is represented as a directed graph that consists of three different levels of abstraction. High-level patterns contain abstract solutions on different ways of combining modalities, medium-level patterns contain solutions to reoccurring problems for certain groups of modalities, and low-level patterns provide solutions for modality specific problems. Relationships between patterns describe how novel patterns are interconnected and to point out the connection to existing patterns from literature. This way, the presented design patterns present a consistent collection of design knowledge for the support of multimodal in-vehicle applications.

9.1.2. Interaction Techniques

We presented interaction techniques that combine speech, gestures, and gaze for in-vehicle interaction, which are based on the existing research and the results from our experiments. Speech input should be integrated as the primary modality for in car interaction. Gestures have the potential to complement speech for simple selection tasks on remote displays, if they allow a fast and effortless selection. Gaze information should be only integrated in a passive way, as multimodal confirmation techniques could not overcome the inherent visual distraction of using gaze as an explicit input modality.

Passive Gaze Integration

Gaze input provides a fast pointing mechanism but is limited to adequate confirmation techniques that avoid the Midas touch problem and reduce visual distraction and the feeling of unnatural staring. We examined the use of speech and gesture input as alternative confirmation modes. However, both options have specific limitations. Hand gestures easily distract the driver's gaze from the target. Speech confirmation is slow and does not provide timely feedback, which results in high visual distraction and gaze input errors due to control glances on the street while waiting for feedback. A better solution is the combination with haptic input on the steering wheel, which leads to reduced visual distraction and increased efficiency compared to the dwell time approach. However, the perceived suitability does not improve compared to the dwell time approach and a high cognitive load persists when gaze is used for pointing input. Therefore, we conclude that gaze should not be used as an explicit input mode, but rather be integrated in passive way to support speech and gesture input.

Optimizing for Speech

The use of speech input is beneficial while driving, but people don't necessarily use this potential when they can choose input modes themselves. We identify explicit textual cues as an effective and easily applicable solution to reduce this problem. They increase the use of speech input and reduce visual distraction without restricting the users' freedom choice or making users feel restricted. Another hurdle for using speech input is the need to explicitly activate the system. We present a prototype that uses a gaze-based activation of the speech system. The evaluation showed that this approach increases efficiency and user experience for display-related tasks, but that it is less ideal for non-display-related tasks. Thus, it cannot replace existing techniques, but provide a valuable addition.

Complementary Gestures

The potential of mid-air gesture input while driving is limited. As gestures are limited in transferring semantic information, we believe that the greatest potential of gesture input in multimodal systems is complementary input to speech. In particular, pointing gestures are a natural way of creating spatial references that are difficult to convey using language. Moreover, they are an efficient option to make simple selections or confirmations. However, gesture input often suffers from a lack of accuracy due to missing feedback and a lack of understanding of the correct execution, which is especially due to the novelty of gestures input for many users. In this regard, we show how peripheral light feedback can effectively support mid-air gestures in terms of perceived support and accuracy. The main limitation of this approach is that the additional feedback results in a greater effort for the driver to make adjustments. For pointing gestures, this results in longer interaction times and is perceived as visually distracting. Therefore, peripheral light feedback should only be used for simple selection and confirmation gestures, but not to increase the accuracy of pointing gestures while driving. To solve this problem, we present a prototype that integrates passive gaze information for pointing gestures. This approach can increase the pointing accuracy, while maintaining an efficient interaction and without increasing cognitive demand.

9.1.3. Empirical Evaluation of Potentials

Finally, we contribute to the research field by providing the results of user experiments that help to assess the potential benefits of multimodal interaction in while driving. One of the greatest potentials of multimodal human-computer interaction is the flexible use of alternative input modalities. This is especially promising in the automotive domain, where driver-vehicle interaction needs to cope with variety of tasks, environmental influences, and user preferences.

Use of Multiple Modalities while Driving

A flexible use of different input modalities requires driver to switch between modes. Therefore, we investigated the cost of modality switches. Our results show that modality switches between touch and speech input can be performed without additional costs when they occur during a task switch. The interaction is rather determined by the combination of the task and the used modality than by the process of switching. This observation encourages the flexible use of different modalities, which allows drivers to choose different input modes based on their current needs. In this regard, we also found that situational demands influence the suitability of speech, gestures, and gaze input.

However, these effects are limited. The demands of the driving task and the influence of the task type have a much bigger influence on the suitability of input modalities. Accordingly, it showed that for many tasks people prefer speech input while driving due to low visual distraction and little cognitive workload, although touch, gestures and gaze allow faster input. Speech was even preferred during increased auditory demand, as there were no objective impairments of speech input, and only slightly reduced perceived suitability. Despite these advantages, we observed that many drivers do not choose speech input when given free choice between touch and speech input, which resulted in inefficient and distracting interaction. Especially for verbal tasks, such as text entry, drivers should be guided to use speech input.

The potential for touch and gesture input, on the other hand, lies in complementary input for specific use-cases in which speech input is less feasible, such as simple direct selection tasks (e.g., selections of objects on the screen). In these cases, speech input may take relatively long, while touch and gestures are much more efficient with low visual distraction. In this regard, it is especially important for gesture input to provide a very easy and effortless form of interaction so that it can effectively compliment speech input.

9.2. Future Work

Many results of this thesis were mainly derived from user experiments with interactive prototypes. These prototypes were built with currently available technology to provide a realistic experience for the users and thus to achieve meaningful results. While most of the presented findings are independent of the technology used, the capability and quality of the available technology will always influence the way users interact with interactive systems. Future work in this research area will therefore focus on the integration of novel technologies and how they can be used to improve driver-vehicle interaction. The following sections show the most relevant areas for future work from our point of view.

9.2.1. Novel Feedback Technologies

This thesis has shown the feasibility of driver-vehicle interaction based on the combined use of speech and complementary gestures. It has also shown that both modalities face the challenge of providing adequate feedback to the user. While users can profit from direct haptic feedback when using touch input or haptic controllers, they often feel decoupled from the vehicle and not adequately supported during speech and gesture input. Sections 5.1 and 5.2 approach this challenge by exploring new ways on how to guide the drivers during the use of speech and gestures. However, more work needs to be done so that drivers fully accept these technologies, especially mid-air gesture interaction. The system needs to provide further feedback that enables are more easy and understandable use. Future research must explore how to create an effective affordance for gesture interaction and how to better support users during the interaction. One promising approach is the application ultrasound feedback. Some researchers have already pointed out the potential of this technology to support mid-air gestures in the car (Harrington et al., 2018; Shakeri et al., 2018). However, it remains to be shown how to best apply this technology for in-car use, if it can expand the application scenarios of gesture input, and how well ultrasound feedback can work in real-world driving scenarios.

9.2.2. Beyond Explicit Interaction Techniques

In the context of manual driving, it is essential to enable efficient, safe, and easy interaction with secondary tasks. At the same time, all forms of non-driving related activities induce a certain amount of driver distraction. Therefore, generally reducing the amount of user effort needed for these forms of interaction is desirable. However, there will always be the need for users to explicitly interact with the system. Typical examples are the choice of music, or the selection of navigation destinations. Sections 6.2 and 6.3 demonstrated ways on how to reduce user input for such use-cases by integrating additional implicit context information (i.e., gaze). Based on these first promising results, our future work will focus on the development of more elaborate fusion algorithm for the integration of gaze information.

Other functions in current vehicles, such as headlights and windshield wipers are already largely automated and operate completely without user input. In contrast, most of the comfort and infotainment functions (e.g., ventilation, temperature, interior light) must be actively controlled by drivers. Further automation of these functions can reduce the risk of driver distraction. With the help of in-vehicle sensors, such as cameras and microphones, the driver state can be modelled, which allows driver-vehicle interaction that goes beyond explicit interaction. Instead, users might interact with the system in an unconscious, implicit way. In order to reach this goal, research work is needed to determine key context factors, which serve as parameters for the automation of vehicle functions. Which functions and features can be automated? How much control has to be left to the user? Which type of context information is needed (e.g., physiological and emotional state), how can it be acquired (e.g., heart rate variability, facial feature recognition, EEG), and how can the user interface make automated adaptions transparent to the user?

9.2.3. New Sensory Modalities

This thesis focused on the efficient combination of natural input modalities based on sensor technologies that can be already found in current vehicles. Microphones are used to capture the driver's speech output, and camera-based systems detect gestures and gaze behavior. However, all current systems suffer from technical limitations regarding accuracy and performance of natural input and therefore restrict possible field of applications. One way to approach this problem is to improve the existing type of sensors and the processing of sensor data. An alternative option is to integrate new sensory modalities that allow a broader range of different data types to be captured. Moreover, these additional sensors do not necessarily have to be installed in the vehicle. In particular, wearables such as smartwatches, fitness trackers, or wireless in-ear headphones are promising. Over the last years, they have become widely accepted in our society and are getting more and more popular. The fact the users wear the sensors directly at their bodies makes them independent of recognition areas of cameras and microphones. This way they can enhance information from existing sensors, but also allow overcome their limitations. For example, smartwatches could be used to support gesture input by providing exact information about the position and orientation of the driver's hand, independent of the users' position in the vehicle. Furthermore, smartwatches allow to integrate new information sources, such as physiological data, which allow to support novel use cases in the car. In-ear headsets could be used to improve local independent speech recognition, capture head-gestures, or provide private audio feedback. Research is needed that explores the potential of these novel sensory modalities and how they can integrate in the existing sensor environment.

9.2.4. Extension of the Pattern Collection

The patterns presented in Chapter 7 represent the first collection for multimodal interaction design patterns that respect the specific requirements of the automotive domain. Just like any other pattern collection or pattern language, it does not represent a final state. Instead, the collection must be continuously developed, updated, and extended based on new empirical result and proven concepts. Future research on novel feedback technologies, new input modalities and interaction techniques in the vehicle have the potential to generate new powerful patterns, which may be integrated into the pattern collection to extend or replace existing solution.

9.3. Concluding Remarks

This dissertation summarizes a variety of insights that we have generated over more than three years of multimodal research. We presented an exhaustive literature research, which summed up the interplay of the two main research fields addressed in this thesis: multimodal interaction and automotive user interfaces. A series of user experiments was conducted with a focus on different aspects of multimodal interaction in the car, such as the cost of modality switches, or the combination of two input modes. The insights and results from both, existing literature and these experiments, resulted in the first collection of design patterns for multimodal in-vehicle interaction. In this last section we summarized our contributions and answered the main research questions of this thesis. Now, at the end of this last section, want to make some brief concluding remarks.

The mere availability of speech, gesture and gaze input does not make the interaction with the system easier, more efficient, or even more natural. All three modalities can definitely be a valuable addition to automotive user interfaces, but their potentials are strongly depending on the use-cases and the way how the modalities are applied. In particular, we do not consider active gaze input when driving manually to be a viable option. This does not mean, however, that the driver's gaze should not be taken into account. It can provide very valuable information about the user's attention, which can be used as a passive input channel. Speech interaction has a great potential to cope with a great variety of use cases and allows for more complex input while driving, but it can be an overkill for simple interactions such as confirmations and selections. Gesture input may serve as a quick and easy alternative input for such tasks, but (just like speech input) gesture input is not self-revealing and users will need guidance by the user interface to make efficient use of them. Furthermore, the combination of input modalities does not guarantee that the advantages of each modality add up. It depends on the concrete implementation and small details if the combination will produce a benefit or even result in a reverse effect. For example, sections 6.1 and 6.3 showed that speech input in combination with temporally critical states, such as pointing gestures or gaze, can have a negative influence on distraction and glance times.

This thesis presents design support for natural multimodal applications in the form of a pattern collection. At the same time the presented contributions are only a small step towards a mature form of natural multimodal communication between humans and their future vehicles. We want to encourage researchers, developers, and designers of multimodal systems to continue working in this large field of research and to increase design knowledge about natural multimodal interaction so that one day we will not only press buttons and tap screens, but really communicate with our cars.

10. References

Ahmad, B. I., & Langdon, P. (2018). How Does Eye-Gaze Relate to Gesture Movement in an Automotive Pointing Task? *Advances in Intelligent Systems and Computing*, *597*(August 2017), 423–434. https://doi.org/10.1007/978-3-319-60441-1

Ahmad, B. I., Langdon, P. M., Godsill, S. J., Donkor, R., & Wilde, R. (2016). You Do Not Have to Touch to Select : A Study on Predictive In-car Touchscreen with Mid-air Selection. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 113–120. https://doi.org/10.1145/3003715.3005461

Ahmad, B. I., Langdon, P. M., Godsill, S. J., Hardy, R., Skrypchuk, L., & Donkor, R. (2015). Touchscreen usability and input performance in vehicles under different road conditions. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 47–54. https://doi.org/10.1145/2799250.2799284

Ahmad, B. I., Murphy, J. K., Godsill, S., Langdon, P. M., & Hardy, R. (2017). Intelligent Interactive Displays in Vehicles with Intent Prediction: A Bayesian framework. *IEEE Signal Processing Magazine*, *34*(2), 82–94. https://doi.org/10.1109/MSP.2016.2638699

Akyol, S., & Canzler, U. (2000). Gesture Control for use in Automobiles. *IAPR Workshop on Machine Vision Applications*, 1–4.

Aldridge, L. C., & Lansdown, T. C. (1999). Driver Preferences for Speech Based Interaction with in-Vehicle Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(18), 977–981. https://doi.org/10.1177/154193129904301807

Alexander, C. (1977). *A pattern language: towns, buildings, construction*. Oxford University Press.

Alliance of Automobile Manufacturers (AAM). (2006). Statement of Principles , Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems. *Driver Focus-Telematics Working Group and Alliance of Automobile Manufacturers*.

Alpern, M., & Minardo, K. (2003). Developing a car gesture interface for use as a secondary task. *Extended Abstracts on Human Factors in Computing Systems - CHI '03*, 932. https://doi.org/10.1145/765891.766078

Alvarez, I., Martin, A., Dunbar, J., Taiber, J., Wilson, D. M., & Gilbert, J. E. (2011). Designing Driver-Centric Natural Voice User Interfaces. *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 42–49.

Antin, J. F., Dingus, T. A., Hulse, M. C., & Wierwille, W. W. (1990). An evaluation of the effectiveness and efficiency of an automobile moving-map navigational display. *International Journal of Man-Machine Studies*, *33*(5), 581–594. https://doi.org/10.1016/S0020-7373(05)80054-9

Aslan, I., Krischkowsky, A., Meschtscherjakov, A., Wuchse, M., & Tscheligi, M. (2015). A leap for touch: proximity sensitive touch targets in cars. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 39–46. https://doi.org/10.1145/2799250.2799273

Ba h, K. M., Jæger, M. G., Skov, M. B., & Thomassen, N. G. (2008). You can touch, but you can't look. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, 1139. https://doi.org/10.1145/1357054.1357233

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/SCIENCE.1736359

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Bayle, E., Bellamy, R., Casaday, G., Erickson, T., Fincher, S., Grinter, B., Gross, B., Lehder, D., Marmolin, H., Moore, B., Potts, C., Skousen, G. and Thomas, T. (1998). Putting It All Together: Towards a Pattern Language for Interaction Design Reports. *ACM SIGCHI Bulletin*, *30*(1), 17–23.

Becker, T., Blaylock, N., Gerstenberger, C., Kruijff-Korbayová, I., Korthauer, A., Pinkal, M., … Schehl, J. (2006). Natural and Intuitive Multimodal Dialogue for In-Car Applications: The SAMMIE System. *Frontiers in Artificial Intelligence and Applications*, *141*, 612.

Becker, T., Poller, P., Schehl, J., Blaylock, N., Gerstenberger, C., & Kruijff-korbayov, I. (2006). The SAMMIE System : Multimodal In-Car Dialogue. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, (July), 57–60.

Benoit, C., Martin, J., & Pelachaud, C. (2000). Audio-visual and multimodal speech systems. *Handbook of Standards and Resources for Spoken Language Systems-Supplement*, *500*, 1–95.

Biguer, B., Jeannerod, M., & Prablanc, C. (1982). The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental Brain Research*, *46*(2), 301–304. https://doi.org/10.1007/BF00237188

Biguer, B., Prablanc, C., & Jeannerod, M. (1984). The contribution of coordinated eye and head movements in hand pointing accuracy. *Experimental Brain Research*, *55*(3), 462–469. https://doi.org/10.1007/BF00235277

BMW. (2019). BMW Connected Drive. Retrieved June 26, 2020, from https://www.bmw.de/de/topics/faszination-bmw/connecteddrive/bmw-connected-drive-uebersicht.html

Boles, D. B., Bursk, J. H., Phillips, J. B., & Perdelwitz, J. R. (2007). Predicting Dual-Task Performance With the Multiple Resources Questionnaire (MRQ). *Human Factors*, *49*(1), 32–45. https://doi.org/10.1518/001872007779598073

Bolt, R. a. (1980). "Put-that-there": Voice and Gesture at the Graphics Interface. *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '80*, *14*(3), 262–270. https://doi.org/10.1145/800250.807503

Borchers, J. O., & O., J. (2000). A pattern approach to interaction design. *Proceedings of the Conference on Designing Interactive Systems Processes, Practices, Methods, and Techniques - DIS '00*, 369–378. https://doi.org/10.1145/347642.347795

Bradford, J. H., & H., J. (1995). The Human Factors of Speech-Based Interfaces. *ACM SIGCHI Bulletin*, *27*(2), 61–67. https://doi.org/10.1145/202511.202527

Brand, D., Meschtscherjakov, A., & Büchele, K. (2016). Pointing at the HUD : Gesture Interaction Using a Leap Motion. *Proceedings of the 8th International Conference on*

Automotive User Interfaces and Interactive Vehicular Applications. https://doi.org/10.1145/3004323.3004343

Cairnie, N., Ricketts, I. W., McKenna, S. J., & McAllister, G. (2000). Using finger-pointing to operate secondary controls in automobiles. *Proceedings of the IEEE Intelligent Vehicles Symposium 2000*, 550–555. https://doi.org/10.1109/IVS.2000.898405

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410. https://doi.org/http://doi.acm.org/10.1145/358886.358895

Chakraborty, A., Wha Hong, K., St Amant, R., Zhao, Y.-L., Hong, K., & Kakkaradi, S. (2012). Pointing at responsive objects outdoors. *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, 281–284. https://doi.org/10.1145/2166966.2167018

Chang, Cheng, & Gao, Y. (2016). 3D medical image interaction and segmentation using Kinect. *Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 498–501. https://doi.org/10.1109/ISBI.2016.7493316

Chang, Chun-cheng. (2016). Don ’ t Speak and Drive : Cognitive Workload of In-Vehicle Speech Interactions. *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 99–104.

Chatterjee, I., Xiao, R., & Harrison, C. (2015). Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 131–138. https://doi.org/10.1145/2818346.2820752

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Statistical Power Analysis for the Behavioral Sciences*, Vol. 2nd, p. 567. https://doi.org/10.1234/12345678

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coram, T., & Lee, J. (1996). Experiences: A Pattern Language for User Interface Design. *Proceedings of Joint Pattern Languages of Programs Conferences (PLOP)*, 1–16.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., & Young, R. M. (1995). Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The Care Properties. In *Human-Computer Interaction* (pp. 115–120). https://doi.org/10.1007/978-1-5041-2896-4_19

Crandall, J. M., & Chaparro, A. (2012). Driver Distraction: Effects of Text Entry Methods on Driving Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1693–1697. https://doi.org/10.1177/1071181312561339

Daimler. (2019). Mercedes Benz User Experience. Retrieved June 26, 2021, from https://www.daimler.com/innovation/case/connectivity/mbux.html

Dillard, J. P., & Shen, L. (2005). On the Nature of Reactance and its Role in Persuasive Health Communication. *Communication Monographs*, *72*(2), 144–168. https://doi.org/10.1080/03637750500111815

Donges, E. (2009). *Handbuch Fahrerassistenzsysteme*. https://doi.org/10.1007/978-3-8348-9977-4

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5440 LNCS*, 3–26. https://doi.org/10.1007/978-3-642-00437-7_1

Ecker, R. (2013). *Der verteile Fahrerinteraktionsraum*. Ludwig-Maximilians-Universität München.

Endres, C., Schwartz, T., & Müller, C. (2011). "Geremin": 2D Microgestures for Drivers Based on Electric Field Sensing. *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI '11*, 327–330. https://doi.org/10.1145/1943403.1943457

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2 SPEC. ISS.), 97–120. https://doi.org/10.1016/j.trf.2005.04.012

European Statement of Principles (ESoP). (2008). Commission recommendation of 26 May 2008 on safe and efficient in-vehicle information and communication systems: update of the European Statement of Principles on human-machine interface. *Official Journal of the European Union*.

European Telecommunications Standards Institute (ETSI). (2003). Multimodal interaction, communication and navigation guidelines. *ETSI EG 202 191*, Vol. 1, pp. 1–53. European Telecommunications Standards Institute.

Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology*, *67*(2), 103–112. https://doi.org/10.1037/h0045689

Fogg, B. J. (2009a). A Behavior Model for Persuasive Design. *Proceedings of the 4th International Conference on Persuasive Technology*, 1. https://doi.org/10.1145/1541948.1541999

Fogg, B. J. (2009b). Creating Persuasive Technologies : An Eight-Step Design Process. *Proceedings of the 4th International Conference on Persuasive Technology*, *91*, 1–6. https://doi.org/10.1145/1541948.1542005

Fono, D., & Vertegaal, R. (2005). EyeWindows. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '05*, 151. https://doi.org/10.1145/1054972.1054994

Freeman, E., Brewster, S., & Lantz, V. (2014). Illuminating Gesture Interfaces with Interactive Light Feedback. *Adjunct Proceedings of the 8th Nordic Conference on Human-Computer Interaction*.

Gable, T. M., Raja, S. R., Samuels, D. P., & Walker, B. N. (2015). Exploring and evaluating the capabilities of Kinect v2 in a driving simulator environment. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15*, 297–304. https://doi.org/10.1145/2799250.2799276

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

Garay-Vega, L., Pradhan, A. K., Weinberg, G., Schmidt-Nielsen, B., Harsham, B., Shen, Y., … Fisher, D. L. (2010). Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. *Accident Analysis & Prevention*, *42*(3), 913–920. https://doi.org/10.1016/J.AAP.2009.12.022

Gärtner, U., König, W., & Wittig, T. (2001). Evaluation of manual vs. speech input when using a driver infomation system in real traffic. *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, (August), 7–14.

Geiser G. (1985). Man machine interaction in vehicles. *ATZ*, *87*, 74–77.

Gibbon, D., Mertins, I., & Moore, R. K. (2000). Audio-visual and multimodal speech-based systems. In *Handbook of Multimodal and Spoken Dialogue Systems* (pp. 102–203). https://doi.org/10.1007/978-1-4615-4501-9_2

Gondan, M., Lange, K., Rösler, F., & Röder, B. (2004). The Redundant Target Effect is Affected by Modality Switch Costs. *Psychonomic Bulletin & Review*, *11*(2), 307–313.

Granlund, Å., Lafrenière, D., & Carr, D. a. (2001). A Pattern-Supported Approach to the User Interface Design Process. *HCI International 2001 9th International Conference on Human-Computer Interaction*, *1*, 282–286.

Grasso, M. A., Ebert, D. S., & Finin, T. W. (1998). The integrality of speech in multimodal interfaces. *ACM Transactions on Computer-Human Interaction*, *5*(4), 303–325. https://doi.org/10.1145/300520.300521

Green, P. (1999). The 15-Second Rule for Driver Information Systems. *Proceedings of the ITS America 9th Annual Meeting*. Washington, DC: Intelligent Transportation Society of America.

Green, P., & Paul. (2013). Standard definitions for driving measures and statistics. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, 184–191. https://doi.org/10.1145/2516540.2516542

Gundelsweiler, F. (2008). *Design-Patterns zur Unterstützung der Gestaltung von interaktiven, skalierbaren Benutzungsschnittstellen*.

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention*, *39*(2), 372–379. https://doi.org/10.1016/j.aap.2006.08.013

Harrington, K., Large, D. R., Burnett, G., & Georgiou, O. (2018). Exploring the Use of Mid-Air Ultrasonic Feedback to Enhance Automotive User Interfaces. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '18*, 11–20. https://doi.org/10.1145/3239060.3239089

He, J., Chaparro, A., Nguyen, B., Burge, R. J., Crandall, J., Chaparro, B., … Cao, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? *Accident Analysis and Prevention*, *72*, 287–295. https://doi.org/10.1016/j.aap.2014.07.014

Hua, Z., & Ng, W. L. (2010). Speech recognition interface design for in-vehicle system. *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '10*, (AutomotiveUI), 29. https://doi.org/10.1145/1969773.1969780

Huckauf, A., & Urbina, M. H. (2011). On Object selection in gaze controlled systems. *ACM Transactions on Applied Perception*, *8*(4), 1–14. https://doi.org/10.1145/1870076.1870081

ISO. (2008). *ISO 9241-210 Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems.*

Jacob, R. J. K., & K., R. J. (1991). The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems*, *9*(3), 152–169. https://doi.org/10.1145/123078.128728

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1–2), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

Jex, H. R. (1967). Two Applications of a Critical-Instability Task to Secondary Work Load Research. *IEEE Transactions on Human Factors in Electronics*, (4), 279–282. https://doi.org/10.1109/THFE.1967.234316

Jex, H. R., McDonnell, J. D., & Phatak, A. V. (1966). A "Critical''' Tracking Task for Manual Control Research." *IEEE Transactions on Human Factors in Electronics*, (4), 138–145. https://doi.org/10.1109/THFE.1966.232660

Kamm, C. (1995). User interfaces for voice applications. *Proceedings of the National Academy of Sciences*, *92*(22), 10031–10037. https://doi.org/10.1073/pnas.92.22.10031

Karray, F., Karray, F., Alemzadeh, M., Saleh, J. A., & Arab, M. N. (2008). Human-Computer Interaction: Overview on State of the Art. *International Journal on Smart Sensing and Intelligent Systems*, *1*(1).

Kern, D., Mahr, A., Castronovo, S., Schmidt, A., & Müller, C. (2010). Making use of drivers' glances onto the screen for explicit gaze-based interaction. *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 110–116. https://doi.org/10.1145/1969773.1969792

Koons, D. B., Sparrell, C. J., & Thorisson, K. R. (1998). Integrating simultaneous input from speech, gaze, and hand gestures. In *Readings in Intelligent User Interfaces* (pp. 53–64).

Kousidis, S., Kennington, C., Baumann, T., Buschmeier, H., Kopp, S., & Schlangen, D. (2014). A Multimodal In-Car Dialogue System That Tracks The Driver's Attention. *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, 26–33. https://doi.org/10.1145/2663204.2663244

Kruschitz, C., & Hitz, M. (2010a). Analyzing the HCI design pattern variety. *Proceedings of the 1st Asian Conference on Pattern Languages of Programs - AsianPLoP '10*, 6. https://doi.org/10.1145/2371736.2371745

Kruschitz, C., & Hitz, M. (2010b). Human-Computer Interaction Design Patterns : Structure , Methods , and Tools. *International Journal on Advances in Software*, *3*(1), 225–237.

Kujala, T. (2017). Visual Distraction Effects of In-Car Text Entry Methods – Comparing Keyboard , Handwriting and Voice Recognition. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 1–10.

Landay, J., & Boriello, G. (2003). Design Patterns for Ubiquitous Computing. *James A Landay, Gaetano Boriello*, *36*(8), 93–95.

Large, D. R., Burnett, G., Anyasodo, B., & Skrypchuk, L. (2016). Assessing Cognitive Demand during Natural Language Interactions with a Digital Driving Assistant. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular*

Applications, 67–74.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. *HCI and Usability for Education and Work. USAB 2008. Lecture Notes in Computer Science*, *5298*, 63–76. https://doi.org/https://doi.org/10.1007/978-3-540-89350-9_6

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. *Mensch Und Computer 2006: Mensch Und Computer Im Strukturwandel*, 125–134.

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based Interaction with In-vehicle Computers: The Effect of Speech-based E-mail on Drivers' Attention to the Roadway. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 631–640.

Lee, J., Lee, C., & Kim, G. J. (2017). Vouch: multimodal touch-and-voice input for smart watches under difficult operating conditions. *Journal on Multimodal User Interfaces*, *11*(3), 289–299. https://doi.org/10.1007/s12193-017-0246-y

Lee, S. H., Yoon, S.-O., & Shin, J. H. (2015). On-wheel finger gesture control for in-vehicle systems on central consoles. *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15*, 94–99. https://doi.org/10.1145/2809730.2809739

Lisseman, J., Diwischek, L., Essers, S., & Andrews, D. (2014). In-Vehicle Touchscreen Concepts Revisited: Approaches and Possibilities. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, *7*(1), 2014-01–0266. https://doi.org/10.4271/2014-01-0266

Lumsden, J. (2008). *Handbook of research on user interface design and evaluation for mobile technology*. Information Science Reference.

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, *41*(5), 924–930. https://doi.org/10.1016/J.AAP.2009.05.007

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. a. (2000). Gaze and Speech in Attentive User Interfaces. *Proceedings of the Third International Conference on Multimodal Interfaces*, *1948*(c), 1–7. https://doi.org/10.1007/3-540-40063-X_1

Martin, D., Rodden, T., Rouncefield, M., Sommerville, I., & Viller, S. (2001). Finding Patterns in the Fieldwork. In *ECSCW 2001* (pp. 39–58). https://doi.org/10.1007/0-306-48019-0_3

May, K. R., Gable, T. M., & Walker, B. N. (2014). A Multimodal Air Gesture Interface for In Vehicle Menu Navigation. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*, 1–6. https://doi.org/10.1145/2667239.2667280

Mayer, S., Le, H. V., Nesti, A., Henze, N., Bülthoff, H. H., & Chuang, L. L. (2018). The Effect of Road Bumps on Touch Interaction in Cars. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '18*, 85–93. https://doi.org/10.1136/bmj.327.7422.1047-f

Mayer, S., Schwind, V., Schweigert, R., & Henze, N. (2018). The Effect of Offset Correction and Cursor on Mid-Air Pointing in Real and Virtual Environments. *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3173574.3174227

Mayer, S., Wolf, K., Schneegass, S., & Henze, N. (2015). Modeling Distant Pointing for Compensating Systematic Displacements. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *1*(3606), 4165–4168. https://doi.org/10.1145/2702123.2702332

McNeill, D. (1992). Guide to Gesture Classification, Transcription and Distribution. *Hand and Mind: What Gestures Reveal about Thought*, pp. 75–104. https://doi.org/10.2307/1576015

Mehler, B., Kidd, D., Reimer, B., Reagan, I., Dobres, J., & McCartt, A. (2016). Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems. *Ergonomics*, *59*(3), 344–367. https://doi.org/10.1080/00140139.2015.1081412

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(3), 396–412. https://doi.org/10.1177/0018720812442086

Miranda, B., Jere, C., Alharbi, O., Lakshmi, S., Khouja, Y., & Chatterjee, S. (2013). Examining the efficacy of a persuasive technology package in reducing texting and driving behavior. *PERSUASIVE 2013LNCS*, *7822*, 137–148. https://doi.org/10.1007/978-3-642-37157-8_17

Mirnig, A. G., Kaiser, T., Lupp, A., Perterer, N., Meschtscherjakov, A., Grah, T., & Tscheligi, M. (2016). Automotive User Experience Design Patterns : An Approach and Pattern Examples. *International Journal on Advances in Intelligent Systems*, *9*(3), 275–286.

Mitrevska, M., Moniri, M. M., Nesselrath, R., Schwartz, T., Feld, M., Korber, Y., … Muller, C. (2015). SiAM - Situation-Adaptive Multimodal Interaction for Innovative Mobility Concepts of the Future. *Proceedings of the International Conference on Intelligent Environments - IE '15*, 180–183. https://doi.org/10.1109/IE.2015.39

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Müller, C., Weinberg, G., & Vetro, A. (2011). Multimodal input in the car, today and tomorrow. *IEEE Multimedia*, *18*(1), 98–103. https://doi.org/10.1109/MMUL.2011.14

Neggers, S. F., & Bekkering, H. (2001). Gaze anchoring to a pointing target is present during the entire pointing movement and is driven by a non-visual signal. *Journal of Neurophysiology*, *86*(2), 961–970.

Neßelrath, R., Moniri, M. M., & Feld, M. (2016). Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. *International Conference on Intelligent Environments*, 190–193. https://doi.org/10.1109/IE.2016.42

Neuss, R. (2001). Usability Engineering als Ansatz zum Multimodalen Mensch-Maschine-Dialog. Technische Universität München.

NHTSA, Campbell, J. L., Brown, J. L., Graving, J. S., Richard, C. M., Lichty, M. G., … Divekar, G. (2016). *Human Factors Design Guidance for Driver-Vehicle Interfaces (DVI)*.

Nigay, L., & Coutaz, J. (1993). A design space for multimodal systems: concurrent processing and data fusion. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 172–178. https://doi.org/10.1145/169059.169143

Norman, D. (2010). Natural user interfaces are not natural. *Interactions*, *17*(3), 6. https://doi.org/10.1145/1744161.1744163

Oh, A., Fox, H., Van Kleek, M., Adler, A., Gajos, K., Morency, L.-P., & Darrell, T. (2002). Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment. *Extended Abstracts on Human Factors in Computing Systems*, 650. https://doi.org/10.1145/506443.506528

Ohn-Bar, E., Tran, C., & Trivedi, M. (2012). Hand gesture-based visual user interface for infotainment. *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12*, 111. https://doi.org/10.1145/2390256.2390274

Ohn-Bar, E., & Trivedi, M. M. (2014). Hand Gesture Recognition in Real Time for Automotive Interfaces: A Multimodal Vision-Based Approach and Evaluations. *IEEE Transactions on Intelligent Transportation Systems*, *15*(6), 2368–2377. https://doi.org/10.1109/TITS.2014.2337331

Ortega, J., Barker, C., Wilson, C., & Kruse, R. (1987). An Interactive, Reconfigurable Display System for Automotive Instrumentation. *IEEE Transactions on Consumer Electronics*, *CE-33*(1), xi–13. https://doi.org/10.1109/TCE.1987.290190

Oviatt, S., & Jacko, J. (2012). Multimodal Interfaces. In *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (Third Edit, pp. 405–423).

Oviatt, Sharon. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, *12*(1–2), 93–129. https://doi.org/10.1207/s15327051hci1201&amp;2_4

Oviatt, Sharon. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, *42*(11), 74–81. https://doi.org/10.1145/319382.319398

Oviatt, Sharon. (2006). Human-centered design meets cognitive load theory. *Proceedings of the 14th Annual ACM International Conference on Multimedia - MULTIMEDIA '06*, 871. https://doi.org/10.1145/1180639.1180831

Oviatt, Sharon. (2012). Multimodal Interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (14th ed., pp. 286–304).

Oviatt, Sharon, & Cohen, P. (2000). Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM*, *43*(3), 45–53. https://doi.org/10.1145/330534.330538

Oviatt, Sharon, Cohen, P., Wu, L., Suhm, B., Bers, J., Winograd, T., & Landay, J. (2009). 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*(May 2014), 37–41. https://doi.org/10.1207/S15327051HCI1504

Oviatt, Sharon, Coulston, R., & Lunsford, R. (2004). When do we interact multimodally?: cognitive load and multimodal communication patterns. *Proceedings of the 2004*

International Conference on Multimodal Interfaces - ICMI '04, 129–136. https://doi.org/http://doi.acm.org/10.1145/1027933.1027957

Oviatt, Sharon, DeAngeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '97*, 415–422. https://doi.org/10.1145/258549.258821

Parush, A. (2005). Speech-Based Interaction in Multitask Conditions: Impact of Prompt Modality. *Human Factors*, *47*(3), 591–597. https://doi.org/10.1518/001872005774860041

Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, *2*(4), 315. https://doi.org/10.1049/iet-its:20080023

Pe ina, M., & Bojani , I. (2003). Forearm and Hand. In *Overuse Injuries of the Musculoskeletal System, Second Edition* (pp. 107–119). https://doi.org/10.1201/b14243-8

Peissner, M., & Doebler, V. (2011). Can voice interaction help reducing the level of distraction and prevent accidents ? *Carnegie Mellon University*, (May 2011), 24.

Petzoldt, T., Bellem, H., & Krems, J. F. (2014). The Critical Tracking Task: A Potentially Useful Method to Assess Driver Distraction? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(4), 789–808. https://doi.org/10.1177/0018720813501864

Pfleging, B., Fekety, D., Schmidt, A., & Kun, A. L. (2016). A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5776–5788. https://doi.org/10.1145/2858036.2858201

Pfleging, B., Kienast, M., & Schmidt, A. (2011). *SpeeT : A Multimodal Interaction Style Combining Speech and Touch Interaction in Automotive Environments*. (November), 2–3.

Pfleging, B., Schneegass, S., & Schmidt, A. (2012). Multimodal Interaction in the Car - Combining Speech and Gestures on the Steering Wheel. *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12*, (c), 0–7. https://doi.org/10.1145/2390256.2390282

Pickering, C. A., Burnham, K. J., & Richardson, M. J. (2007). A research study of hand gesture recognition technologies and applications for human vehicle interaction. *Institution of Engineering and Technology Conference on Automotive Electronics*, 1–15.

Pickering, C. a. C. a., Burnham, K. J. K. J., & Richardson, M. J. M. J. (2007). A Review of Automotive Human Machine Interface Technologies and Techniques to Reduce Driver Distraction. *2nd IET International Conference on System Safety*, 223–228. https://doi.org/10.1049/cp:20070468

Pickering, J. A. (1986). Touch-sensitive screens: the technologies and their application. *International Journal of Man-Machine Studies*, *25*(3), 249–269. https://doi.org/10.1016/S0020-7373(86)80060-8

Pieraccini, R., Dayanidhi, K., Bloom, J., Dahan, J., Phillips, M., International, S., … Co, F. M. (2003). A Multimodal Conversational Interface for a Concept Vehicle. *The New School Psychology Bulletin*, *1*(1), 9–24. https://doi.org/10.1037/e741632011-002

Pieraccini, R., Dayanidhi, K., Bloom, J., Dahan, J., Phillips, M., International, S., … Co, F. M. (2004). Multimodal Conversational Systems for Automobiles. *Communications of the ACM*, *47*(1), 47–49. https://doi.org/10.1037/e741632011-002

Plaumann, K., Weing, M., Winkler, C., Müller, M., & Rukzio, E. (2017). Towards accurate cursorless pointing: the effects of ocular dominance and handedness. *Personal and Ubiquitous Computing*, 1–14. https://doi.org/10.1007/s00779-017-1100-7

Poitschke, T. (2011). Blickbasierte Mensch-Maschine Interaktion im Automobil. Technische Universität München.

Poitschke, T., Laquai, F., Stamboliev, S., & Rigoll, G. (2011). Gaze-based interaction on multiple displays in an automotive environment. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 543–548. https://doi.org/10.1109/ICSMC.2011.6083740

Printz, L. (2017). Infotainment systems? That's so 1986. Retrieved February 5, 2020, from www.hegerty.com website: https://www.hagerty.com/articles-videos/articles/2017/12/27/1986-buick-riviera-gcc-touchscreen

Ramm, S., Giacomin, J., Robertson, D., & Malizia, A. (2014). A First Approach to Understanding and Measuring Naturalness in Driver-Car Interaction. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*, 1–10. https://doi.org/10.1145/2667317.2667416

Ratzka, A. (2008). Steps in identifying interaction design patterns for multimodal systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5247 LNCS*, 58–71. https://doi.org/10.1007/978-3-540-85992-5-5

Ratzka, A. (2013). User interface patterns for multimodal interaction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7840*, 111–167. https://doi.org/10.1007/978-3-642-38676-3-4

Ratzka, A., & Wolff, C. (2006). A Pattern-Based Methodology for Multimodal Interaction Design. *International Conference on Text, Speech and Dialogue*, 677–686. https://doi.org/10.1007/11846406_85

Reeves, L. M., Martin, J.-C., McTear, M., Raman, T., Stanney, K. M., Su, H., … Kraal, B. (2004). Guidelines for multimodal user interface design. *Communications of the ACM*, *47*(1), 57–59. https://doi.org/10.1145/962081.962106

Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2013). The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance. *MIT AgeLab Technical Report No. 2013-17A. Massachusetts Institute of Technology, Cambridge, MA.*, 293. https://doi.org/2013-18A

Riener, A. (2012). Gestural Interaction in Vehicular Applications. *Computer*, *45*(4), 42–47. https://doi.org/10.1109/MC.2012.108

Riener, A., Ferscha, A., Bachmair, F., Hagmüller, P., Lemme, A., Muttenthaler, D., … Weger, F. (2013). Standardization of the in-car gesture interaction space. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular*

Applications (AutomotiveUI '13). https://doi.org/10.1145/2516540.2516544

Riener, A., & Wintersberger, P. (2011). Natural, intuitive finger based input as substitution for traditional vehicle control. *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11*, (August 2014), 159. https://doi.org/10.1145/2381416.2381442

Roider, F., & Gross, T. (2018). I See Your Point : Integrating Gaze to Enhance Pointing Gesture Accuracy While Driving. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*, 351–358. https://doi.org/10.1145/3239060.3239084

Roider, F., & Raab, K. (2018). Implementation and Evaluation of Peripheral Light Feedback for Mid-Air Gesture Interaction in the Car. *International Conference on Intelligent Environments (IE)*, 87–90. IEEE.

Roider, F., Reisig, L., & Gross, T. (2018). Just Look : The Benefits of Gaze-Activated Voice Input in the Car. *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*, 210–214.

Roider, F., Rümelin, S., Pfleging, B., & Gross, T. (2017). The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 94–102. https://doi.org/10.1145/3122986.3122999

Roider, F., Rümelin, S., Pfleging, B., & Gross, T. (2019). Investigating the Effects of Modality Switches on Driver Distraction and Interaction Efficiency in the Car. *Journal of Multimodal User Interfaces*, 1–9. https://doi.org/10.1007/s12193-019-00297-9

Rümelin, S., Gabler, T., & Bellenbaum, J. (2017). Clicks are in the Air: How to Support the Interaction with Floating Objects through Ultrasonic Feedback. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '17*, 103–108. https://doi.org/10.1145/3122986.3123010

Rümelin, S., Marouane, C., & Butz, A. (2013). Free-hand pointing for identification and interaction with distant objects. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, 40–47. https://doi.org/10.1145/2516540.2516556

Salvucci, D. D., & Anderson, J. R. (2000). Intelligent gaze-added interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '00*, 273–280. https://doi.org/10.1145/332040.332444

Sanderson, C., & Paliwal, K. K. (2002). Information Fusion and Person Verification Using Speech & Face Information. *Digital Signal Processing*, *14*(5), 449–480. https://doi.org/10.1016/j.dsp.2004.05.001

Sas, C., Whittaker, S., Dow, S., & Forlizzi, J. (2014). Generating Implications for Design through Design Research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1971–1980. https://doi.org/10.1145/2556288.2557357

Schneegaß, S., Pfleging, B., Kern, D., & Schmidt, A. (2011). Support for modeling interaction with automotive user interfaces. *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11*, 71. https://doi.org/10.1145/2381416.2381428

Schnelle-Walka, D., Duarte, C., & Radomski, S. (2016). Multimodal Fusion and Fission within the W3C MMI Architectural Pattern. In *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything* (pp. 1–422). https://doi.org/10.1007/978-3-319-42816-1

Schrepp, M., Hinderks, A., Thomaschewski, J., & Germany, S. A. P. A. G. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *Ternational Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

Seffah, A. (2010). The evolution of design patterns in HCI: From pattern languages to pattern-oriented design. *Proceedings of the 1st International Workshop on Pattern-Driven Engineering of Interactive Computing Systems*, (October), 4–9. https://doi.org/10.1145/1824749.1824751

Sezgin, T. M., Davies, I., & Robinson, P. (2009). Multimodal Inference for Driver-Vehicle Interaction. *Proceedings of the 2009 International Conference on Multimodal Interfaces - ICMI '09*, 193–197. https://doi.org/10.1145/1647314.1647348

Shakeri, G., Williamson, J. H., & Brewster, S. (2017). Novel Multimodal Feedback Techniques for In-Car Mid-Air Gesture Interaction. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '17*, 84–93. https://doi.org/10.1145/3122986.3123011

Shakeri, G., Williamson, J. H., & Brewster, S. (2018). May the Force Be with You: Ultrasound Haptic Feedback for Mid-Air Gesture Interaction in Cars. *Proceedings of the 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '18*, *47*(6), 673–674. https://doi.org/10.1016/j.devcel.2018.11.041

Sharma, R., Pavlovic, V. I., & Huang, T. S. (2002). Toward Multimodal Human–Computer Interface. In *Advances in Image Processing and Understanding* (pp. 349–365). https://doi.org/10.1142/9789812776952_0014

Shi, Y., Park, T., Ruiz, N., Taib, R., Choi, E. H. C., & Chen, F. (2007). Galvanic Skin Response (GSR) as an Index of Cognitive Load. *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2651–2656. https://doi.org/10.1145/1240866.1241057

Society of Automotive Engineers. (2000). Definitions and experimental measures related to the specification of driver visual behavior using video based techniques. *SAE J2396*.

Society of Automotive Engineers. (2015). *Navigation and Route Guidance Function Accessibility While Driving - SAE International*. https://doi.org/https://doi.org/10.4271/J2364_201506

Stanney, K., Samman, S., Reeves, L., Hale, K., Buff, W., Bowers, C., … Lackey, S. (2004). A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations. *International Journal of Human-Computer Interaction*, *17*(2), 229–257. https://doi.org/10.1207/s15327590ijhc1702_7

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance: New developments and findings. *Journal of Psychology*, *223*(4), 205–214. https://doi.org/10.1027/2151-2604/a000222

Strayer, D. L., Drews, F. A., & Crouch, D. J. (2006). A comparison of the cell phone driver and the drunk driver. *Human Factors*, *48*(2), 381–391. https://doi.org/10.1518/001872006777724471

Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive Distraction While Multitasking in the Automobile. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 54). https://doi.org/10.1016/B978-0-12-385527-5.00002-4

Suhm, B., Myers, B., & Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, *8*(1), 60–98. https://doi.org/10.1145/371127.371166

Teichner, W. H., & Krebs, M. J. (1974). Laws of visual choice reaction time. *Psychological Review*, *81*(1), 75–98. https://doi.org/10.1037/h0035867

Tidwell, J. (1997). A pattern language for human-computer interface design. *Available via DIALOG*. https://doi.org/2011-15904-001 [pii]\r10.1037/a0024376

Tidwell, J. (2010). *Designing Interfaces: Patterns for Effective Interaction Design*. O'Reilly Media, Inc.

Tsimhoni, O., Smith, D., & Green, P. (2004). Address Entry While Driving: Speech Recognition Versus a Touch-Screen Keyboard. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 600–610. https://doi.org/10.1518/hfes.46.4.600.56813

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*(1), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Van Welie, M., & Traetteberg, H. (2000). Interaction Patterns in User Interfaces. *Proceeding of the 7th Pattern Languages of Programs Conference*, 13–16.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '01*, 301–308. https://doi.org/10.1145/365024.365119

Vilimek, R., Hempel, T., & Otto, B. (2007). Multimodal Interfaces for In-Vehicle Applications. *International Conference on Human-Computer Interaction*, 216–224. https://doi.org/10.1007/978-3-540-73110-8_23

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Welie, M. Van, & Van Der Veer, G. (2003). Pattern Languages in Interaction Desing: Structure and Organization. *Proceedings of INTERACT2003*.

Wickens, C. D. (1980). The Structure of Attentional Resources. *Attention and Performance VIII*, *8*, 239–257.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/DOI 10.1080/1463922021012380 6

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors*, *25*(2), 227–248. https://doi.org/10.1177/001872088302500209

Wierwille, W. (1993). Visual and manual demands of in-car controls and displays. In *Automotive ergonomics* (pp. 299–320).

Wierwille, W. W. (1993). Demands on driver resources associated with introducing advanced technology into the vehicle. *Transportation Research Part C: Emerging Technologies*, *1*(2), 133–142. https://doi.org/10.1016/0968-090X(93)90010-D

Wong, N., & Gutwin, C. (2010). Where are you pointing? The accuracy of deictic pointing in CVEs. *Proceedings of the 28th of the International Conference on Human Factors in Computing Systems - CHI '10*, 1029–1038. https://doi.org/10.1145/1753326.1753480

Wu, R., Rossos, P., Quan, S., Reeves, S., Lo, V., Wong, B., … Morra, D. (2011). An evaluation of the use of smartphones to communicate between clinicians: a mixed-methods study. *Journal of Medical Internet Research*, *13*(3), e59. https://doi.org/10.2196/jmir.1655

Xiao, B., Girand, C., & Oviatt, S. (2002). Multimodal Integration Patterns in Children. *Proceedings of International Conference on Spoken Language Processing*, 629–632. https://doi.org/10.1.1.4.6072

Xiao, B., Lunsford, R., Coulston, R., Wesson, M., & Oviatt, S. (2003). Modeling multimodal integration patterns and performance in seniors. *Proceedings of the 5th International Conference on Multimodal Interfaces - ICMI '03*, 265. https://doi.org/10.1145/958432.958480

Yankelovich, N. (1996). How Do Users Know What to Say? *Interactions*, *3*(6), 32–43. https://doi.org/http://dx.doi.org/10.1145/242485.242500

Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 246–253. https://doi.org/10.1145/302979.303053

Zhang, Y., Stellmach, S., & Sellen, A. (2015). The Costs and Benefits of Combining Gaze and Hand Gestures for Remote Interaction. *International Conference on Human-Computer Interaction*, Vol. 1, pp. 570–577. https://doi.org/10.1007/978-3-319-22698-9

Zinchenko, K., Wu, C.-Y., & Song, K.-T. (2017). A Study on Speech Recognition Control for a Surgical Robot. *IEEE Transactions on Industrial Informatics*, *13*(2), 607–615. https://doi.org/10.1109/TII.2016.2625818