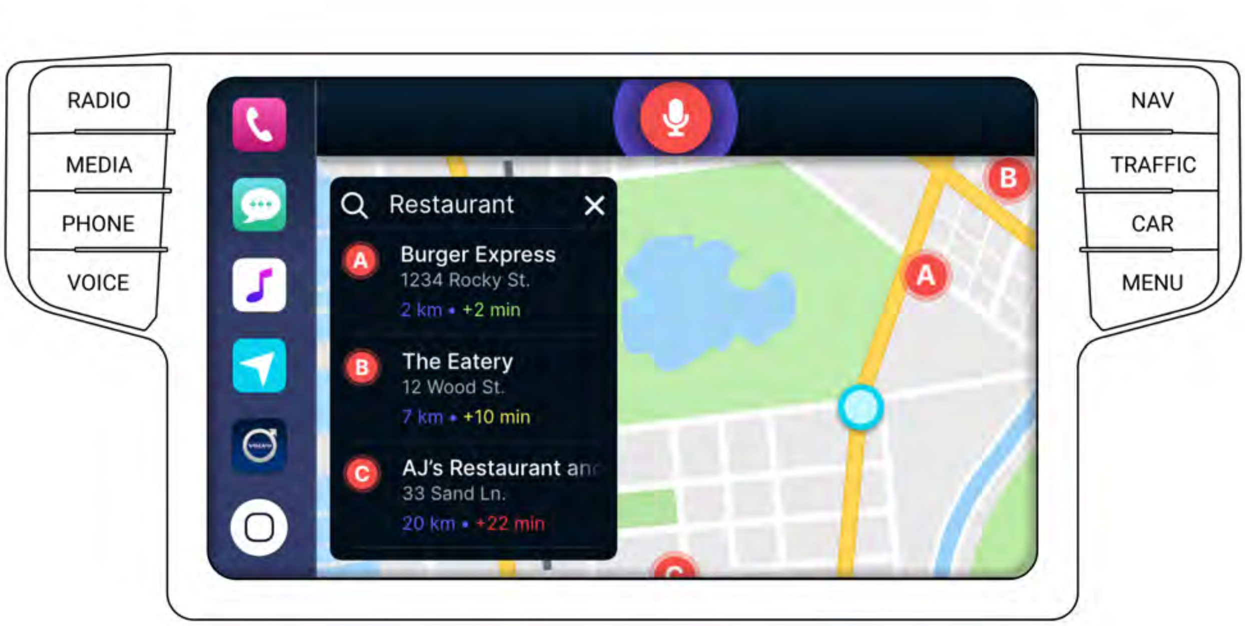
CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG



Designing in-vehicle voice assistants

Creating safer, integrated driver experiences

Master's thesis in Computer science and engineering

CONNIE (KHANH) NGUYEN AND WILLIAM FALKENGREN

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Designing in-vehicle voice assistants

Creating safer, integrated driver experiences

CONNIE (KHANH) NGUYEN AND WILLIAM FALKENGREN

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Designing in-vehicle voice assistants
Creating safer, integrated driver experiences
CONNIE (KHANH) NGUYEN AND WILLIAM FALKENGREN

Supervisor: Fang Chen, Department of Computer Science and Engineering
Advisor: Jenny Wilkie, Volvo Cars
Examiner: Staffan Björk & Olof Torgersson , Department of Computer Science and Engineering

Master's Thesis 2019
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: In-vehicle infotainment system with prototype voice assistant interface.

Typeset in LaTeX
Gothenburg, Sweden 2019

Designing in-vehicle assistants
Creating safer, integrated driver experiences
CONNIE (KHANH) NGUYEN AND WILLIAM FALKENGREN
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Voice assistants are increasing in popularity with the rise of devices like smart speakers and screens. As people grow accustomed to using these assistants, it is likely they would want use the same voice assistant in their car. Many modern cars already support integration of voice assistants from both Apple and Google. In this project, voice assistants integrated into the vehicle and their effects on safety in terms of increased diverted attention and cognitive load are examined. Current voice assistants are also reviewed. Apple Siri and Google Assistants, two commercial voice assistants, are evaluated under the conditions of manual driver, as well as with longitudinal and lateral assistive drive features. New, improved design solutions and guidelines were evaluated through two prototypes with different approaches to solving found problems in existing voice assistants. The results indicate several similarities and differences in the existing design guidelines for the different voice assistants. Users provide input and thoughts about the existing solutions. New design solutions for decreasing distraction and cognitive load are presented. These new solutions can help continued research and further improvement of voice assistants within cars in the future to come.

Keywords: Voice Assistant, Voice Interaction, Driving, Safety, Attention, Cognitive Load, Design Guidelines.

Acknowledgements

We would like to thank all personnel of Volvo Cars who in any way participated and helped with the project. We would like to especially thank Jenny Wilkie for her expertise and all her supervision, feedback and guidance throughout the project. We would like to thank all test participants who participated in our studies. We would like to thank Chalmers for providing us with material and equipment and lastly, we would like to thank Fang Chen for her supervision throughout the project.

Connie (Khanh) Nguyen & William Falkengren, Gothenburg, June 2019

Abbreviations

NLP - Natural Language Processing
IVI - In-Vehicle Infotainment
HMI - Human-Machine Interface
VUI - Voice User Interface
ADS - Autonomous Driving System
NHTSA - National Highway Traffic Safety Administration
SAE - Society of Automotive Engineers
PA - Pilot Assist
VA - Voice Assistant
CSD - Center-Stack Display
DIM - Driver Information Module
AA - Android Auto
AC - Apple CarPlay

Contents

List of Figures

List of Tables

1

Introduction

Voice assistants have exploded in popularity in recent years thanks to smart speakers. *Natural Language Processing* (NLP) allows these smart speakers to communicate with their users in a convenient and natural way and makes them suitable for helping their users with a large and varied set of tasks. One report predicts that 47% of American homes will have a smart speaker by 2022 [34]. As people grow accustomed to the voice assistants in their homes and on their phones, it is not unreasonable to assume that drivers will use the same voice assistant from their daily lives in their cars. This possibility is in fact a reality in many modern cars that offer voice assistant integration directly into the in-vehicle infotainment system (IVI) through Android Auto or Apple CarPlay. Voice assistants, and voice interaction at large, offer drivers an eyes-free, hands-free way to complete secondary tasks while driving. However, voice assistants are susceptible to recognition issues and have transient, paced interaction flows that require immediate response from the driver. Despite the integration of voice assistants into vehicles, set guidelines for safe voice interaction are not well defined. As voice assistant integrated IVIs become increasingly prevalent, it is necessary to evaluate the existing commercially available voice assistant integrated IVIs in relation to distracted driving.

1.1 Purpose

Voice interaction in vehicles have long been a topic of research [25]. However, many of the previous studies on voice interaction in vehicles focus on evaluating voice interactions developed by the OEM, such as the Chevrolet MyLink and the Volvo Sensus [28, 39]. Generally, voice interaction has been found to be a safer alternative to standard HMI inputs to the IVI [25]. However, the landscape of voice interaction in vehicles is expanding with voice assistants developed by software giants including Google and Apple. These voice assistants have seldom been examined in relation to distracted driving and have yet to be studied in a setup where they are fully integrated in an IVI. A deeper understanding of voice assistants' effect on distracted driving is critical as voice assistants increasingly integrate into available IVIs.

Despite the heralded safety benefits of voice interaction, standards for safe voice interaction in vehicles are largely undefined. The National Highway Traffic Safety Administration (NHTSA), which has set guidelines and standards for safe manual interactions with IVIs, has yet to publish similar guidelines for voice interaction [30, 31]. With many voice interaction systems already commercially available, designers cannot continue to put off considerations for safe voice interactions while driving. Voice assistants in particular introduce the possibility for third-party designers and developers to create and distribute in-vehicle applications, such as navigation apps. Today's guidelines treat the design of the IVI and voice assistant as two separate entities rather than one integrated voice-driven, multimodal experience [2, 14, 13].

As an added layer of consideration is the development of autonomous vehicles in parallel to voice assistants. Society of Automotive Engineers (SAE) Level 2 autonomous driving offers drivers support in the primary task of driving through features such as adaptive cruise control and lane keeping [41]. However, a misunderstanding of how these systems work and their limitations may lead to overly trusting or relying on these support systems, thereby causing drivers to divert their attention from the primary task to a secondary one. The affect of voice assistants to complete secondary tasks in combination with driver support systems should also be considered for safer interaction.

1.2 Aim

The primary aim of this project is to improve voice assistant interactions to complete secondary tasks without compromising driver safety. Thus, it is necessary to assess the current state of the art of voice assistant integrated IVIs commercially available today and the design patterns they employ. These systems will be assessed in relation to their effect on distracted driving, which includes visual distraction and cognitive load. While voice interaction can reduce visual distraction, it has some possible drawbacks. It is transient, meaning it is non-persistent, and it can potentially increase the cognitive load of completing secondary tasks, such as mentally visualizing navigation instructions in an unfamiliar area. Both visual distraction and cognitive load must be carefully balanced to create what may be considered a safe interaction while driving. This project also aims to produce design guidelines for multimodal voice assistant-driven interactions for performing secondary tasks.

1.2.1 Scope

This project is limited to the performance of secondary tasks using a voice assistant in a vehicle equipped with a voice assistant integrated IVI. Guidelines for currently existing integrated voice assistants will be evaluated and a new set of guidelines will be suggested. The new suggested guidelines will consist of currently existing guidelines as well as new guidelines developed in this project. The project is further delimited to situations with a driver using a Level 2 or lower autonomous passenger vehicle with no additional passengers. The drivers are defined as civilian drivers, people who may drive as part of their daily commute, but not those who drive for extended periods of time as part of their profession, such as a taxi cab driver or a cargo trucker. As such, the design guidelines that will be produced as a result of this project may be limited to driving scenarios that also match the scope of the project. As this project considers the current state of voice assistants in vehicles, the produced guidelines would be directly applicable for the near future.

As autonomous driving improves, reaching high or full level of automation, the types of tasks users will perform in the vehicle will likely shift to be more entertainment focused. However, even with the advent of fully autonomous vehicles, complete market adoption of such vehicles will not happen overnight. Thus the guidelines produced by this project will remain relevant for the remaining vehicles on the road that are Level 2 and under.

1.2.2 Stakeholders

This project is carried out as part of a larger project known as SEER (Seemless, Efficient and Enjoyable user-vehicle inteRaction). SEER is a joint collaboration between Volvo Cars, Volvo Technology, RISE Viktoria, and Semcon; the project is funded by Vinnova [44]. The SEER project is focused on improving the experience of completing secondary tasks in low-level autonomous vehicles (up to SAE level 2). General findings and projects developed under the umbrella of SEER are available to the public to promote knowledge sharing and innovation in the automotive industry.

1.2.3 Ethical Concerns

This project will use on-road tests to assess the current state of voice assistant integrated IVIs. In such testing conditions, test participant safety is paramount and shall take precedent over the test itself. Additional measures, such as using a specially equipped vehicle for facilitator interjection may be necessary in the interest of safety.

An additional concern is the handling of personal user data. The commercially available voice assistants send and retrieve data to and from external servers owned by parties outside of this project, such as the voice assistant author company and third-party app services. Also concerning personal user data is the collection of data as video footage. Video footage collected as part of this project was done so with full consent from the test participants, where participants also had the option to have any personally identifying footage removed once the collected data was analyzed.

1.3 Research Questions

In the context of SAE Level 2 (and lower) vehicles, this project addresses the following questions:

1. What adjustments to existing NLP-based voice assistant design guidelines should be made for safer interaction while driving?

    (a) What existing design guidelines and patterns are implemented in voice assistant integrated infotainment systems?

    (b) What improvements to existing voice assistants can be made to minimize diverted attention from the primary task of driving?

    (c) What improvements to existing voice assistants can be made to minimize cognitive load while executing a secondary task during the primary task of driving?

In the primary research question, safer interactions are defined with respect to how well they comply to NHTSA design guidelines for human-machine interfaces (HMIs) and reduce distracted driving [30, 31]. While the NHTSA guidelines explicitly do not consider voice interaction, they may serve as a starting point as the infotainment systems examined are multimodal. Moreover, NHTSA has yet to define safe interactions with respect to voice interaction, though it has plans to do so in the future. The current NHTSA guidelines related to this project are covered in section 2.4 and 2.5.1.

2

Background

This project delves into several areas including voice interaction, autonomous cars, and distracted driving. This chapter provides a brief history of each area and previous scientific work researching the intersection of all three.

2.1 Voice Interaction

The field of voice interaction has experienced a recent increase in interest thanks to the introduction of smart speakers to market. However, voice interaction long predates smart speakers and early interactive voice systems were first introduced in the 1990s [22]. These early systems were known as finite state *voice user interfaces* (VUIs). VUIs are typically categorized as either *finite state* or *natural language processing* (NLP), but hybrids do exist [22].

Finite state VUIs are characterized by a limited set of commands for each point in the interaction flow, typically in a tree menu [22]. Most people encounter finite state VUIs on the phone, in the form of automated customer service systems. These systems are usually met with frustration as many users have difficulty finding the information or action they want in a tree menu.

Natural language processing VUIs improve upon their finite state predecessors by recognizing a wider array of user input for the same action through statistical language modeling. Prime examples of NLP VUIs are the voice assistants available on smartphones and smart speakers. Voice assistants typically process voice input off-site through cloud-computing. The most popular voice assistants include Amazon Alexa, Google Assistant, Apple Siri, and Microsoft Cortana. These voice assistants allow users to interact with them in a more natural, conversational manner. Moreover, thanks to their off-site processing, voice assistants can learn and improve over time as more users interact with them [22].

2.1.1 Voice Assistants

Voice assistants take NLP VUIs to the next level. Not only can they understand and respond to conversational input, they can use information about the user to provide relevant responses. For example, users can ask a voice assistant, "What are upcoming Robyn concerts?" and the assistant can respond with upcoming concert dates in the user's city with the option to hear about concerts in other cities. However, not all tasks completed through a voice assistant take advantage of this contextual information and may require users to repeat information, introducing frustration into the input process.

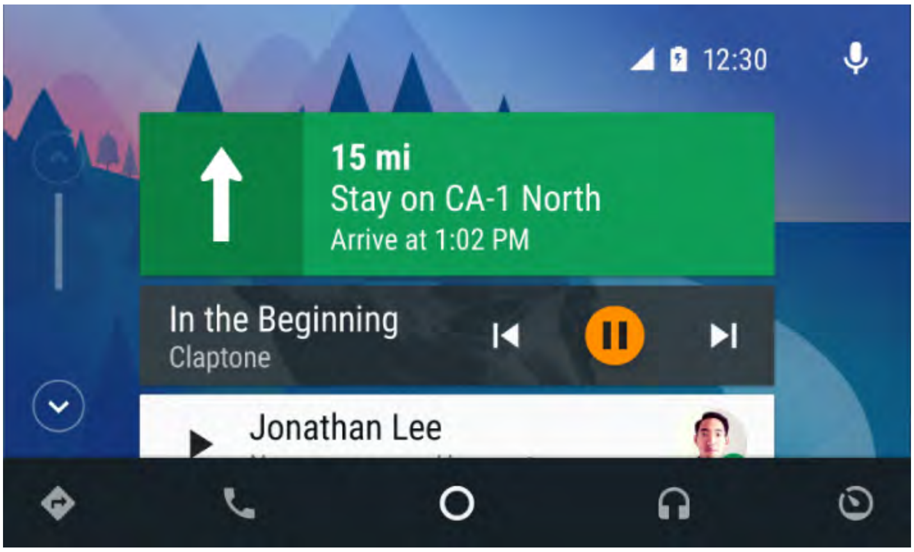


Figure 2.1: The Android Auto GUI

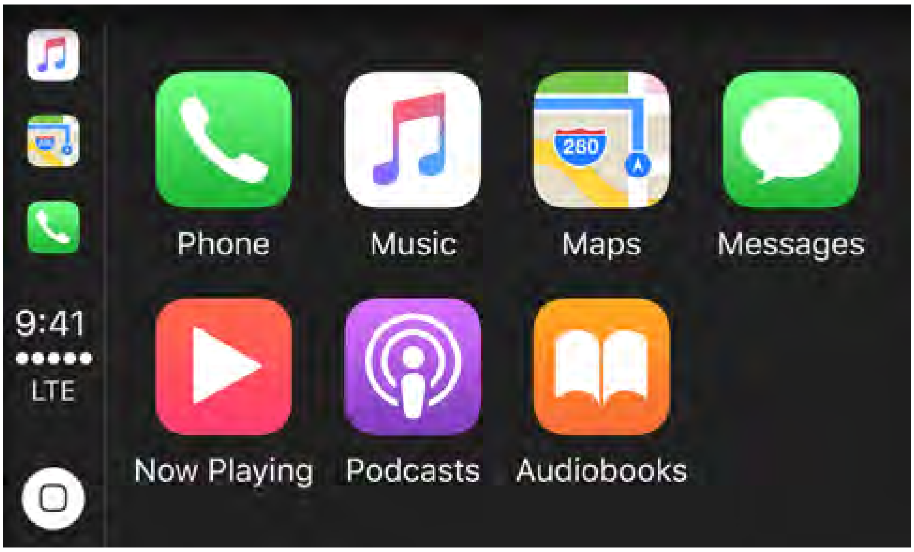


Figure 2.2: The CarPlay GUI

Voice assistants have recently made the leap from the phone to the car through IVI integration interfaces. Integration interfaces allow drivers to connect their smartphone to their car's IVI, enabling drivers to access some of the functionality of their phones directly on the IVI, including voice assistants. Two platforms that offer IVI integration interfaces are Google and Apple. Google's Android Auto allows drivers to connect their Android phone and Google Assistant to the IVI. The Android Auto home screen can be seen in figure 2.1. Similarly, Apple CarPlay allows iOS users to integrate their iPhone and Siri assistant into the IVI. The Apple CarPlay home screen can be seen in figure 2.2. Both Android Auto and Apple CarPlay enable drivers to use select apps from their phone on the IVI. Only apps which belong to an enabled category and have been developed for in-vehicle use may be available on the IVI. Integration interface authors, such as Apple and Google, dictate which categories may be enabled.

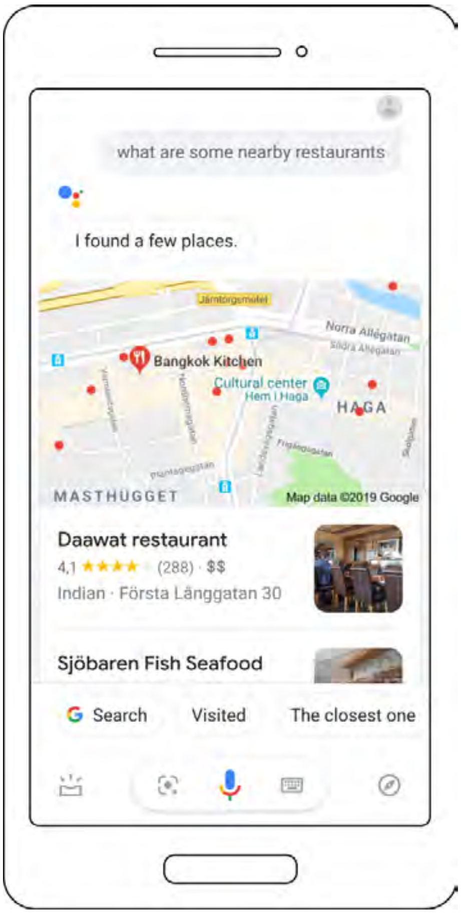
Categories enabled on both Android Auto and Apple CarPlay are communication, navigation, audio, and automaker [2, 13]. The communication category includes apps with messaging and VoIP calling features. Navigation apps allow drivers to locate points of interest and provide driving directions. Audio apps cover an array of audio services which include music streaming services, podcast stations, and sports news. Automaker apps allow drivers to get information about their car and adjust car settings through the integration interface. If a driver's voice assistant is enabled on the phone, then enabled apps may be used via voice assistant. However, the degree of voice interaction is left to the discretion of the app developer. If a developer has chosen not to include voice interaction, then some features of the app may not be steered by the voice assistant, instead requiring manual interaction.

2.2 Designing with Voice

While there are pure VUIs, only allowing voice input and output, many interfaces provides multiple modes for both input and output. Screens, keyboards, and other types of input are combined with voice to produce multimodal interfaces. There are several approaches to using voice in interaction which can be categorized as screen-first, voice-only, and voice-first [10, 47].

The screen-first approach prioritizes the screen first and utilizes voices to enhance screen functionality [47]. The screen-first approach is currently applied to most smartphones as the voice assistants are highly dependant on the screen. In many cases, the user is unable to complete a voice-initiated interaction without manual input through the screen [47]. For example, if a user requested nearby restaurant recommendations, a screen-first system may read aloud the first recommendation and output the remaining alternatives on the screen for the user to manually select an option. Screen results from asking Google Assistant for nearby restaurants can be seen in Figure 2.3.

A voice-only interaction uses only voice for both input and output, unlike screen-first and voice-first. Early screenless smart speaker models such as the Amazon Echo,



what are some nearby restaurants

I found a few places.

Daawat restaurant
4,1 ★★★★ (288) · $$
Indian · Första Långgatan 30

Sjöbaren Fish Seafood

Figure 2.3: Google Assistant displaying results for nearby restaurants on a Android phone

Google Home, and Apple HomePod are examples of voice-only design [47]. Due to the singular mode of input and output, using voice-only interactions to complete simple tasks can become tedious [47].

A voice-first approach is the inverse of the screen-first approach. In a voice-first design, a complementary display is used to visually supplement the voice interaction and a user can complete an interaction through voice alone [47]. Voice-first has been widely embraced in the latest models of voice assistants like the Amazon Echo Show and the Google Home Hub which include touchscreens. The voice-first approach is different in that many traditional graphical user interface elements, such as heavily-nested menus and visually dense content, are completely eliminated in favor of contextualizing information to enhance whatever the voice is communicating [47]. Moreover, a voice-first approach assumes that the user may not always have access to look at or touch the screen; therefore, voice carries the bulk of the interaction in a voice-first approach [47].

2.3 Autonomous Cars

As the field of voice interaction continues to develop, so does the field of autonomous driving. According to the SAE International, there are six levels which describe the level of autonomous driving a car is capable of, as shown in Table 2.1. It is worthy to note that Levels 2 and under still require a human driver to perform part or all of the driving task, even with the *autonomous driving system* (ADS) engaged [41]. In contrast, vehicles classified as Level 3 and up are able to fully takeover the primary task of driving, under varying scenarios [41].

Table 2.1: SAE levels of driving automation [41]

Level	**Autonomous Driving System Role**
Human driver monitors driving environment	
Level 0 No Driving Automation	Does not perform any of the driving task on a sustained basis
Level 1 Driver Assistance	Performs part of the driving either in the longitudinal OR lateral motion and can be disengaged immediately upon driver request
Level 2 Partial Driving Automation	Performs part of the driving in both the longitudinal AND lateral motion and can be disengaged immediately upon driver request
Autonomous Driving System monitors driving environment while engaged	
Level 3 Conditional Driving Automation	Performs all of the driving under select driver-manageable conditions and can be disengaged immediately by the driver or issue a request for the driver to intervene
Level 4 High Driving Automation	Performs all of the driving under most diver-manageable conditions and may delay driver-requested disengagement
Level 5 Full Driving Automation	Performs all of the driving under all driver-manageable conditions and may delay driver-requested disengagement

For Level 2 and under autonomous driving, ADS can provide drivers assistance with the primary task of driving. This can in turn free up some of the driver's attention and cognitive load to complete secondary tasks, such as tuning the radio, replying to a text message, or getting directions to a nearby point of interest. However, in Level 3 and up autonomous driving, handing off the primary task of driving to the ADS from the human driver may introduce a new interaction paradigm in the vehicle. In scenarios where the human driver is no longer responsible for the driving, the primary task may shift dramatically from driving to other tasks, such as entertainment or work.

2.4 Distracted Driving

The National Highway Traffic Safety Administration (NHTSA) is the U.S. governmental agency responsible for setting and enforcing safety standards in vehicles [30]. In 2016, the NHTSA reported that 3,450 deaths in the United States were reportedly due to distracted driving [29]. The year prior, a staggering 391,000 people suffered injuries from distracted driving related incidents [29]. With statistics like these, distracted driving has become a key traffic safety issue.

According to the NHTSA, distracted driving refers to the inattention of drivers from the primary task of driving to other activities or secondary tasks [30]. Electronic devices in particular are an area of concern for the NHTSA as more and more technology is incorporated into modern vehicles. Electronic devices can influence drivers by causing visual distraction, manual distraction, and cognitive distraction [30].

In an effort to combat distracted driving from electronic devices, the NHTSA has thus far issued two phases of guidelines for designing in-vehicle electronic devices. Phase One of the design guidelines concerns the design of *original equipment* (OE), such as the in-vehicle infotainment system that already comes installed on a vehicle [30]. Phase Two extends the guidelines from the first phase to include portable and aftermarket devices, which includes smartphones with a car mode [31]. Both guidelines use eye glance metrics as acceptance criteria where eye glances away from the road for more than 2.0 seconds are correlated with an increased crash risk [30, 31]. While both Phase One and Phase Two acknowledge voice interaction as an alternative to traditional HMIs, both guidelines explicitly do not include voice interaction. The NHTSA has announced plans for Phase Three of the guidelines, which would provide recommendations specifically for voice interaction; however, there is currently no set date for when these guidelines will be published, leaving the definition of safe voice interaction in vehicles largely undefined.

While the jurisdiction of the NHTSA is limited only the United States of America, its safety recommendations extend beyond those borders. In following the guidelines set by the NHTSA for vehicles in the American market, car manufacturers in practice also apply these guidelines to vehicles in markets outside of the United States. Alternate guidelines for designing in-vehicle interfaces include those published by Japan Automobile Manufacturers Association (JAMA), Alliance of Automobile Manufacturers (AAM), and the EU [20, 26, 9]. However, the NHTSA guidelines are the most recent guidelines and likely the most relevant when considering voice interaction as an emerging technology.

2.5 Guidelines for In-vehicle and Voice Interfaces

At present, few guidelines consider the holistic interface of a voice-assistant integrated IVI. However, the existing guidelines for both in-vehicle and voice interfaces outline important considerations for each respective interface that should be taken into account.

2.5.1 NHTSA Interface Guidelines

Phase One of the NHTSA interface guidelines are applicable to original IVIs [30]. Recommendations in the Phase One guidelines include where to place the IVI, what tasks should not be allowed on the IVI, and IVI response time. The guidelines also describe a number of best practices for interacting with an IVI manually. Some notable interaction guidelines include single-handed operation, interruptibility, and disablement [30]. Drivers should be able to operate the IVI with a single hand and while driving and the IVI should not require the driver to complete an uninterrupted sequence of tasks [30]. Drivers should be able to stop a task mid-way and then resume the task if not completed [30]. Additionally, IVIs should have the ability to disable the display of any non-safety related information through methods including dimming, blanking, or changing the state of the display [30].

Phase Two of the NHTSA guidelines expand upon those covered in Phase One to include the interfaces of portable and aftermarket devices [31]. Notable additions from the Phase Two guidelines include pairing devices, driver mode, and access to emergency services and alerts [31]. For devices that can be paired with the original IVI, the pairing and disconnection should be easy to complete. When paired and using the IVI display, guidelines from Phase One should also be followed [31]. For unpaired devices, there must be a driver mode which conforms to the Phase One recommendations [31].

As the second set of guidelines are an expansion, portable and aftermarket devices described in Phase Two of the guidelines must also follow the guidelines defined in Phase One. Notable additions from the Phase Two guidelines include pairing devices, driver mode, and access to emergency services and alerts [31]. In both scenarios, emergency services and alerts must be easily accessible [31]. However, the guidelines do not state what additional notifications should also be accessible, such as communication notifications.

2.5.2 Android Auto

Android Auto is the integration interface made available by Google for compatible Android phones. Only apps which fall into the navigation, communication, media, or automaker categories can be enabled for use through Android Auto [13]. The Android Auto design guidelines are primarily concerned with the appearance and structure of visual content on the IVI. Android Auto uses a global UI, which means the visual interfaces of each app uses a template provided by Google [13]. By using a template approach, drivers using Android Auto do not need to learn app-specific UIs when switching between two apps in the same category. The Android Auto guidelines make almost no mention of designing for voice interaction, save for constructing or replying to a message [13].

The Android Auto guidelines prescribe recommendations for user input, menu organization, and notification display. The pace of input into the IVI should be determined by the user [13]. This recommendation aligns with the NHTSA guidelines for interruptibility. The Android Auto guidelines also suggest items in the drawer menu be context specific [13]. For example, rather than displaying broad categories such as "All Songs" and "All Artists" the menu items should be more specific such as "Top Hits" or "Favorite Artists". The guidelines also state that notifications may be used if they are appropriate to driving or important enough to interrupt the driver [13]. However, Android Auto provides little guidance on what is considered "important enough" and leaves it up to the discretion of the designer.

2.5.3 Apple CarPlay

Like Android Auto, the Apple CarPlay guidelines use a global set of UI elements and a template system [2]. Voice integration is briefly described for automaker and communication apps, though Apple does have a separate guideline for custom Siri voice commands [2]. When CarPlay is active, interactions on the iPhone should be eliminated and CarPlay interactions should never require input from the iPhone [2]. The Apple guidelines also provide a number of test conditions for designing a CarPlay enabled app [2]. For example, apps should be tested in an actual car, not a simulator alone, and in varying network conditions [2].

Generally, the Apple CarPlay guidelines provide more guidance to designers regarding the architecture of apps including badging, error handling, and navigation structure. The Apple CarPlay guidelines also provide detailed recommendations for content writing, organization, and notifications. Written content in CarPlay should be succinct and avoid accusatory or judgmental tones [2]. Content and navigation should require as few inputs as possible, either through flat or hierarchical navigation [2]. Moreover, there should only be one path for manual input to a specific view [2]. Alerts should be minimized and used only when there is error so users will take them seriously [2].

2.5.4 Google Assistant

Google's design framework for voice interaction is called Conversation Design. It is an extensive framework with a lot of detailed information and examples. Google highlights the framework as being multimodal and consisting of many different disciplines of design such as voice, audio and visual design. Google argues that all of these disciplines are required to design real conversations as, according to them, real conversation is a multimodal activity.

The Conversation Design framework is built upon Grice's *Cooperative Principle*. This principle states that conversation is shaped by the social context and that this shaping of the conversation relies on a type of subconscious cooperation between the conversing parts, Grice's Cooperative Principle is covered in depth in section 3.7 in this report.

The design framework provides extensive guidelines regarding the aspects of context of conversation, variations of phrases and turn-taking during dialogues. A shorter list of visual components to be used together with voice assistants is also provided. Information regarding how and when graphical components are to be used in combination with conversation is however very limited and the few guidelines related to this that exists, are very general.

2.5.5 Siri

The Siri voice guidelines describe how to integrate the voice assistant in a variety of contexts for a seamless voice-driven experience [3]. Moreover, the guidelines describe when Siri would enhance an interaction and how to create Siri responses. The Siri framework supports shortcuts which can perform useful or frequent actions without much navigation [3]. Shortcuts should be short and concise, but also not context-specific [3]. An example shortcut could be "Order clam chowder". Designers can make shortcuts more relevant and accurate using custom vocabulary or providing examples on the screen [3]. Like Google Assistant, Apple recommends that Siri responses are conversational. Apple additionally recommends that actions should be voice-driven with as little manual input as possible, a voice-first approach. Verbal responses from Siri should be accurate and relevant to the user's request [3].

2.6 Related Research

Voice interaction in vehicles has been well-researched in terms of distracted driving and usability. However, as the voice interfaces continue to evolve, so do research opportunities in the field. Previous research of automotive VUIs has generally focused in-vehicle VUIs. In other words, VUIs that are built into the car by the OEM, instead of portable alternatives such as modern day voice assistants.

In their 2013 review, Lo and Green surveyed key researched in-vehicle VUIs [25]. The VUIs covered by Lo and Green all used NLP, but they did not utilize cloud-computing as voice assistants do [25]. Core functionality between the systems surveyed included communication, media, and navigation, not unlike the enabled app categories on both Android Auto and Apple CarPlay [25]. However, some of these systems had extended functionality, such as climate control via voice command [25].

More recent studies compared different VUIs against each other to identify the effect of different voice-driven multimodal interactions on distracted driving. Mehler et al. compared the Chevrolet MyLink and Volvo Sensus against each other, where the former allows for 'one-shot' voice input while the latter requires input through a series of menus and sub-menus [28]. For most tasks, 'one-shot' input performed better than guided, menu-based input given no recognition errors [28]. However, if there were recognition errors, the 'one-shot' input, similar to that of current voice assistants, increased driver workload and caused user frustrations [28]. Reimer et al. further expanded upon the work by comparing a Samsung S-Voice assistant against the two in-vehicles systems evaluated by Mehler et al. [28, 39]. Reimer et al. found that the smartphone assistant actually performed worse that the embedded in-vehicle systems [39]. However, they proposed that perhaps coupling the smartphone into the embedded IVI to create one holistic experience may reduce workload and visual demand [39].

One study that does examine the holistic experience of a voice-assistant integrated IVI on distracted driving was conducted by Strayer et al. for the AAA Foundation for Traffic Safety [42]. Motivated by the lack of Phase Three guidelines from the NHTSA, this study investigated how Apple's Siri affects distracted driving [42]. The study found that the use of a voice assistant to carry out a secondary task significantly increased the crash risk; however, the study has yet to be corroborated and does not provide suggestions to address the issue of increased risk [42].

Beyond voice interaction, distracted driving has been studied in many capacities. A 2009 review by Bach et al. surveyed 100 papers related to attention understanding within automobiles [4]. Despite the extensive studying of attention and cognitive load while performing secondary tasks, the review makes it clear that there is no one singular method for assessing attention and cognitive load [4]. Previous studies have used primary task performance, secondary task performance, eye glance behavior, physiological measures, and subject assessments to measure attention and cognitive load [4]. The variety of methods and the lack of a singular standard illustrate the difficulty in capturing and measuring what goes on in the mind while performing multiple tasks.

3

Theory

Voice interaction, especially for in-vehicle use, sits at the crux of many fields including design research, attention, cognitive load, and linguistics. This chapter covers the theory and domain-specific knowledge from these fields that are related to this project.

3.1 Research Approach

The research approach of this project relies on *human-centered design* (HCD), where user involvement and testing with users is central to the design and development of a product. The idea that the solutions to a problem is held within the very people who face this problem is a core idea of HCD [19]. Social research principles also support the frequent user involvement in this research project. One prevalent idea in the social research approach is that if enough people agree on a subjective opinion, it can become an objective fact [46]. This can be said of the design field, where many designers and researchers consider involving users as part of the design process or design research to be standard, thus objectively validating HCD as a approach. HCD largely focuses on understanding the users and evaluating with and for users throughout the process [12]. Another characteristic of a HCD approach is applying a wide range of disciplinary skills and perspectives [12]. This project especially applies theory from psychology and cognition to be able to properly research the user's attention and cognitive load. Applying a varied set of theories and concepts from different fields is, according to Gaver, a way to both inspire and articulate new and already existing designs [11].

3.2 Wickens' Attention Model

The attention of a human being is a limited resource. When it comes to the task of driving and all of the secondary tasks that follow in a modern car, managing attention and distributing it correctly becomes very important. There are various theories explaining the complexities of human attention resources. One which has proven to be especially relevant to mental workload in relation to multitasking is Wickens' *Multiple Resource Theory* [48]. According to this theory, the attention of humans can be divided into different resource pools. The different resource pools represent the humans ability to process different types of stimuli. The internal processes are divided into perception, cognition and response. Figure 3.1 shows a four-dimensional model of the resource model.

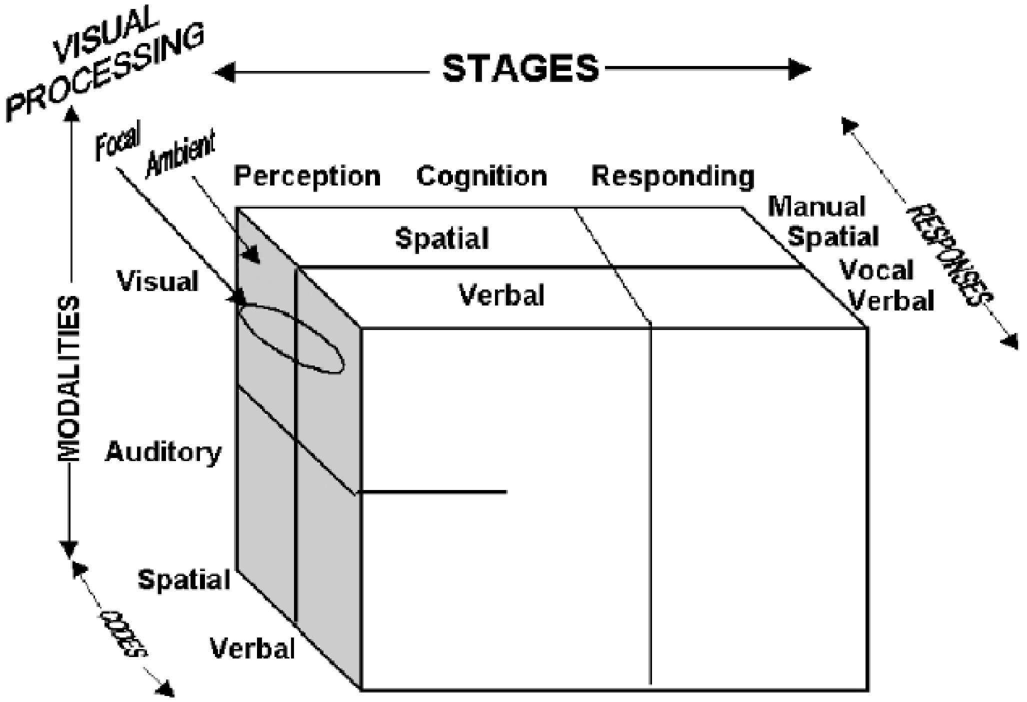


Figure 3.1: Wickens' Multiple Resource Model [48]

According to Wickens, humans are able to perceive four different types of input: spatial-auditory, verbal-auditory, visual-spatial and visual-verbal [48]. Multiple Resource Theory posits that multiple simultaneous inputs are better perceived if they are of different types. When internal mental processes move from the perception of input to the cognition of it, humans are capable of simultaneously processing verbal and spatial input. In the final internal process, humans are capable of deciding a response to manual-spatial and vocal-verbal input at the same time. However, the ability to simultaneously process input is still affected by the weight and complexity of the individual inputs. This means that very complex spatial input will affect a person's ability to process other input at the same time, even if the additional input is of another modality.

Multiple Resource Theory helps to reinforce the findings of previous research which concludes voice interfaces as being a safer input method in vehicles [28, 39, 25]. According to the theory, verbal information from a voice interface would never interfere with the visual information from looking at the road as both inputs are processed in the driver's mind.

3.3 Intensive and Selective Attention

Another theory for explaining human attention is Kahneman's work on effort and attention [21]. According to Kahneman, the two most important factors affecting attention are intensity and selectivity [21].

Intensity is directly connected with the effort one applies to their current focus of attention [21]. A person may direct greater effort into a specific focus of attention when motivated by arousal or personal choice [21].

Selectivity describes how a person decides to distribute their effort toward different sources of attention [21]. Ultimately, the total amount of effort available at a given moment is limited [21].

Problems occur when different sources of attention and their demand of effort interfere with each other. This explains the difficulty behind dividing attention, such as in multitasking. The idea of interference in distribution of attention is interesting, as it provides a contextual explanation of the ideas presented by Wickens' Attention Model which were summarized in section 3.2. explaining the difficult task of dividing attention.

3.4 Cognitive Load

There are many different definitions of cognitive load, sometimes also referred to as cognitive workload. Waard decomposes cognitive workload into two parts: demand and load [45]. Demand is the specific external task demand a task places upon a user. Load is the individual effect of the task demand placed upon a user. Task demand is highly dependant on the complexity of the task. Increased task complexity increases the demand of the task. Perceived load is more complex and depends on a variety of factors including skill, experience, and current mood of the person performing the task. When examining cognitive load, both task demand and task load should be considered, as the two are closely related. In a driving situation, the main task of driving places a certain demand on the driver. Depending on the driver's skill level and experience, the perceived load will vary. When adding secondary tasks, like making phone calls and playing music, the total load of the driver further increases.

When analyzing the cognitive load, there are several aspects to consider. Cognitive load essentially is a measure of how many mental processing *resources* are available. The upper limit of resources is referred to as the *capacity* [45]. In a practical scenario where cognitive load is measured, a researcher tries to measure how many resources are available and how close the test participant is to their capacity limit. In a driving scenario, the driver always needs to have enough resources to handle the primary task of driving.

In section 3.2 the concept of attention resources was introduced. Attention resources are closely related to mental processing resources. Perceiving input, the first step of the previously mentioned Wickens' Attention Model [48], is a prerequisite to processing input through the consumption of mental resources. The stages of cognition and response in Wickens' model correspond with the mental processing concepts that are central to discussing cognitive load.

3.5 Eye Movement

In order to assess visual distraction, it is important to understand how to analyze a person's eye movements, through four basic movements. These four eye movements are saccades, smooth pursuit movements, vergence movements, and vestibulo-occular movements [37].

Saccades is the most basic type of movement. Saccades are quick movements that occur when a person changes their eye's fixation point from one to another. [37]. Saccades may be short or long depending on the situation. When driving, the moment between saccides can be interesting to analyze as the user's fixation points are likely to switch between on-road and on the various interfaces within the vehicle.

Smooth pursuit movement occurs when a person fixates their view on a moving object. Smooth pursuit movement is difficult to perform without a moving object. Attempts to perform this eye movement by the untrained may actually instead be a series of short saccades [37].

Vergence movements occurs when a person fixates on a point that moves either closer or further away from the person [37]. Vergence movements are different from the two mentioned above, since the eyes during this movement moves in different directions from each other compared to moving in the same direction during saccades and smooth pursuit movement [37].

Vestibulo-ocular movements are made in order to stabilize the eyes during movements from the outside world such as fixating on a point while the head is moving in some direction [37].

When working eye tracking, several types of data can be analyzed. One type of data is glances. A glance is a fixation on a specific point in the world between two saccades. By this definition, glances have both a duration and a direction. With respect to this project, glances are a highly relevant type of data as they are used in part by the NHTSA to define safe task interactions [30]. Glance directions can be divided into glance areas of interest in order to more easily measure glances on specific areas of interest within the car.

3.6 Elements of Voice Interfaces

To understand and discuss VUIs, it is important to know the basic elements of a voice interface. These elements are: utterances, responses, prompts, and intents. Together, these elements create a dialog, a linguistic exchange between the user and the VUI [16].

An utterance is a natural unit of speech which can range from a single word to a small cluster of sentences [16]. With respect to VAs, utterances are usually inputs from the user.

A response is the second utterance in a summons/response pair [16]. If a summons is a request from a VA user, such as "What's the weather today?" then a response manifests as information related to the day's weather.

A prompt is a system utterance that helps guide user input [16]. Prompts are most often in the form of questions which can be explicit ("Which flowers would you like to order, roses or daisies?"), implicit ("Which type of music would you like to listen to?"), or open-ended ("What can I do for you?") [16]. Inferential prompts are typically statements that convey to the user the capabilities of the VUI ("I can answer questions about train arrivals, departures, and on-board amenities.") [16].

An intent is a representation of action or a feature that fulfills a user's spoken request. Intents may include variable information to complete a user's request. In the previous example for responses, the intent is to get weather information where "today" was a variable that enables the VUI to respond with relevant information.

Utilizing these elements, and mimicking a VUI's way of processing these, will be necessary when trying and testing Wizard of Oz style prototypes.

3.7 The Cooperative Principle

NLP VUIs aim to function through conversation between the user and the system. To design computers to converse in a natural way, VUI designers must understand the underlying principles of conversation. The semantics of conversation have been carefully studied by H. Paul Grice who has defined the underlying mechanics of conversation through a set of principles [15]. Together, these principles are know as The Cooperative Principle, which is made up of four sets of subprinciples or maxims [15]. The maxims describe the subconscious cooperation the occurs as a person formulates sentences in a conversation [15]. Grice's Maxims are as follow [15]:

Quality

1. Make your contribution as informative as required (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

Quantity

1. Try to make your contribution one that is true.

    (a) Do not say what you believe to be false.

    (b) Do not say that for which you lack adequate evidence.

Relation

1. Be relevant.

Manner

1. Be perspicuous.

    (a) Avoid obscurity of expression.

    (b) Avoid ambiguity.

    (c) Be brief (avoid unnecessary prolixity).

    (d) Be orderly.

These maxims can be used to formulate the output of a VUI. They can also be applied when designing the VUI to anticipate different user inputs and how the system should respond to them. This applies for designing the dialog of any prototypes developed as a part of this project.

4

Methods

This chapter covers all methodology relevant to the project. Usage details regarding the methods, suitable contexts of use and alternative methods are discussed. The methods are varied ranging from purely evaluative to creatively stimulating and can be utilized at different points throughout the project.

4.1 Wicked Problems and Iterative Design

Many of the challenges and problems designers aim to solve are known as wicked problems. Rittel and Webber were the first to define wicked problems, which are problems that are unique, have no definitive formulation, have no stopping rule and whose solutions are not true-or-false but good-or-bad [40]. By comparison, there are tame problems which have a definite formulation and solution, such as math problems which have stopping rules to indicate when a solution has been reached and equations by which the solution can be verified as true or false. Solutions to wicked problems are rated on a scale of good or bad, where some solutions are better than others and some maybe be considered a good enough solution to the problem. Thus, as many designers tackle wicked problems, they may use an iterative design process to explore several solutions to find a better or good enough solution.

There are four basic activities in a design process: establishing requirements, designing alternatives, prototyping, and evaluating [36]. Iterative design is the process by which a design is refined by user feedback through the repetition of these four design activities. The iterative design process has been visualized as a design funnel, where at the start of the process, designers begin at the wide end of the funnel and explore a broad number of potential design solutions [7]. As designers progress through the design process, they move towards the narrow end of the design funnel, reducing the number of possible design solutions and ultimately arriving upon a design solution [7].

In an iterative design process, each iteration is a step toward narrowing the design funnel. However, each iteration in itself is not narrowing, or reducing [7]. In fact,

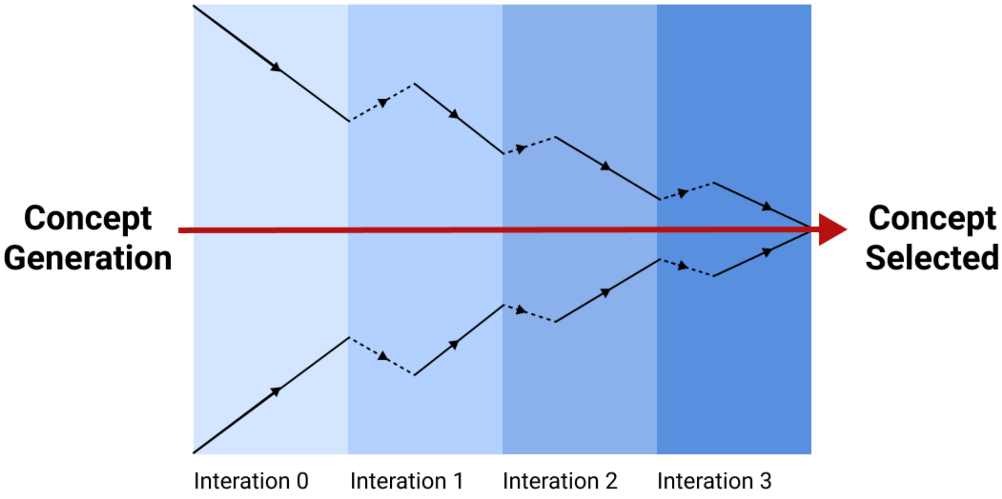


Figure 4.1: Design funnel as described by Bill Buxton where dashed lines indicate divergence and solid lines indicate convergence in the design process [7]

each iteration is a combination of divergent and convergent thinking where the divergence comes from the generation of new ideas and improvements to a design and convergence is the reduction of those solutions into an iteration or prototype of the design [7]. With respect to wicked problems, each iteration adds knowledge and is an attempt to define and solve the problem.

4.2 Literature Reviews

Literature review is conducted by researching and reviewing research literature relevant to the field of study [27]. The purpose of a literature review is to gather knowledge from previous research or findings to guide new research within a related field [27]. A literature review can vary in its result, from establishing a theoretical framework for discussing previous and future research to practical information, such as guidelines for designing for a specific context. Literature reviews enable researchers and designers to make connections and cross-references between several literature sources in order to understand the larger context behind their own work as well as how their own research can provide new knowledge.

4.3 Summative and Formative Evaluation

Evaluative testing can be divided into two types: evaluative and summative [33]. A summative evaluation is focused on evaluating the quality of a system or a product [33]. It is typically suitable in the end of a design process, evaluating a finished system, but also when two alternatives are available or when market competitors are analyzed. Summative evaluations tend to be focused on measuring quantitative data [33]. A formative evaluation is focused providing input to improve a system of a product [33]. It is typically done in an iterative design process, driving the design forward and motivating design choices and improvements [33]. Formative evaluations are more focused on providing qualitative input [33].

4.4 Field and Lab Testing

There are several different possible approaches to testing the voice assistants in cars. For this project, the considered options are: in a car simulator, in a real car on a test track or in a real car on real roads. There are specific pros and cons of each method but the contextual aspects of sitting in a real car are weighted as being especially important. Simulations have the great benefit of being a completely controlled environment where the scenario can be completely consistent between tests. A large disadvantage of using simulation is that the participants never feels the sense of real danger as a consequence of their driving, this might lead to the driver adapting a more reckless driving style than their usual, affecting the overall outcome of the test [8].

Doing testing in a real car while driving on actual roads with traffic has the benefit of providing real, contextual information and performance shaping factors but at the same time, the environment is completely uncontrollable. Traffic situations, weather, red and green lights are all factors that would be completely random. Knowing exactly how these factors affect the results is very difficult.

Conducting tests in a real car on a closed off controlled test circuit allows for some of the benefits of both previously mentioned methods. The environment can be better controlled. Real traffic situations can be mimicked and since the participants are driving real cars, the sense of consequence and danger is there, forcing the driver to always pay close attention to their driving. Weather still is an uncontrollable factor.

4.5 A/B Testing

A/B testing means testing of two different version of a design so that results can be compared and it can be determined which one who performs better [27]. The A/B testing method is not qualitative, the two versions A and B are only measured by how much they fill a certain quantitative criteria. An example could be two versions of a voice assistant where the time to complete a specific task is measured. The A/B test would in the case of the example only result in knowledge about which one is faster, not why it is faster [27]. In order to cope with this lack of qualitative data, it is recommended that it is combined with other, qualitative methods.

4.6 Interviews

Conducting interviews is a method for design research that allows direct interaction with users and allows researchers to take part and explore the user's personal views, experiences and perceptions about a subject [27]. Interviews are best done in person so that the researcher may collect information in the form of body language and facial expressions as well as what is actually said by the user [27].

Interviews can be structured or unstructured. Structured means that all questions are planned in advance and unstructured has the questions made up as the interview is active [27]. There are combinations where topics and some base questions are formed in advance but the the interviewer is allowed to ask new unplanned questions if he or she wishes, this is sometimes called semi-structured interview.

Interviews is a very flexible method and allows customization and tweaking for specific uses. Interviews can be done in groups or individually and it can be focused on attaining information from specific roles or user groups [27].

4.7 Cognitive Workload Measuring

This section covers details and differences of four different methods that have been developed for the purpose of measuring a subject's cognitive workload.

4.7.1 NASA-Task Load Index

The NASA-Task Load Index (NASA-TLX) is a rating based measurement method for assessing the subjective experience of workload during activities [17]. The method divides the workload into several specific workload sources which allows specific sources of workload to a specific task to be identified [17].

The method has two steps, first a set of rating scales, then pairwise comparisons. The first step consists of rating all possible sources of workload on a 20-point scale representing 0 to 100 in steps of 5. The different sources of mental can be seen in table 4.1.

Table 4.1: The NASA-TLX measurement factors and their descriptions [17]

Title	Endpoints	Description
Mental Demand	Low/High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low/High	How much physical activity was required (e.g.. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Low/High	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, con- tent, relaxed and complacent did you feel during the task?

The second part of the NASA-TLX is a weighting process to be able to weight the ratings in accordance to how much the influenced the task. Possible pairwise combinations of the sources of workload are compared and the user gets to choose which one out of the two influenced the task more than the other. This leads to a weighted rating for each one of the sources of workload and the total weighted task load score is calculated through the average value of the weighted scores.

4.7.2 The Driving Activity Load Index

The *Driving Activity Load Index* (DALI) is a subjective evaluation method for evaluating the cognitive workload of car drivers [35]. The method is largely based on the NASA-TLX but is revised to more carefully evaluate aspects that are specifically relevant to driving, ruling out aspects like e.g. physical demands [35]. A complete list of all the measurements factors of the DALI method and corresponding descriptions can be seen in table 4.2.

Table 4.2: The DALI measurement factors and their descriptions [35]

Title	**Endpoints**	**Description**
Effort of Attention	Low/High	To evaluate the attention required by the activity – to think about, to decide, to choose, to look for and so on.
Visual Demand	Low/High	To evaluate the visual demand necessary for the activity.
Auditory Demand	Low/High	To evaluate the auditory demand necessary for the activity.
Temporal Demand	Low/High	To evaluate the specific constraint owing to timing demand when running the activity.
Interference	Low/High	To evaluate the possible disturbance when running the driving activity simultaneously with any other supplementary task such as phoning, using systems or radio and so on.
Situational Stress	Low/High	To evaluate the level of constraints/stress while conducting the activity such as fatigue, insecure feeling, irritation, discouragement and so on.

The method is used after a user has performed a task or a set of tasks related to driving. The user ranks each of the measurement factors on how big of an impact they had on the task on a two point scale, ranging from very low to very high. The measurement factors are then weighted in relation to each other, the user is shown two factors at a time and chooses the one of the two which had the most impact. This is repeated with factors until all possible combinations has been shown. The total number of times a certain factor has been chosen as the most impactful is its weight number. The original rank score, that the user filled out, is multiplied with its corresponding weight to produce that aspect's adjusted score. The sum of all weighted scores divided by 15 represent the weighted rating for the whole task

4.7.3 Subjective Workload Assessment Technique

The Subjective Workload Assessment Technique (SWAT) is a scaling procedure that allows test participant to put number on their subjective experience of mental workload during a task [38]. It was originally developed for the U.S. Air Force to be used to assess their pilots mental workload [38]. The SWAT method measures the workload in three different dimensions, these are *Time Load, Mental Effort Load* and *Psychological Stress Load* [38]. These three dimensions are combined to give a measure of the total workload of a task.

The SWAT method divides the three previously named dimensions into three different levels. Where one would indicate a low level while three indicates the highest, e.g. a time load rating of one would indicate low levels of time load, where the user has a lot of time to perform the task while a rating of three would indicate very high level of time load where the user has no spare time and has to deal with overlapping activities.

The first step of the SWAT methods is a card sorting process. Cards representing all different combinations of levels for each of the three dimensions are to be sorted and ordered from the combination that represent the lowest workload to the highest. The lowest would logically be a rating of 1, 1 and 1 for time load, mental effort load and psychological stress load respectively while the highest would be 3, 3 and 3. The steps in between would typically vary with users and tasks. The user would then perform a task and rate it on the three dimensions of workload. By seeing where this rating places in the order of the sorted cards, a weighted workload score ranging from 0 to 100 can be calculated.

4.7.4 Rating Scale Mental Effort

The Rating Scale Mental Effort (RSME) methods is a simple, one dimensional subjective scale method for measuring mental effort required for a task [49]. It is more simple than a lot of other mental workload measuring methods due to the fact that it only requires the user to answer one single scale question.

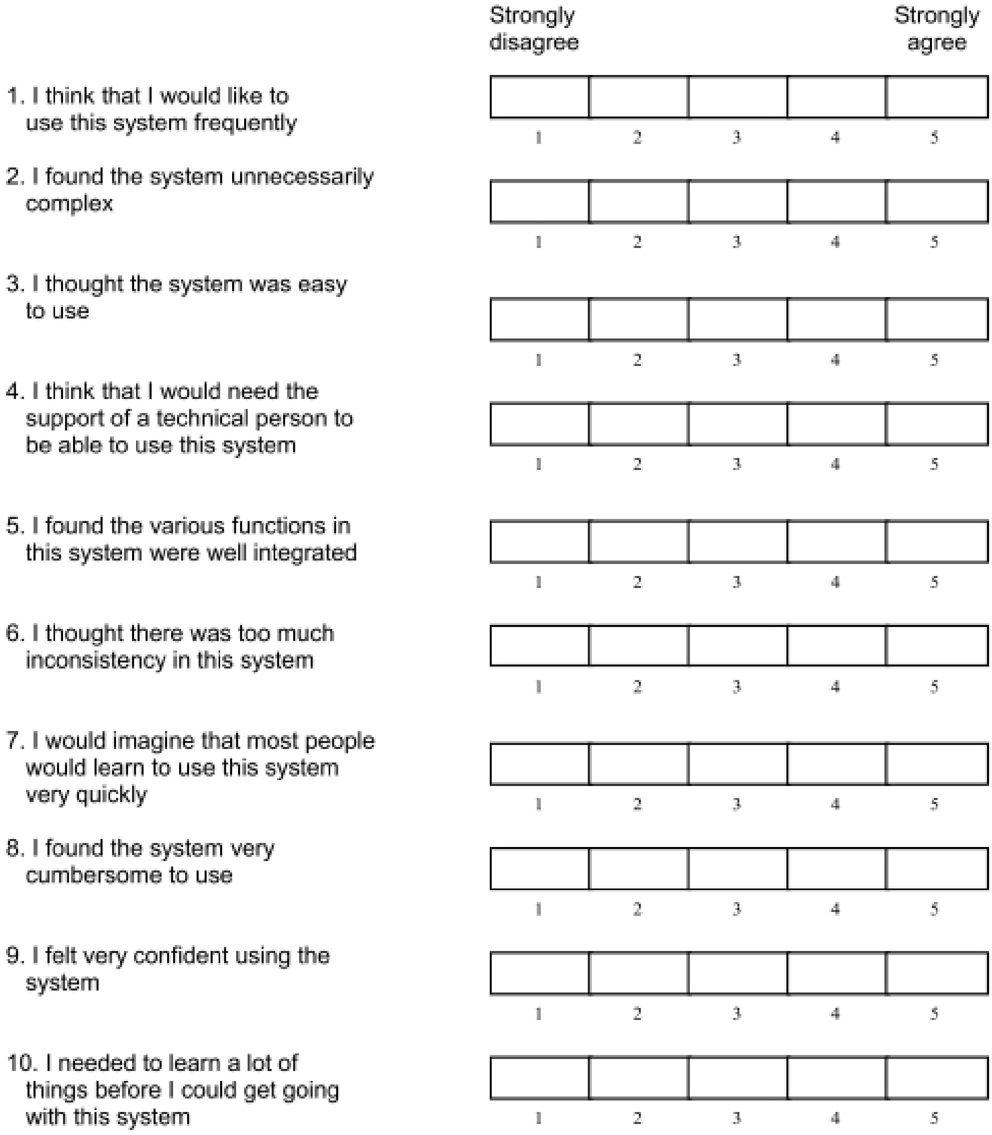
The scale is made up by a 15 cm long line with every 1 cm indicated. The line is accompanied by verbal descriptors of the level of mental effort, examples are "almost no effort" and "extreme effort". The position of the verbal descriptors along the scale has been carefully adjusted after many user tests during the initial development of the RSME method [49].

In comparison to other mental workload measurement methods the RSME lacks some of the more complex aspects that make up the total workload, it does not consider different dimension of mental workload like NASA-TLX, DALI and SWAT does [45].

4.8 System Usability Scale

The System Usability Scale (SUS) is a simple usability scale for subjective assessment of a system's usability [6]. The SUS is made to be quick an allow users to very quickly convey their experienced usability of a system they have just used.

The SUS is a likert scale and it utilizes ten 5-point scales ranging from strongly disagree to strongly agree. The SUS contains scales covering topics like the complexity of the system, integration of functions and whether it was cumbersome to use etc. A full example of the SUS including all scales can be seen in Figure 4.2.



Strongly disagree — Strongly agree

1. I think that I would like to use this system frequently □ □ □ □ □ 1 2 3 4 5
2. I found the system unnecessarily complex □ □ □ □ □ 1 2 3 4 5
3. I thought the system was easy to use □ □ □ □ □ 1 2 3 4 5
4. I think that I would need the support of a technical person to be able to use this system □ □ □ □ □ 1 2 3 4 5
5. I found the various functions in this system were well integrated □ □ □ □ □ 1 2 3 4 5
6. I thought there was too much inconsistency in this system □ □ □ □ □ 1 2 3 4 5
7. I would imagine that most people would learn to use this system very quickly □ □ □ □ □ 1 2 3 4 5
8. I found the system very cumbersome to use □ □ □ □ □ 1 2 3 4 5
9. I felt very confident using the system □ □ □ □ □ 1 2 3 4 5
10. I needed to learn a lot of things before I could get going with this system □ □ □ □ □ 1 2 3 4 5

Figure 4.2: A full example of a SUS [6]

The user's inputted values on the scale goes through a calculation process where the scores are converted to lower values if they indicate bad usability or higher values if they indicate good usability. This is done by simply subtracting 1 from all even question and subtracting the score of all even questions from 5. After summarizing and multiplying with 2.5, a final SUS score between 0 and 100 emerges.

4.9 Subjective Assessment of Speech System Interfaces

The Subjective Assessment of Speech System Interfaces (SASSI) method, is a Likert scale based questionnaire for subjective evaluation of speech system interfaces [18]. The SASSI consists of 34 different scales related to the user's experience with the speech interface [24]. Each scale is a seven point Likert scale. The scales are divided into six different topics: System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability and Speed [24]. A full SASSI questionnaire can be seen in Figure 4.3.

4.10 Eye Tracking

Eye tracking is the process of measuring the eye movements in relation to different points of fixations in the world. Eye tracking can be used as a measure of visual attention. There are two types of eye tracking: automated and manual.

Automated eye tracking refers to all eye tracking technology that automatically record and translate eye movements into data. Models specifically suitable for in-vehicle eye tracking are remote eye trackers, which do not require the user's head to be locked in position. Makers of popular remote eye trackers include Tobii, EyeTribe, and SMI [32]. While automated eye trackers may benefit from the advantages of technology, such as increased precision, they have their limitations, such as a smaller area of focus. Automated eye trackers may also experience issues with inconsistencies, accuracy, and precision of the collected data, which may require manual review of the collected data [32]. Moreover, these eye trackers may demand consistent lighting conditions and additional configuration, sometimes for each user [32].

Manual eye tracking refers to tracking and measuring user visual attention through the visual analysis of video recordings. This is done by having a researcher manually code or annotate segments of a video recording according to relevant glance areas, usually with the aid of some annotating software. With respect to driver attention, relevant glance areas may include parts of the road or areas of the vehicle's interior. The precision of this method is lower compared to automated eye tracking, but allows for examining larger areas of glance interest. Moreover, manual eye tracking does not require advanced camera equipment or configuration. Cameras used in manual eye tracking should have sufficient video quality to see the user's eyes under all expected light conditions. Depending on the desired time-precision of the eye tracking analysis, different frequencies of video capture may be considered.

Automated and manual eye tracking each have their advantages and disadvantages. Automated eye tracking is most suitable for situations where the glance area is relatively small and requires high precision. For example, when examining areas of interest in the driver information module (DIM), the area behind the steering wheel. For larger areas of glance interest, such as multiple areas within a vehicle, manual eye tracking may be more suitable. Manual eye tracking also requires less setup, but requires additional labor to manually code eye glances in the video footage.

4.11 Affinity Diagramming

Affinity diagramming is a method used for analyzing and structuring results from research [27]. The results are structured so that themes emerge allowing designers to better understand and categorize data, this ultimately leads to a good understanding of major problems or other important details [27].

The method is conducted by first letting all participant start writing down all relevant details gathered through research on notes. Each participants may have their own unique color on their notes to make them easier to distinguish. The notes are all put on a wall and the participants can then start moving them trying to group them into relevant groups and come up with group titles and even subgroup titles if they feel the need.

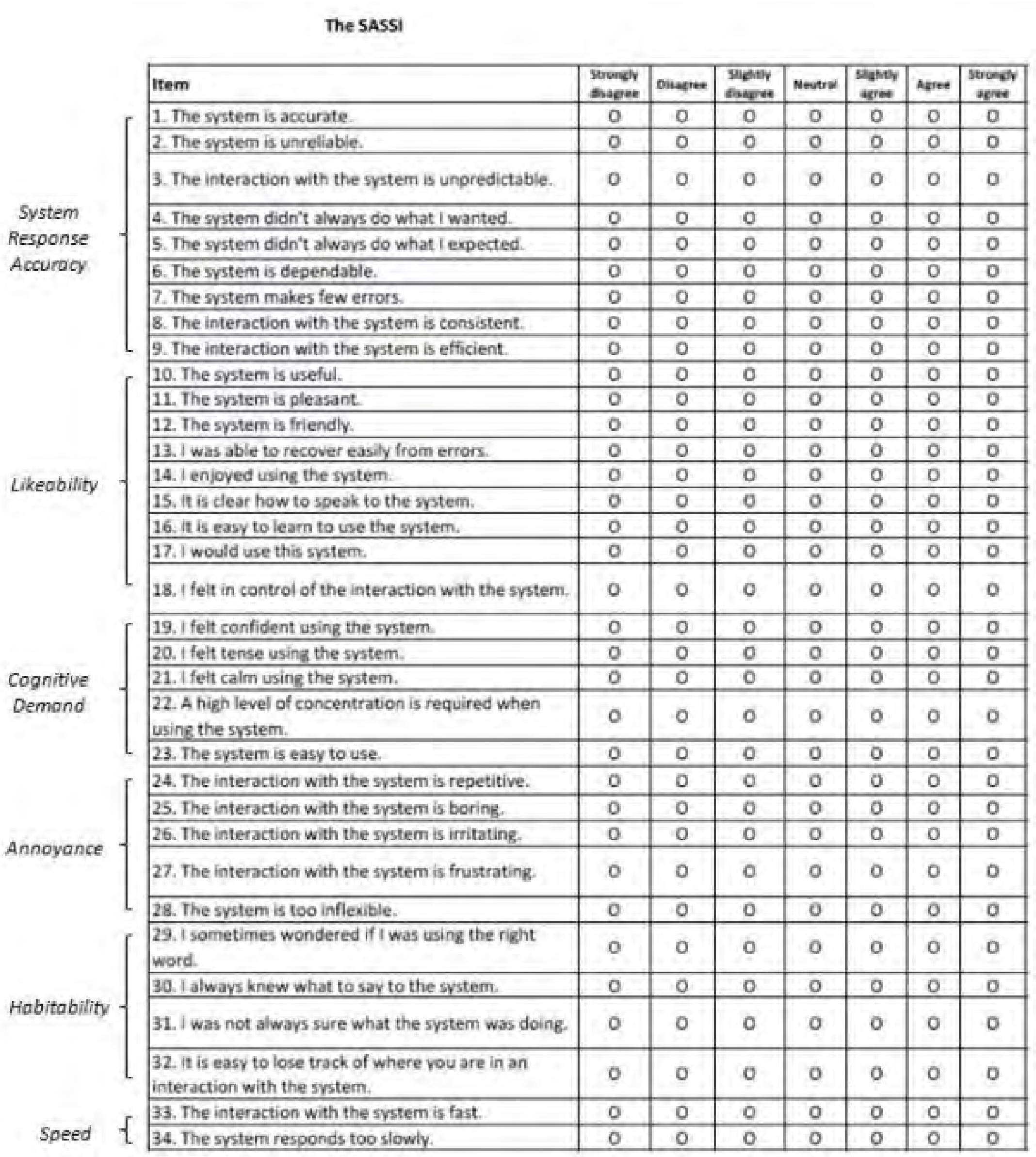
A popular method for making affinity diagrams is the *KJ method* [27]. The KJ method is done in a similar way as the above written description but with a big emphasis that talking is not allowed while writing, placing and organizing the sticky notes. No speaking allows all participants to minimize any possible influence of group pressure [27].

4.12 Wizard of Oz

The Wizard of Oz (WOz) technique is performed by simulating a working prototype or system by letting a researcher or a "wizard" operate and control the prototype from behind the scenes [27]. Developing a fully working prototype is time and resource intensive. The WOz technique allows researchers and designers to evaluate a design concept without having to spend as much resources as building a fully functional prototype would have demanded [27].

From the user's or test participant's perspective, WOz prototypes and implemented features are indistinguishable. This is achieved by preparing system responses for potential paths of interaction in advance, so that the prototype operator, or wizard, can quickly respond to user input. For WOz prototypes to be successful, the prototype operator must be able to see or hear the user so that appropriate responses can be provided based on user input. Moreover, users should be unaware that the prototype operator is controlling the WOz prototype.

The WOz technique has a long history with the development of speech recognition and voice user interfaces [16]. WOz prototypes can be used throughout the design process of voice interfaces and is invaluable for resource for understanding users' vocabulary, utterance structures, and interactive patterns [16]. While WOz prototypes allow voice interface designers to bypass developing speech recognition systems to evaluation a design, the value of designed errors is not to be discounted. In fact, there are tools for creating WOz prototypes that randomly assign speech recognition errors to understand user reactions to such scenarios [23].



The SASSI

	Item	Strongly disagree	Disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree
System Response Accuracy	1. The system is accurate.	○	○	○	○	○	○	○
	2. The system is unreliable.	○	○	○	○	○	○	○
	3. The interaction with the system is unpredictable.	○	○	○	○	○	○	○
	4. The system didn't always do what I wanted.	○	○	○	○	○	○	○
	5. The system didn't always do what I expected.	○	○	○	○	○	○	○
	6. The system is dependable.	○	○	○	○	○	○	○
	7. The system makes few errors.	○	○	○	○	○	○	○
	8. The interaction with the system is consistent.	○	○	○	○	○	○	○
	9. The interaction with the system is efficient.	○	○	○	○	○	○	○
Likeability	10. The system is useful.	○	○	○	○	○	○	○
	11. The system is pleasant.	○	○	○	○	○	○	○
	12. The system is friendly.	○	○	○	○	○	○	○
	13. I was able to recover easily from errors.	○	○	○	○	○	○	○
	14. I enjoyed using the system.	○	○	○	○	○	○	○
	15. It is clear how to speak to the system.	○	○	○	○	○	○	○
	16. It is easy to learn to use the system.	○	○	○	○	○	○	○
	17. I would use this system.	○	○	○	○	○	○	○
	18. I felt in control of the interaction with the system.	○	○	○	○	○	○	○
Cognitive Demand	19. I felt confident using the system.	○	○	○	○	○	○	○
	20. I felt tense using the system.	○	○	○	○	○	○	○
	21. I felt calm using the system.	○	○	○	○	○	○	○
	22. A high level of concentration is required when using the system.	○	○	○	○	○	○	○
	23. The system is easy to use.	○	○	○	○	○	○	○
Annoyance	24. The interaction with the system is repetitive.	○	○	○	○	○	○	○
	25. The interaction with the system is boring.	○	○	○	○	○	○	○
	26. The interaction with the system is irritating.	○	○	○	○	○	○	○
	27. The interaction with the system is frustrating.	○	○	○	○	○	○	○
	28. The system is too inflexible.	○	○	○	○	○	○	○
Habitability	29. I sometimes wondered if I was using the right word.	○	○	○	○	○	○	○
	30. I always knew what to say to the system.	○	○	○	○	○	○	○
	31. I was not always sure what the system was doing.	○	○	○	○	○	○	○
	32. It is easy to lose track of where you are in an interaction with the system.	○	○	○	○	○	○	○
Speed	33. The interaction with the system is fast.	○	○	○	○	○	○	○
	34. The system responds too slowly.	○	○	○	○	○	○	○

Figure 4.3: The SASSI questionnaire [24]

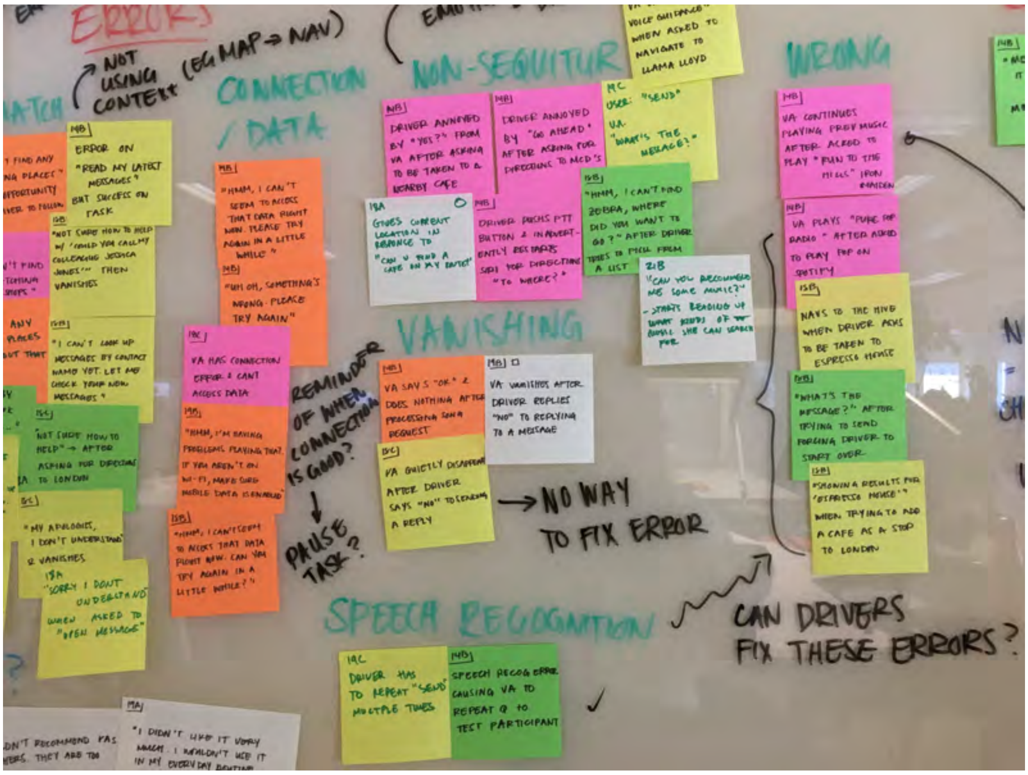

ERRORS
NOT USING CONTEXT (EG MAP → NAV)
CONNECTION / DATA
NON-SEQUITUR
WRONG
VANISHING
SPEECH RECOGNITION
REMINDER OF WHEN CONNECTION IS GOOD?
PAUSE TASK?
→ NO WAY TO FIX ERROR
CAN DRIVERS FIX THESE ERRORS?
"HMM, I CAN'T SEEM TO ACCESS THAT DATA RIGHT NOW. PLEASE TRY AGAIN IN A LITTLE WHILE."
"UH OH, SOMETHING'S WRONG. PLEASE TRY AGAIN"
VA HAS CONNECTION ERROR & CANT ACCESS DATA
DRIVER ANNOYED BY "YES?" FROM VA AFTER ASKING TO BE TAKEN TO A NEARBY CAFE
DRIVER ANNOYED BY "GO AHEAD" AFTER ASKING FOR DIRECTIONS TO MCD'S
VA CONTINUES PLAYING PROG MUSIC AFTER ASKED TO PLAY "RUN TO THE HILLS" IRON MAIDEN
VA SAYS "OK" & DOES NOTHING AFTER PROCESSING A REQUEST
VA VANISHES AFTER DRIVER REPLIES "NO" TO REPLYING TO A MESSAGE
VA QUIETLY DISAPPEARS AFTER DRIVER SAYS "NO" TO SENDING A REPLY
DRIVER HAS TO REPEAT "SEND" MULTIPLE TIMES
SPEECH RECOG ERROR CAUSING VA TO REPEAT Q TO TEST PARTICIPANT
"I DIDN'T LIKE IT VERY MUCH. I WOULDN'T USE IT IN MY EVERYDAY ROUTINE"

Figure 4.4: Affinity diagram (partial) used to analyze qualitative data

5

Process

This chapter describes the research and design process carried out as part of this thesis work. Several methods were carried out as part throughout the process, and the specifics of those methods with respect to the purpose of this project are discussed here. For details about the methods themselves, see Chapter 4: Methods.

5.1 Pre-study and Preparation

The pre-study phase was the first phase of the research project and focused on a review of related research. This pre-study was done to understand what research has already been done and what gaps in the research exist which this project could aim to answer.

In addition to developing a contextual understanding of the research area, the pre-study helped to identify methods and test setups that are frequently used when examining distracted driving in terms of visual distraction and cognitive load. Methods for data analysis and theories related to attention and cognition were also identified. The knowledge gathered during the pre-study phase was used to plan the project execution.

5.2 Project Planning

The project planning phase was focused on developing a schedule for the execution of the project. The distribution of time between the pre-study and preparation, project execution, and project finalization phases were based on recommendations from thesis examiners at Chalmers University of Technology and spread over a 20 week time period. During the project planning phase, methods were selected for their suitability to the research question developed in the pre-study phase. A full GANTT schedule of the project process with calender week numbers can be seen in Appendix A.

5.3 Literature Review of Existing Guidelines

During the pre-study phase, a brief review of three existing design guidelines was done to gain a general understanding of what each set of guidelines covered with respect to voice assistant interaction in vehicles. The guidelines reviewed in the pre-study stage were *Android Auto Design Guidelines*, *Apple CarPlay Human Interface Guidelines*, and *Google Conversation Design* [2, 1, 14, 13]. These guidelines were selected for review since they are directly tied to the two commercially available integration interfaces.

To answer the first sub-question of the research question and understand what current guidelines exist for voice assistant interaction in vehicles, a more in-depth literature review of the existing guidelines was required. This literature review aimed to summarize and understand the collective wisdom of the industry when it comes to in-vehicle voice assistant interaction. In addition to the guidelines once reviewed during the pre-study phase, this literature review also included *Amazon Alexa Design Guide* [1]. Although Amazon does not have an integration interface on the market, it has announced plans to do so in the coming years. A review of existing NHTSA guidelines was also done, as those guidelines specifically deal with traffic safety [30, 31].

The results of this literature review would be used in later phases to identify established guidelines which work well to decrease visual distraction and cognitive load. The review was also used to identify areas where the guidelines were not followed by existing voice assistants and to identify gaps in the guidelines with respect to distracted driving and voice assistant interaction.

5.4 Summative Evaluation

In order to understand the efficacy of the existing guidelines, a summative evaluation of existing voice assistants in vehicles was conducted. The summative evaluation also served to identify any issues in the voice assistant integrated IVI that may contribute to visual distraction and increased cognitive load, thereby decreasing safe driving. The evaluation consisted of three parts: an on-road test, data collection and handling, and analysis of data collected from the test.

A total of 8 test participants completed the on-road test. A ninth participant began an on-road test, but the test was ended prematurely due to concerns for traffic safety.

The 8 test participants had been licensed drivers for a mean time of 12.6 years. Frequency of driving was evenly spread out among participants between driving every day to less than once a month. Only two participants had previous experience with driving with PA, the Level 2 ADS used in the test. All but one participant had previous experience with VAs and the large majority of these previous experience were with VAs on smartphones, a screen-first solution.

5.4.1 On-road Test Setup

The on-road test was done on public roads in Torslanda, Gothenburg. Test participants drove along a predefined route that measured 10.3 kilometers with roundabouts at each end which made for a continuous driving experience. Speed limits along the route varied between 50 and 70 kilometers per hour.

Figure 5.1: Interior of the test car model, equipped with Android Auto and Apple CarPlay

Test participants were recruited internally at Volvo but were not limited to employees. Students and consultants placed at Volvo were also invited. Due to liability issues with the test car, only employees, students, and consultants with Volvo access could participate. The test car used was a Volvo V90 with automatic transmission and Pilot Assist (PA), a lane keeping and adaptive cruise control feature which makes it a SAE Level 2 autonomous vehicle. The implementation of the PA feature on the test car is common to other Level 2 vehicles. The V90 test car was also equipped with both Android Auto and Apple CarPlay.

The on-road test was designed to compare the performance of the two voice assistants, Apple Siri and Google Assistant. The test was also designed to determine if there was a decrease in visual distraction and cognitive load when test participants were aided by Pilot Assist. Thus, there were four conditions for test participants to complete:

- Android Auto with manual driving
- Android Auto with Pilot Assist
- Apple CarPlay with manual driving
- Apple CarPlay with Pilot Assist

Under each condition, test participants were asked to perform 9 secondary tasks while driving. These tasks were selected due to their relation to categories of app enabled on integration interfaces. Moreover, functionality for all tasks exist on both voice assistants tested. The tasks were:

1. Open a new received text message
2. Send text message to a contact
3. Make a call to a contact
4. Make a call to a contact with multiple phone numbers
5. Play a genre of music
6. Play a specific song by a specific artist
7. Start navigation to a street address
8. Add a café to the current route
9. Start navigation to the nearest McDonald's

For each test, participants began by signing a consent form for their data to be collected and used for this project's research. Next, they completed a survey about their previous experience with driving and using voice assistants. On the drive from Volvo Headquarters to the designated test route, test participants were trained on using PA and had a chance to get familiar with driving the car, with and without PA. Then, the test participants performed the 9 secondary tasks while driving along the test route for each of the four test conditions. The order of test conditions was randomized to minimize any bias from the order the test conditions were completed. Prior to starting each test condition, test participants were given training on the voice assistant for each condition in relation to the types of tasks they would be asked to perform. The order of the 9 tasks was not randomized, since some tasks built upon the output of a previous task and the overall difference in voice assistants was the focus of each test condition. After completing each driving condition, test participants were asked to complete a DALI survey to assess the cognitive load of each test condition. After the test, participants were asked a set of follow-up questions about their overall experience in a semi-structured interview.

For each task, participants were permitted up to 3 attempts in the case of task failure. Task failure is defined as the end of an interaction with the VA that does not trigger the desired intent or action. Task success is defined as the successful completion of a task using the voice assistant. For example, an utterance for Task 7 which results in navigation to the wrong address would be considered a task failure. Test participants were not required to make repeated attempts in the case of task failure.

The survey about the participant's previous experience can be found in Appendix B. A complete protocol of the on-road test can be found in Appendix C. The test schedule and the randomized condition permutations can be found in Appendix D.

5.4.2 Data Collection and Handling

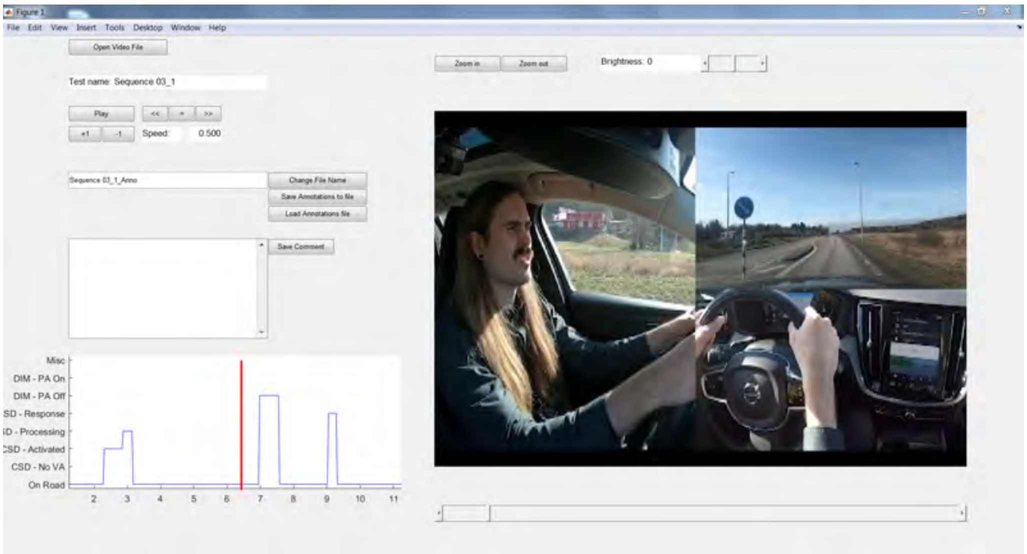
Visual distraction during the on-road test was measured by manual eye tracking, which is further described in section 4.10. The primary focus of this data was to distinguish between on-road and off-road glances. Moreover, cognitive load was assessed using DALI surveys completed during the on-road test. The DALI survey used can be found in Appendix I.

Eye glance data was collected during the on-road test via three video cameras mounted throughout the car. The cameras recorded a view of the driver's face, the IVI display, and the road. The three views of the cameras can be seen in Figure 5.2. Audio was included in the video recordings. These videos were then synchronized for each condition. The synchronized videos made it possible to code the eye glances of the test participant and understand that context of glances with the added road and IVI views.

The synchronized video was then manually coded through a custom tool created in Matlab, one task at a time. The software used is seen in Figure 5.2. Task eye glance analysis began from the end the test facilitator's prompt to complete the task to task success or the end of the last attempt to complete the task. Thus, task footage analyzed may include more than one attempt to complete a task. The tool allowed the video to be analyzed at a 30 Hz frequency. The tool made it possible to assign a glance code to each frame analyzed. Once the glance codes were assigned, duration for each glance was calculated in preparation for data analysis.

Several codes were used to annotate the eye glance data. These codes were:

0. **On-road** Glances on the road



Test name: Sequence 03_1

Speed: 0.500

Sequence 03_1_Anno

Brightness: 0

Figure 5.2: Eye tracking software and video with three camera views

1. **IVI VA Inactive** Glances at the IVI when the VA is not active

2. **IVI VA Active** Glances at the IVI when the VA is actively awaiting a driver utterance

3. **IVI VA Processing** Glances at the IVI when the VA is processing an utterance

4. **IVI VA Response** Glances at the IVI when the VA is presenting a response or prompt

5. **DIM PA Off** Glances at the driver information module (DIM) when PA is off

6. **DIM PA On** Glances at the DIM when PA is on

7. **Miscellaneous** Glances that are directed at the road, IVI, or DIM

DALI data was collected for each test participant, for each test condition, totally 4 completed DALIs per test participant. Adjusted ratings from each DALI were calculated for the individual dimensions of the DALI. Combined, the adjusted ratings resulted in a weighted rating also used in later data analysis. DALI scores were weighted according to the established protocol described in Chapter 4.

In addition to the quantitative data collected above, qualitative observational data was also collected. The synchronized videos were reviewed and qualitative observations, such as emotional reactions and scenarios of high frustration, were recorded. Test participant answers from the debrief interview were also transcribed for qualitative analysis.

5.4.3 Data Analysis

The data analysis was done in two parts: qualitative and quantitative. The qualitative analysis deals with observational notes of the on-road test and transcribed answers from the debrief interview. The quantitative analysis concerns the eye glance and DALI data.

Qualitative data from observation notes and interview transcriptions were combined and analyzed using the affinity diagramming method. This allowed for connections between different data points and recurring themes in the data to be identified. The insights from this analysis would later guide the development of new design guidelines and actualizations of these guidelines as prototypes.

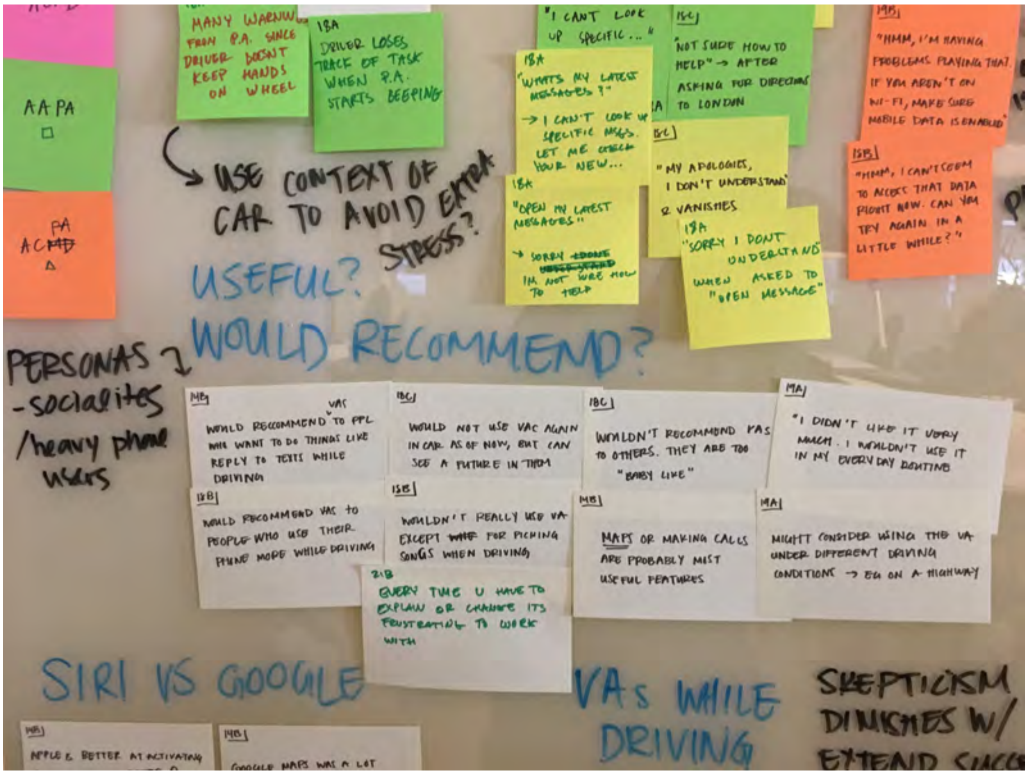

AA PA
□
PA
AC~~AD~~
△
MANY WARNINGS FROM P.A. SINCE DRIVER DOESN'T KEEP HANDS ON WHEEL
18A
DRIVER LOSES TRACK OF TASK WHEN P.A. STARTS BEEPING
USE CONTEXT OF CAR TO AVOID EXTRA STRESS?
USEFUL?
WOULD RECOMMEND?
PERSONAS
-socialites
/heavy phone users
"I CANT LOOK UP SPECIFIC..."
18A
"WHATS MY LATEST MESSAGES?"
→ I CAN'T LOOK UP SPECIFIC MSGS. LET ME CHECK YOUR NEW...
18A
"OPEN MY LATEST MESSAGES"
→ SORRY ~~I DONT UNDERSTAND~~ IM NOT SURE HOW TO HELP
15C
"NOT SURE HOW TO HELP" → AFTER ASKING FOR DIRECTIONS TO LONDON
15C
"MY APOLOGIES, I DON'T UNDERSTAND" & VANISHES
18A
"SORRY I DONT UNDERSTAND" WHEN ASKED TO "OPEN MESSAGE"
19B
"HMM, I'M HAVING PROBLEMS PLAYING THAT. IF YOU AREN'T ON WI-FI, MAKE SURE MOBILE DATA IS ENABLED"
15B
"HMM, I CAN'T SEEM TO ACCESS THAT DATA RIGHT NOW. CAN YOU TRY AGAIN IN A LITTLE WHILE?"
14B
WOULD RECOMMEND VAS TO PPL WHO WANT TO DO THINGS LIKE REPLY TO TEXTS WHILE DRIVING
18C
WOULD NOT USE VAC AGAIN IN CAR AS OF NOW, BUT CAN SEE A FUTURE IN THEM
18C
WOULDN'T RECOMMEND VAS TO OTHERS. THEY ARE TOO "BABY LIKE"
17A
"I DIDN'T LIKE IT VERY MUCH. I WOULDN'T USE IT IN MY EVERYDAY ROUTINE
18B
WOULD RECOMMEND VAS TO PEOPLE WHO USE THEIR PHONE MORE WHILE DRIVING
15B
WOULDN'T REALLY USE VA EXCEPT ~~WHE~~ FOR PICKING SONGS WHEN DRIVING
21B
EVERY TIME U HAVE TO EXPLAIN OR CHANGE ITS FRUSTRATING TO WORK WITH
14B
MAPS OR MAKING CALLS ARE PROBABLY MOST USEFUL FEATURES
17A
MIGHT CONSIDER USING THE VA UNDER DIFFERENT DRIVING CONDITIONS → EG ON A HIGHWAY
SIRI VS GOOGLE
14B
APPLE & BETTER AT ACTIVATING
14B
GOOGLE MAPS WAS A LOT
VAs WHILE DRIVING
SKEPTICISM DIMINISHES W/ EXTEND SUCC

Figure 5.3: Affinity diagram (partial) of qualitative summative evaluation data

The quantitative data was analyzed using Minitab, a statistical data analysis software. Eye glance and DALI data was plotted in order to identify any trends in the data. The plots where also used to help determine whether the different VAs had a discernible difference on visual distraction and cognitive load. The plots were also used to determine if the use of SAE Level 2 ADS, in this case PA, also had an effect on visual distraction or cognitive load. The results from the quantitative data analysis were also used to support and motivate new design guidelines for voice assistant interaction in vehicles.

5.5 Prototype Development and Evaluation

Following the summative evaluation, ideation for improvement to address the issues identified in the summative evaluation began. The ideation process resulted in two prototypes, Prototype 1 and 2. These prototypes embodied interstitial, new guidelines for voice assistant interaction in vehicles. These prototypes were then tested in a simulator. The results from the simulator test were then collected and analyzed.

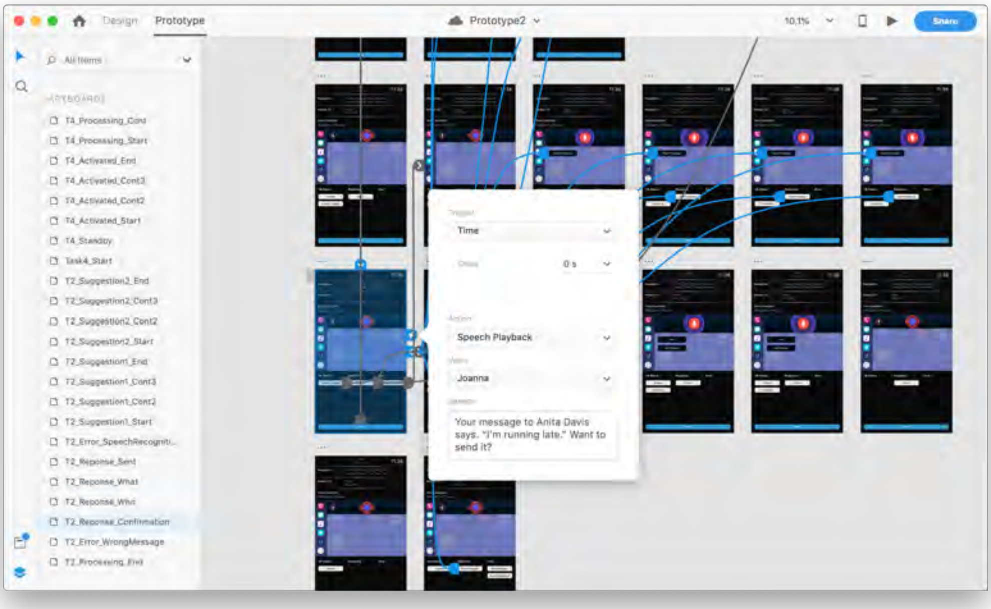
Both prototypes were tested by 10 participants, but eye glance data is only available for 9 participants due to file corruption. The participants has been licensed drivers for a mean time of 9.1 years. Participants were mostly infrequent drivers, driving every other week or less. Three test participants drove at least once a week. All but two participants had previous experience with VAs. The majority of participants had previously experienced VAs on smartphones.

5.5.1 Ideation and Prototype Development

The aim of the ideation process was to come up with potential solutions to improve the problems in the existing guidelines and voice assistants. The ideation process began by narrowing the number of tasks that would be performed by the test participant during the simulation test. Tasks from the on-road test were recycled, but tasks which had little interaction and little data results, were removed, such as the task of calling a contact by name. A voice-first approach was taken, the ideation process focused first on generating many conversation dialogs for the test tasks. Eventually, two concepts emerged, which will be labeled as Prototype 1 and Prototype 2. Conversation dialogs for both prototype were further refined to reflect the two concepts. This includes having multi-turn and one-shot dialogs for applicable tasks. Moreover, since errors were found to be a factor in visual distraction and cognitive in the summative evaluation, error handling was a key re-designed element in both prototypes.

Both prototypes were then implemented in Adobe XD, as shown in Figure 5.4, which allows designers to assign a pre-written system response for each screen. Adobe XD is able to output both visual information as well as speech. The prototypes were developed with high-fidelity graphics for the purposes of using the Wizard of Oz (WOz) technique when testing the prototypes. The two prototypes shared a visual language, in order to focus on discerning any differences or preferences between to two voice interaction concepts developed.

In order to later use the WOz technique with the prototypes, a control panel was designed for both prototypes. The control panel allows the wizard to control the flow of system responses to test participant input. During the simulator test, the



Time

0 s

Speech Playback

Joanna

Your message to Anita Davis says. "I'm running late." Want to send it?

Figure 5.4: Prototype development in Adobe XD

control panel would be hidden from the test participant.

5.5.2 Simulator Test Setup

The two prototypes were tested in a truck simulator located at Chalmers Johanneberg campus. Test participants drove along a highway, following in-game navigation directions in the trucking-driving simulation game Euro Truck Simulator 2. The game included multiple lanes of traffic, which participants were free to switch between while avoiding collision with any of the other in-game vehicles.

Test participants were recruited through an online survey. Participants must have a valid driver's license to participate. Upon completing the test, participants were compensated with a gift card for 250 kronor.

The simulator setup included a large TV display positioned in front of the driver's seat. The seat was a full adjustable car seat, which helped to acclimate experienced drivers to use the simulator. The simulator was also equipped with steering wheel, gear shift, and pedal game controls to drive the truck. The setup can be seen in Figure 5.5. The steering wheel was equipped with some force feedback to simulate bumps in the road and the simulator was set to an automatic transmission. The simulator was not equipped with any autonomous driver features.

A Windows Surface Book was mounted to the right of the steering wheel to simulate the IVI. The simulated IVI displayed the two tested prototypes, one at a time, and

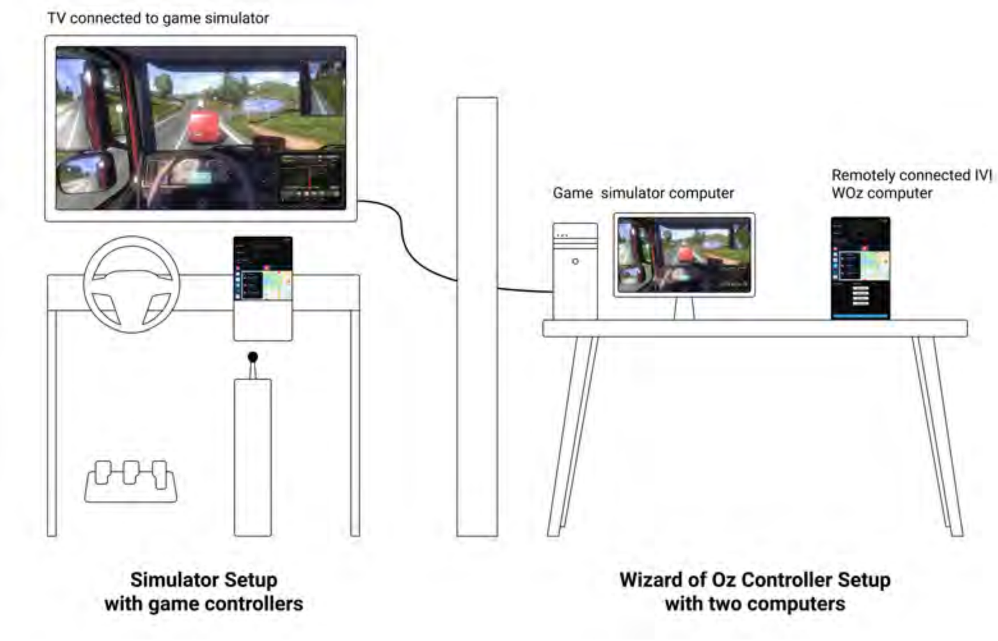

TV connected to game simulator
Game simulator computer
Remotely connected IVI WOz computer
Simulator Setup with game controllers
Wizard of Oz Controller Setup with two computers

Figure 5.5: Simulator test setup with a dividing wall between the test participant and wizard (not to scale)

was connected to the wizard computer by remote desktop. This allowed the wizard to control the prototype's responses to test participant input on the fly. The wizard was situated in the same room as the test participant and test facilitator, but behind a partition so participants were not aware that the wizard was controlling the IVI prototype. Only one person acted as the wizard to minimize systematic bias. As shown in Figure 5.5, the control panel of the IVI prototype was hidden from the test participant but visible to the wizard. This control panel on the prototypes allowed the wizard to remotely control the prototypes in real time, in direct response to the utterances made by the test participant.

The simulator test was designed to compare the two prototypes developed based on findings from the summative evaluation. There were two test conditions for test participants to complete:

- Prototype 1 with manual driving

- Prototype 2 with manual driving

Under each condition, test participants were asked to perform 7 tasks. These tasks were re-used from the previous evaluation. Tasks 3 and 6 were removed from this evaluation since they are typically done in a single-shot interaction and offer little data as to how users interact with voice assistants. The tasks, using the same number as the corresponding tasks of summative evaluation, were:

1. Open a new received text message

2. Send a text message to a contact

4. Make a call to a contact with multiple phone numbers

5. Play a genre of music

7. Start navigation to a street address

8. Add a café to the current route

9. Start navigation to the nearest McDonald's

For each test, participants began giving their informed consent and completing a survey about their previous driving and voice assistant experience. Then they were given a walk-through of the driving simulator, after which they drove the simulator for ten minutes to get used to simulator driving. The participants then got to perform the secondary tasks listed above for one of the test conditions. The order of the test conditions were randomly assigned to minimize bias. Prior to each test condition, participants were given training for the voice assistant for each prototype, with special attention paid to any differences in possible interactions between the two prototypes. The order of the 7 tasks was not randomized, since the focus of the test was discerning differences between the systems, not the tasks. After completing each test condition, participants were asked to complete a DALI. Test participants were also asked to to fill out a SUS for each prototype. After both test conditions were completed, test participants were asked a set of follow-up questions in a semi-structured debrief interview. The full test protocol can be found in Appendix E.

Throughout the test, a set number of randomized errors were scripted into each participant's interaction with both prototypes. The randomized errors were evenly distributed across tasks. A maximum of one error was scripted into a task. The scripted errors included:

- **No recognition** the VA does not recognize the user's input at all
- **No function** the VA recognizes and understands part of the user's input
- **Bad connection** the VA is unable to complete a task due to an insufficient data connection

The survey about the user's previous experiences can be found in Appendix F. For a full list of the different interactions possible during the different tasks, see Appendix G. For the randomized testing schedules containing the planned errors, see Appendix H.

5.5.3 Data Collection and Handling

Like in the summative evaluation, visual distraction was measured in the simulator test using manual eye-tracking. The purpose of the data was to distinguish between on-road and off-road glances. The prototype evaluation also uses the DALI surveys to understand the cognitive of the two prototypes. The DALI survey that was used can be found in Appendix I.

The eye glance data was collected during the simulator test via two cameras mounted in the simulator room with one directed at the test participant's face and the other at the TV showing the traffic situation and the IVI prototype display. A screen shot from a video showing the two camera views can be seen in Figure 5.6. Just as in the summative evaluation, the eye glance footage was synchronized, analyzed in MatLab, and prepared for analysis in MiniTab. In the simulator, the DIM is displayed on the screen. Due to the low precision of manual eye tracking and close proximity of the road and the DIM, both displayed on TV screen, glances between the DIM and the road are indistinguishable. Thus, glance codes relating to the DIM were not used and all on-screen glances were considered on-road. The same coding scheme from the summative evaluation was used in the analysis of the prototype evaluation footage.

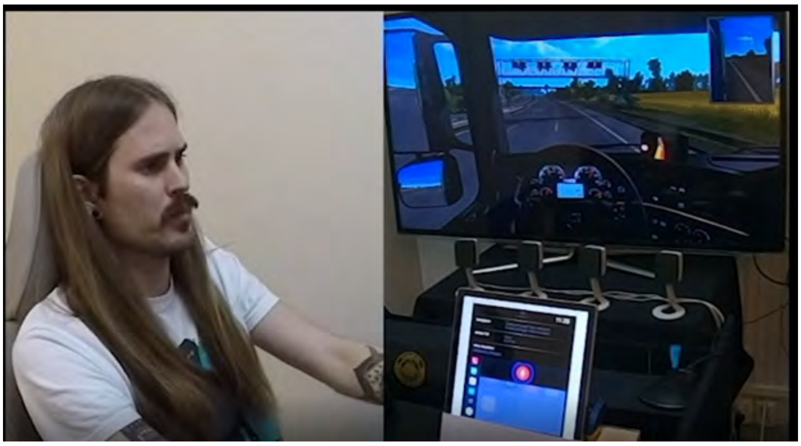


Figure 5.6: Video used for eye tracking

The DALI data from the simulator tests were prepared in the same manner as the DALI data from the summative evaluation. In this phase, the data was prepared for statistical analysis in Minitab.

Qualitative data was collected and gathered for analysis. As in the summative evaluation, footage from the simulator tests was reviewed and observational notes were taken regarding how test participants interacted with and reacted to the two prototypes. Notes of participant answers during the debrief interview were taken during the interview.

Unlike the summative evaluation, the prototype evaluation also includes SUS scores for each prototype, given by each test participant. SUS was incorporated into the prototype evaluation to also measure test participants' acceptance and perception of usability for each prototype. The SUS survey can be found in Appendix J.

5.5.4 Data Analysis

Analysis of the data from the prototype evaluation was done in two parts: qualitative and quantitative. The qualitative analysis was done to identify common themes, problems with the prototypes, and interaction patterns across participants. The purpose of the quantitative analysis is to determine if there is discernible difference between the two prototypes and how they compare to each other with respect to visual distraction, cognitive load, and usability.

The qualitative data from the debrief interviews and test observations were analyzed using an affinity diagram. An affinity diagram is suitable for organizing qualitative data and drawing insights from clusters of qualitative data. The insights from this analysis would later guide the finalization of this work's suggestions for new guidelines for voice assistant interaction in vehicles.

Figure 5.7: Affinity diagram (partial) of qualitative prototype evaluation data

The quantitative analysis concerns the eye glance, DALI, and SUS data. Quantitative data was organized into tables and analyzed using Minitab. Glances frequency and the relation between on- and off-road glances were analyzed. The weighted values of each prototypes DALI and SUS scores were analyzed in Minitab. The DALI and SUS scores were calculated according to the standard methodology for each respective method.

5.6 New Guidelines

The data analysis results and consequent findings were used to develop a new set of guidelines for designing voice assistant interaction in vehicles. This was done by comparing the results of the two evaluations with the existing guidelines and assessing whether the existing guidelines were sufficient. In cases where they were deemed insufficient, the guidelines were altered to better suit the context of in-vehicle voice assistant interaction. Guidelines were deemed insufficient if they were correlated with an increase in visual distraction or cognitive load.

Both the summative and prototype evaluations uncovered design challenges not covered under the existing guidelines. In this case, new guidelines were added. These new guidelines include suggestions on what designers should and should not do when designing for in-vehicle VAs.

The resulting new guidelines answer the research questions posed at the start of this thesis project.

6

Result

This chapter describes the results of methods and evaluations carried out during the project. This chapter also discusses some of the implications of the results.

6.1 Literature Review of Existing Guidelines

A review of the existing guidelines determined that the current industry guidelines do not provide guidance for designing voice assistant interaction in vehicles. While most of the guidelines reviewed acknowledge multiple modal interaction as part of voice interaction or interaction with an IVI, specific guidelines for these multimodal interactions are nearly non-existent. One exception is *Google Conversation Design* which supplement its guidelines with many multi-modal examples [14]. However, all of these multimodal examples are shown on the mobile phone or at-home smart screens. There are no guidelines for an integrated Google Assistant experience on Android Auto.

The lack of an integrated set of guidelines or holistic voice-manual experiences is due in part to the separation of modalities in guidelines. The Android Auto and Apple CarPlay guidelines primarily focus on the visuals of the integrated IVI and manual interaction with the IVI [2, 13]. The recommendations from these guidelines echo the safety recommendations of the NHTSA guidelines with added considerations for information perception and organization [30]. For example, recommendations to create high-contrast VUIs with only essential information displayed.

On the other hand are the Amazon Alexa and Google Conversation design guidelines which focus on voice interaction and structuring prompts and responses [1, 14]. Both guidelines are based on a version of Grice's Maxims [15], Google more explicitly so. Furthermore, both guidelines advocate a voice-first approach, assuming the user is unable to have regular view of the information on the accompanying screen.

Overlapping guidelines and themes were identified in the review of the four industry guidelines. The themes identified by the review are designing car apps, voice and manual input, general response, situation awareness, presenting choice, error handling, discoverability, display, and notifications. A complete list of the guidelines identified by the literature review as relevant to the design of in-vehicle voice assistant interaction are available in Appendix M.

6.1.1 Designing Car Apps

Car versions of apps should focus on the main information or features of the original mobile app or website. Limiting the features available helps to minimize distraction while driving. Car versions of apps may add car-specific functionality if it will benefit the driver.

6.1.2 Voice and Manual Input

Designers should create multi-turn dialogs for beginner voice interface users and single-shot commands for experts. Voice input is not suited for entering complex answers, like an alphanumeric password. With voice input, assume users may reference anything on screen as part of an utterance. For drivers, design IVI interfaces so that minimal touch interaction is necessary. For both voice and manual input, it is important that inputs to actions and intents are flexible and adaptive to uses, especially when it comes to voice utterances.

6.1.3 General Voice Responses

Responses from a connected voice assistant should use conversational language and be concise. Designers can use prompts to guide users when it is their turn in the dialog. Responses should be contextually relevant to a user's voice input.

6.1.4 Situation Awareness

Keep track of previous utterances in a dialog to build a contextual understanding of generic references like "it" or "there." Information collected from previous dialog may be used to shorten and guide subsequent interactions. Utilize the available technology, like location services, to provide relevant and personalized responses.

6.1.5 Presenting Choice

When presenting users with a decision, use a set of clear, simple options to avoid user confusion and unexpected answers. Narrow-focus questions can be used to set expectations and help users provide an appropriate answer to complete a task. Lists enable users to selection one out of several items, but should be presented by voice with only essential content related to content selection, such as the title of an item.

6.1.6 Error Handling

Error handling is used to minimize attention to an error and guide users to an appropriate utterance to trigger the right action or intent. Use variations on a prompt to provide additional details that a user might have missed or misunderstood when the prompt was first presented. Dialogues for No Input and No Match errors should be designed for every turn in a dialog.

6.1.7 Discoverability

With voice interaction, it may be difficult to discover features especially in situations where users are unable to look at the screen often. Voice assistants can use hints that leads and helps users discover new features of an app. Suggestions and signposts throughout a dialog may also help users discover relevant features.

6.1.8 Display

IVI displays should follow NHTSA guidelines and prevent videos, animated images, and scrolling text from being displayed. Most items in a list should be presented without having to scroll and only visual components relevant to the driver's current task should be displayed. Minimize the required navigational levels to reach any content item or action on-screen.

6.1.9 Notifications

Notifications should be provided in a succinct and timely manner to drivers. The number of notifications should be limited to minimize content lists and make room for more relevant and newer content. Transactional notifications which enable drivers to engage in human-human interaction, function better in their daily lives, and control transient app states should be prioritized.

6.2 Summative Evaluation

This section cover the results of the summative evaluation. The results are here divided into qualitative results and quantitative results.

6.2.1 Qualitative Results

Analysis of the qualitative data from interviews and observations collected during the summative evaluation was analyzed using affinity diagramming, which identified several problem categories where in-vehicle VAs could be improved. The major insights from the qualitative data are:

- Error responses from the VA increased visual distraction
- Error responses deteriorate trust between the user and the VA
- Repetition of error messages is a source of irritation and annoyance
- Presenting information with only a single channel, e.g. showing results on screen but not reading them out, leave users confused as to how to proceed
- Drivers expect visually presented information can be referenced by voice
- Audio tones draw attention to the IVI, potentially increasing visual distraction
- Drivers' selective focus on the VA seemingly decreases as the driving task intensity increases, e.g. when entering a busy roundabout
- Data connection status of the VA is not easily visually perceived or readily audibly communicated
- Drivers expect quick VA responses and have trouble interpreting when a VA is still processing a request
- Drivers felt that using a Level 2 ADS increased the pressure to take their eyes off the road to monitor the ADS system

The above mentioned findings are only the main findings. For a full list of all findings of the qualitative analysis see Appendix K.

6.2.2 Quantitative Results

Eye glance data from all 8 participants for all 4 test conditions and all 9 tasks were analyzed in Minitab. Task eye glances were analyzed from the end of the test facilitator's prompt to complete a task to the end of the final attempt for the task or achieving task success. The data shows that there is a greater variance in the duration of the on-road glances in comparison to off-road glances. On-road glances measured a mean value of 2.05 seconds, with a median of 0.93s, and a standard deviation of 3.04 s. Off-road glances have a mean value of 0.66 s, with a median of 0.57 s, and a standard deviation of 0.53 s. This indicates that the off-road glances are generally shorter than on-road glances, which is expected.

In Figure 6.1, a histogram visualizing frequencies of off-road glance duration times can be seen.

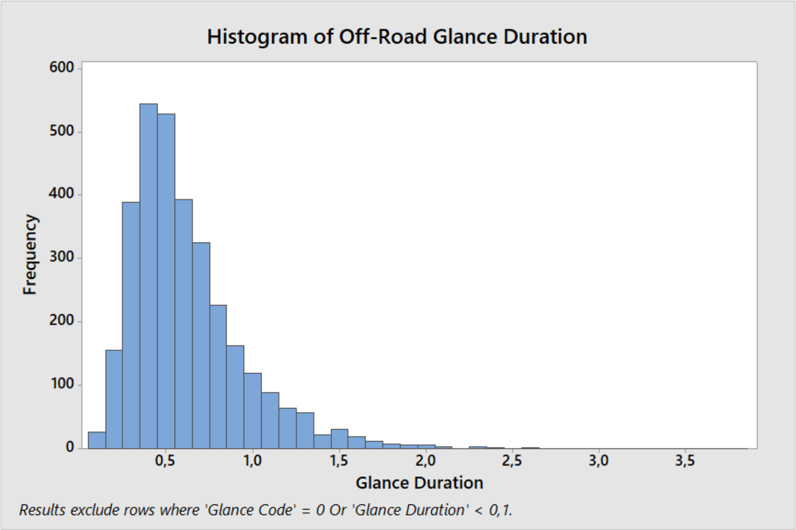


Figure 6.1: Frequency of off-road glances

The histogram depicted in Figure 6.1 has a clear peak between 0.35 to 0.55 seconds. This duration of off-road glances is quite short in comparison to similar studies. A previous report documenting glance duration times for manually performing secondary tasks on an HMI, including adjusting the radio, temperature gauge, and defroster settings had mean glance times above 1 second, more than twice the mean reported here [43]. The same study reported a mean glance duration of 0.62 s for glances toward the speedometer, requiring no manual interaction, which is more in line with off-road glances while using VAs. A separate study by Mehler et al. reported glance duration between 0.75s and 1s while manually making calls on a smartphone while driving [28]. These results indicate that off-road glances while interacting with VAs are comparatively short compared to glances coinciding with manual interaction.

The four test conditions, *Android Auto with manual driving* (AAMD); *Android Auto with Pilot Assist* (AAPA); *Apple CarPlay with manual driving* (ACMD); and *Apple*

CarPlay with Pilot Assist (ACPA), can be compared in a several ways.

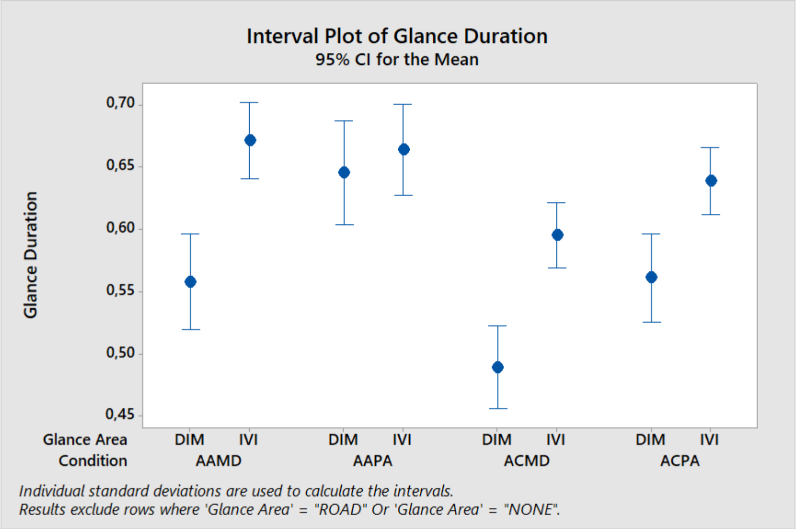


Figure 6.2: Intervals of DIM and IVI glances during four conditions

Glance duration intervals of DIM and IVI glances under each of the four testing conditions are shown in Figure 6.2. The duration intervals of the IVI glances overlap with each other under the four conditions. This suggest that IVI glances times are not directly affected by the use of PA, a Level 2 ADS. The DIM glance times generally increase when using the Level 2 ADS, which is to be expected as information regarding the status of the ADS is located on the DIM. These results support the qualitative findings where users reported an increase in checking the DIM when using an ADS. Figure 6.3 shows that the number for DIM glances increases when using the Level 2 ADS PA. However, there is no indication that the number of IVI glances is tied to either MD or PA driving conditions. Thus, while the use of a Level 2 ADS may increase off-road glances to the DIM, as expected, it does not conclusively increase off-road glances to the IVI.

Regarding the two VAs used in the summative evaluation, Figure 6.2 suggests that using Apple CarPlay (AC) requires less visual attention that Android Auto (AA), regardless of driving condition. This may be related to findings from the brief interviews where participants expressed that they felt Siri felt more human, was easier to talk to, and understood participants better when compared to Google Assistant in Android Auto.

Figure 6.4 shows the number of glances by test condition and glance code, where glance code marks the direction of the glance. Glance codes 5 and 6 respectively indicate a glance toward the DIM when the Level 2 ADS is off and when it is on. The number of glances towards the DIM under manual driving and PA reiterate that using a Level 2 ADS increases the number of off-road glances, specifically to the DIM where ADS status information is placed.

Furthermore, glance codes 1 and 2 respectively indicate glances toward the IVI

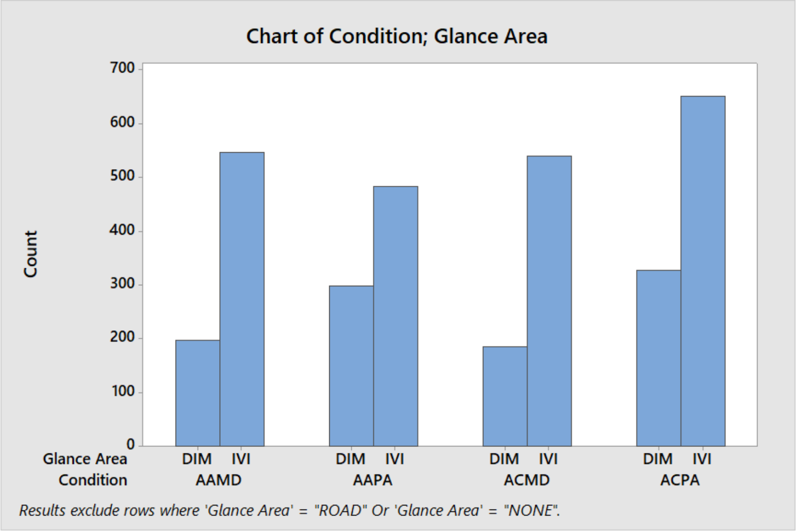


Figure 6.3: The count of DIM and IVI glances during the four conditions.

before the VA is activated or when it is actively listening. These glances reinforce the qualitative observation that drivers felt unsure about when the VA was listening and would use glances to IVI to make sure their utterances would be heard. The number of 1 and 2 glances in Figure 6.4 help to support this claim.

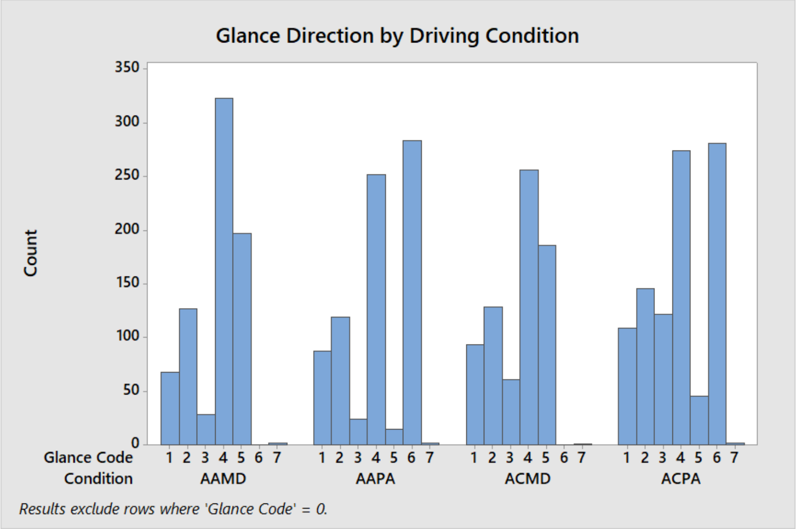
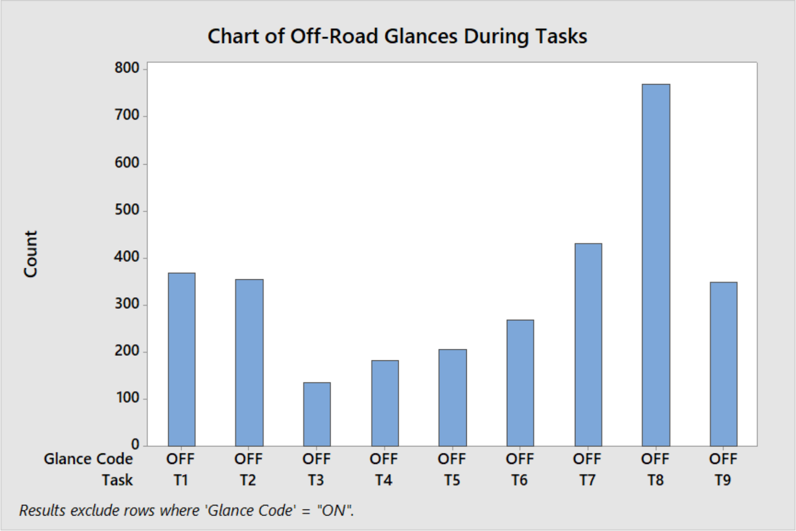


Figure 6.4: Count of off-road glances by direction and condition.

The glance codes shown in Figure 6.4 corresponds to: 1 - IVI VA Inactive, 2 - IVI VA Active, 3 - IVI VA Processing, 4 - IVI VA Response, 5 - DIM PA Off, 6 - DIM PA On, 7 - Miscellaneous. These are all further explained in section 5.4.2.

When considering the individual tasks performed, the number of off-road glances vary. Figure 6.5 shows the variation of counts of off-road glances over the different test tasks.

Tasks 3 through 6 amounted to the smaller number of glances, which correlates well



Results exclude rows where 'Glance Code' = "ON".

Figure 6.5: Count of off-road glances during the various test tasks.

to the qualitative observations. These tasks were more often than not, performed quickly by single-shot commands. Tasks 1 and 2 dealt with message while Tasks 7 through 9 dealt with navigation. Observational data showed that users consistently struggled more with navigational tasks more than any other types of tasks. The data in Figure 6.5 supports this claim with the higher counts of off-road glances in Tasks 7 through 9.

Findings from the qualitative analysis pointed to error as being the most important factor in increasing the number of off-road glances. Errors are defined as all situations where the VA fails to the trigger an intent from a user's utterance, regardless if it is the fault of the user or the system. This include speech recognition errors and data connection problems. Figure 6.6 show the number of glances for each task for situations with errors compared with those without. The figure supports the qualitative finding the errors increase the number of off-road glances.

A descriptive analysis of the DALI data was done using various box plots. The plots show which conditions the test participants believe demanded the most cognitive load. The plots also visualize other contextual data about participants' experience. These plots were used to determine which conditions demanded the greatest cognitive load from users and to understand what aspects influence their perception of cognitive load.

A total of 32 DALIs were completed by the 8 test participants, one for each test condition.

The weighted ratings of the DALIs showed that AA rated slightly higher than AC. This means that AA was rated as having a higher cognitive load than AC. This is visualized in Figure 6.7.

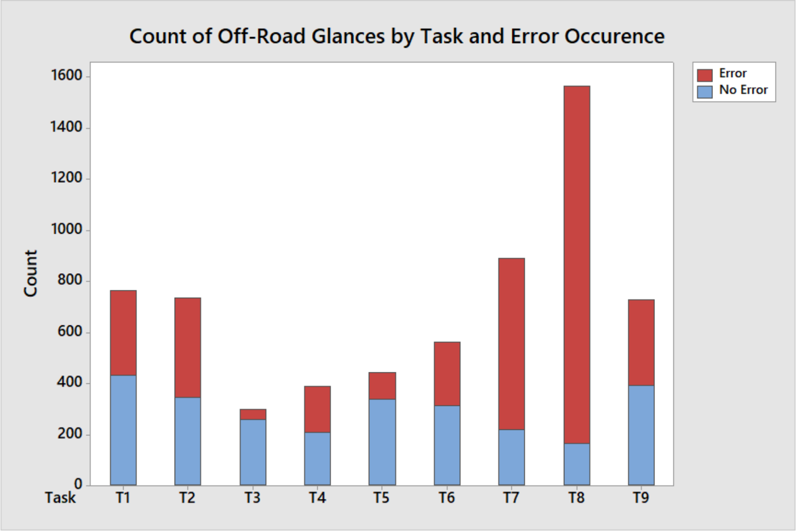


Figure 6.6: Count of off-road glances during tasks with error indications

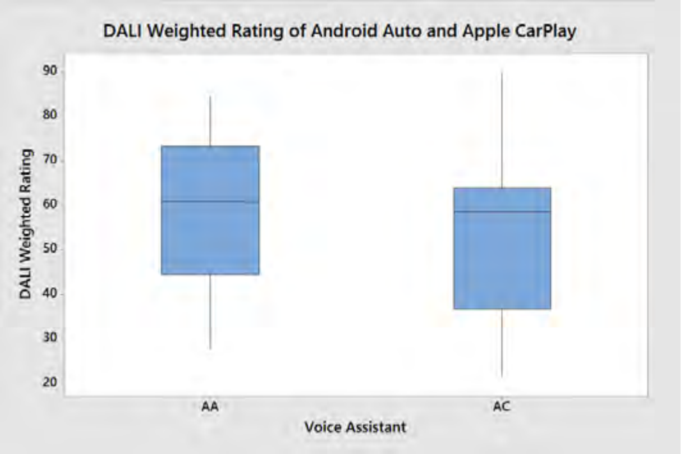


Figure 6.7: DALI weighted rating of Android Auto and Apple CarPlay

The adjusted ratings of the individual DALI dimensions were used to analyze which dimensions of the VA had a greater impact on cognitive load. The ratings for these dimensions are depicted in Figure 6.8. The ratings for the DALI dimensions are generally lower for AC with the exception of temporal demand. In the debrief interviews, test participants expressed greater acceptance of AC with Siri, because they felt the VA seemed smarter and allowed for more flexibility in interpreting utterances. Siri being perceived as smarter than Google Assistant may explain the lower ratings for visual and auditory demand, if they felt that Siri was easier to use and understood them better. The higher rating in temporal demand is reflected in results from the interview questions where several test participants noted that AA felt slower and allowed them to speak and interact with the VA at a slower pace.

Figure 6.8 shows that visual demand for both VAs is low, especially when compared to auditory demand. This suggests that test participants generally used audio and voice to complete the secondary tasks rather than the screen. This correlate with the

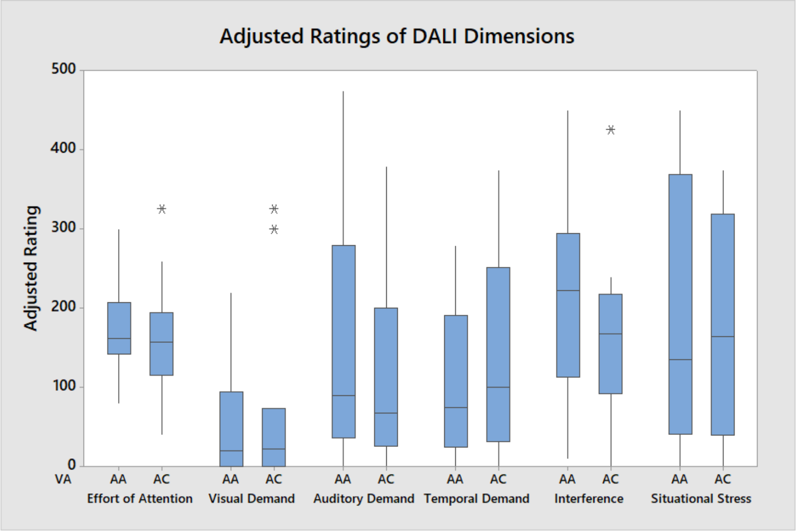


Figure 6.8: Adjusted ratings of the individual DALI dimensions

observations, where participants generally used utterances to trigger the intended unless they reached a tipping point where they were unsuccessful or felt they had to used manual input to complete a task.

The DALI ratings for PA, or Level 2 ADS, conditions shown in Figure 6.9 reinforce previous findings that the use of a Level 2 ADS increases cognitive load. This is indicated by the higher ratings shown in the figure for the Level 2 ADS tested. It also corresponds with interview answers where test participants stated they felt that they had to actively monitor the ADS to ensure it was active. However, only 2 participants had previous experience PA. Participants' lack of experience may be a contributor for the higher DALI ratings when using a Level 2 ADS.

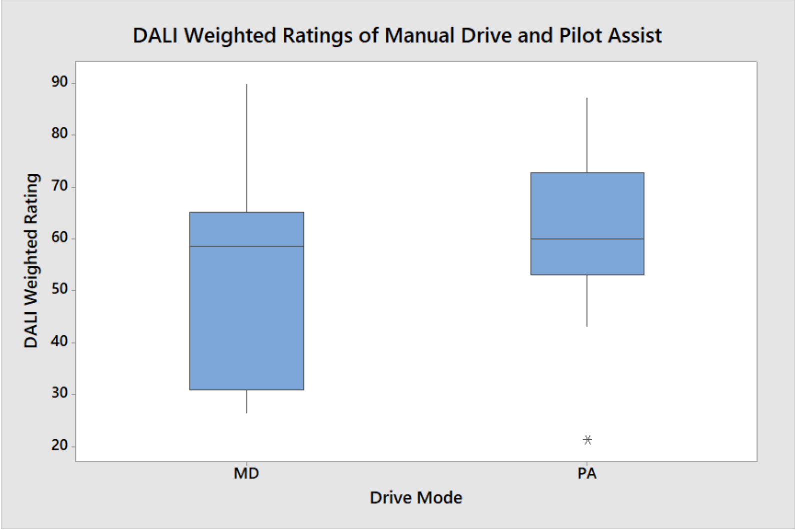


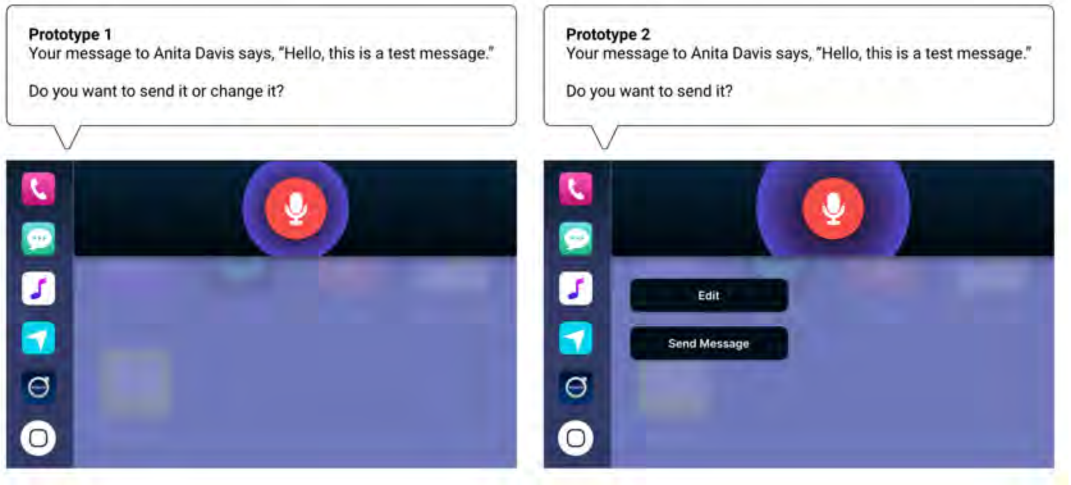
Figure 6.9: Weighted ratings of manual drive and pilot assist

6.3 Prototype Development and Evaluation

Two prototypes were developed to address the problems identified by the summative evaluation. These prototypes were then tested by users and analyzed.

6.3.1 Prototypes Developed

The prototypes developed as part of this project had a shared foundation and primarily differed in prompts and responses. The aim of *Prototype 1* was to facilitate more informed voice interaction. This was done with a more conversation approach, where the VA would provide the user with responses and prompts through voice with little visual support. The prompts for Prototype 1 heavily guided the conversation, especially in content selection scenarios, such as selecting an item from a list.



Prototype 1
Your message to Anita Davis says, "Hello, this is a test message."

Do you want to send it or change it?

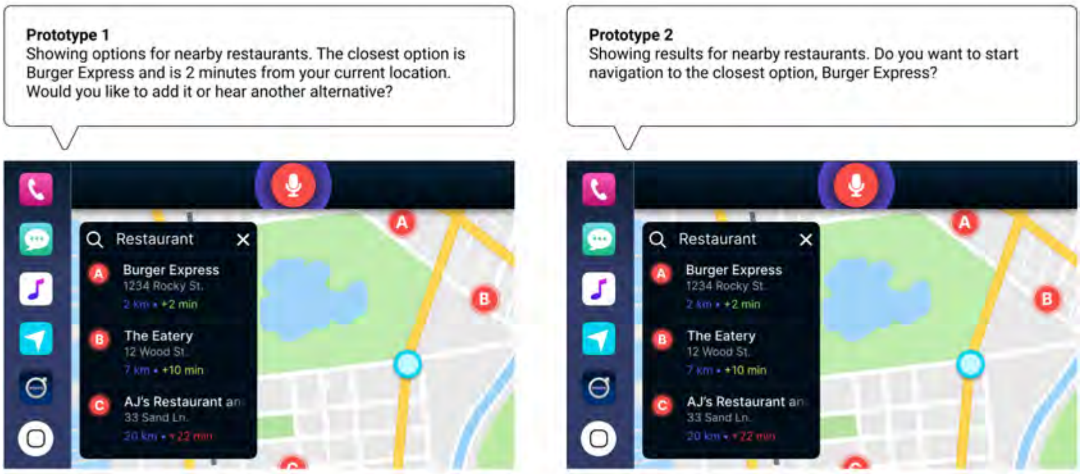
Prototype 2
Your message to Anita Davis says, "Hello, this is a test message."

Do you want to send it?

Figure 6.10: Prototype 1, left, and Prototype 2, right, and their differences when sending a text message

Prototype 2 focused on facilitating shorter voice interactions. This was done by using sparse voice output where only the most relevant content was presented and guiding prompts were not used for content selection. Prototype 2 also had additional visual components, such as suggestion chips, discrete options presented at a relevant turn in a dialog, to help guide users. The suggestion chips were intended to improve discoverability and ease the cognitive demand of formulate accurate utterances. Suggestion chips on Prototype 2 can be seen in Figure 6.10.

Table 6.1 below lists the differences and similarities of the two prototype systems.



Prototype 1
Showing options for nearby restaurants. The closest option is Burger Express and is 2 minutes from your current location. Would you like to add it or hear another alternative?

Prototype 2
Showing results for nearby restaurants. Do you want to start navigation to the closest option, Burger Express?

Figure 6.11: Prototype 1 and Prototype 2 with their differences in voice interaction for showing results in a list

Table 6.1: Prototype similarities and differences

Prototype 1	**Prototype 2**
Presenting Lists Presents option and asks if user want to hear other option.	**Presenting Lists** Presents option, shows other options on screen.
Interjections Interjections not possible.	**Interjections** Users can stop and/or change the VA's actions as they are being executed.
Errors Follows up errors with guiding questions if possible.	**Errors** Follows up errors with guiding questions if possible.
Reference Items on Screen Users are able to reference items/elements on screen.	**Reference Items on Screen** Users are able to reference items/elements on screen.
Remembering Errors Does not repeat error messages the user has previously heard.	**Remembering Errors** Does not repeat error messages the user has previously heard.
Connection Problem Asks user if VA should automatically retry command when connection is better and asks once again once the connection is better.	**Connection Problem** Automatically retries command when the connection is better.
VA Status Looping sound when processing.	**VA Status** No sound when processing.

6.3.2 Qualitative Results

Qualitative observations and interview notes were combined and analyzed using an affinity diagram. The major insights from the qualitative data are:

- Users feel more in control when the VA presented options by voice rather than visual output

- Auto-starting navigation after reading out only one possible route generally caused confusion

- Auto-selecting options for users was met with resistance by participants, citing a loss of freedom and control

- Neither prototypes' solution for communicating VA status was immediately perceptible by participants

- Interjects were a powerful feature, but only for cancelling commands

- VAs should reflect a driver's utterance where concise, direct utterance should trigger the intent as quickly as possible

- VAs must communicate the distinguishing characteristic of an option from many

- Drivers want to know the cause of an error so they can adjust their utterances accordingly

- Partial recognition of an utterance should lead to additional prompts to help trigger the right intent

- Further research and development is required for handling bad connection errors since neither prototype was robust enough

- Participants prefer that the VA require a confirmation before retrying an action after connection issues

- VA processing sound in Prototype 1 was confusing

- Half of the participants found a benefit in the visual suggestion chips

- Participants sometimes gave unexpected affirmative answers to prompts for a selection of one item, e.g. answering "Yes" to "Do you want to send it or change it?"

- Discoverability of features by voice was poor

The above mentioned results are only the main findings. A full list of all findings can be found in Appendix L.

6.3.3 Quantitative Results

Eye glance data from 9 participants were analyzed. The prototype evaluation data, like the summative evaluation data, indicate that off-road glances are shorter and less varied while on-road glances are longer with greater variation in duration while carrying out secondary tasks. In this evaluation, on-road glances are glances directed to the simulator screen, which also includes the DIM. Off-road glances in this evaluation are directed toward the IVI. The mean on-road glance duration was 9.63 seconds with a median of 5.41 s, and a standard deviation of 10.88 s. Off-road glances had a mean duration of 0.65 s, a median of 0.57 s, and a standard deviation of 0.31 s.

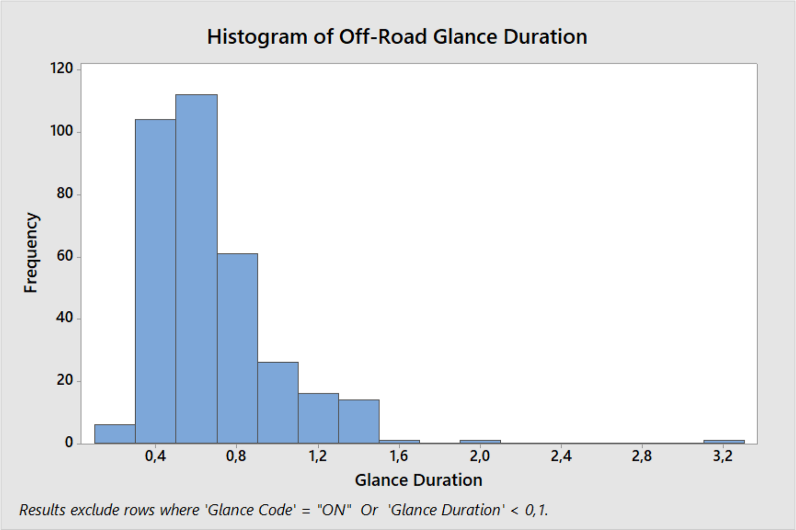


Figure 6.12: Frequency of off-road glance duration times

Figure 6.12 shows the distribution of off-road glance duration times. The peak duration time is between 0.3 to 0.7s. This is in line with off-road glance times for non-manual interaction reported by Taoka [43] and the results of the summative evaluation.

The number of glances varies over the different tasks. The count of off-road glances for the different tasks is shown in Figure 6.13. Tasks 7 through 9 are the navigation tasks and are correlated with the great amount of off-road glances. Task 5, a music task, has a large amount of off-road glances, contradicting the results from the findings in the summative evaluation for single-shot commands. However, this may be due to the scripted error in the prototype evaluation which was distributed evenly rather than on the most difficult as on the actual VAs in the summative evaluation.

The correlation between error and increased number of off-road glances is not as

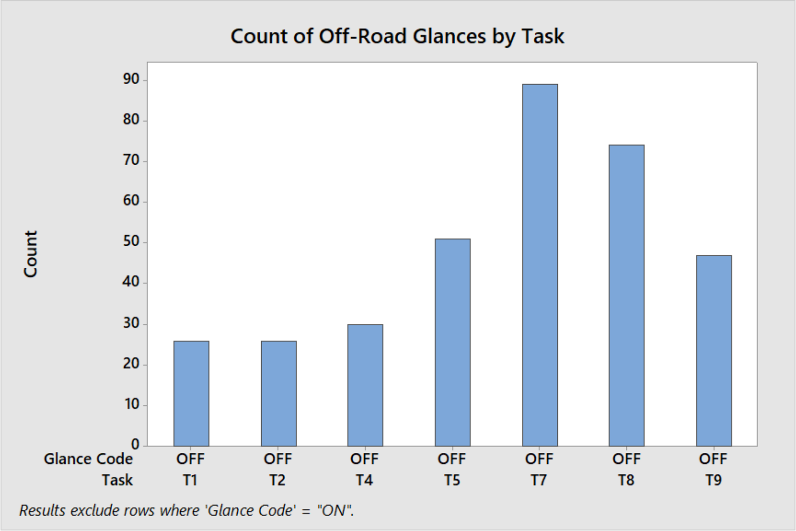


Figure 6.13: Count of off-road glances by task

prevalent as in the summative evaluation. This can be seen in the Figure 6.14. Tasks with the largest proportion of errors have the most off-road glances. Despite Task 5 being a relative easy task, it has the largest proportion of of errors, which explains its high glance count in the previous figure.

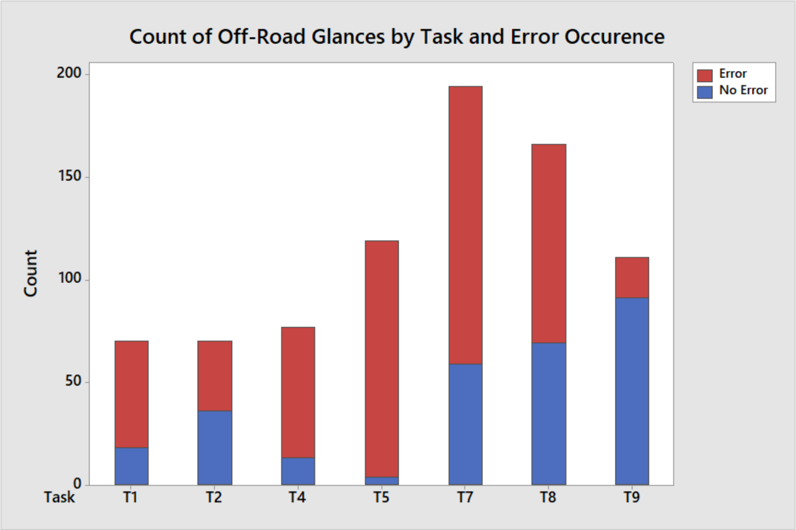


Figure 6.14: Count of off-road glances during tasks with error indications

The glance data does not suggest that either of the tested prototypes is better than the other in terms of off-road glances by task. This is shown in Figure 6.15. The largest differences in glance count exists in Tasks 1 through 4. However, there is no qualitative findings from the prototype evaluation that correlate with this data.

Data from the DALIs were analyzed through the use of box plots. Findings from these plots are checked against the qualitative findings for a relationship or explanation. The first plot of the DALI data in Figure 6.16 show that Prototype 2 had

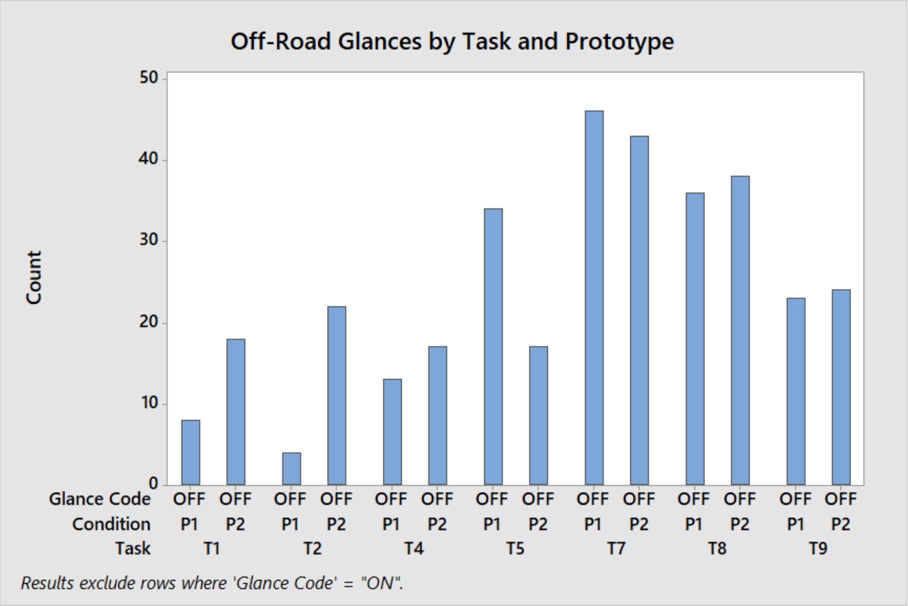


Figure 6.15: Count of off-road glances by task and prototype.

a slightly higher and more concentrated rating for cognitive demand. This correspond with data from the brief interviews where test participants expressed that the conversational approach of Prototype 1 made them feel more in control. This could possibly be linked to participants experiencing a lower cognitive load, but further study is required to confirm these results.

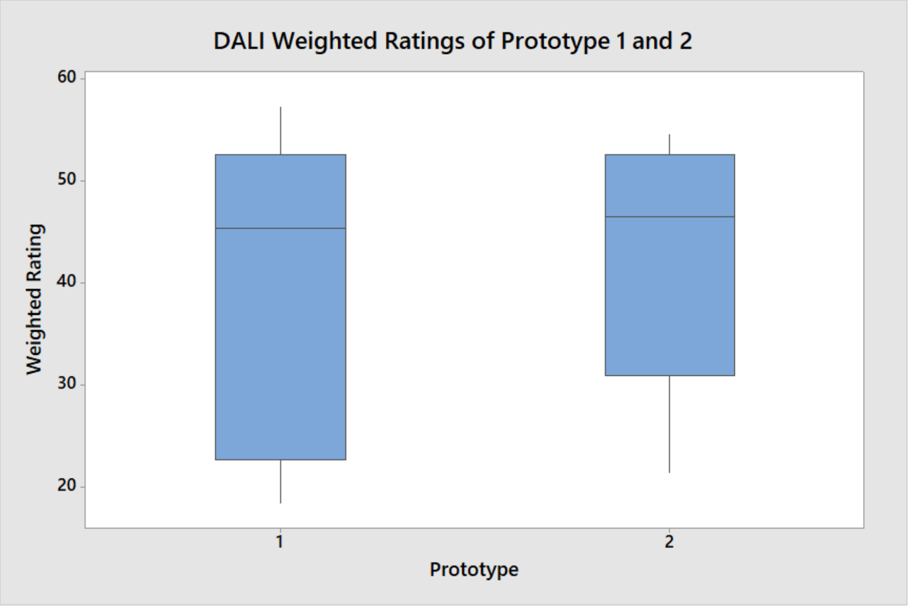


Figure 6.16: Weighted DALI ratings for Prototype 1 and Prototype 2

The adjusted ratings for the individual dimensions of the DALI can be seen in Figure 6.17. The results for the two prototypes are quite similar and there are no large differences. Visual demand is slightly higher for Prototype 2, which was expected due to the suggestion chips. Alternatively, it could be that the conversational nature of Prototype 1 required less visual demand from participants.

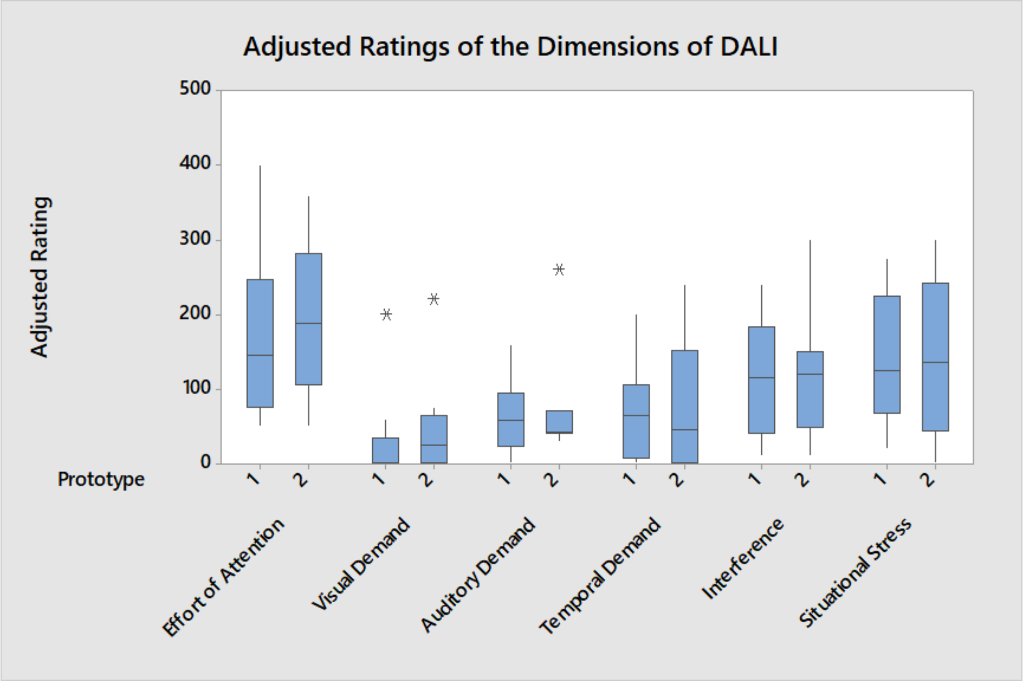


Figure 6.17: Adjusted rating of the dimensions of the DALI

Figure 6.17 also shows that auditory demand is higher for Prototype 1. This is also expected, the prototype takes a more conversational approach and has a higher voice output. Overall, most test participants preferred Prototype 1 because the additional prompts made them feel like they were more in control and making informed decisions. However, Prototype 1 also rated slightly higher in interference, which could be linked to the increased voice communication necessary to complete the secondary tasks.

The SUS data was analyzed by a calculation of mean and standard deviation. This data was compared with related research. The data analyzed comprised 20 individual ratings, two from each test participant for each prototype. Table 6.2 shows all of the scored together with a mean value and standard deviation.

Table 6.2: SUS Results

User	Prototype 1 SUS Score	Prototype 2 SUS Score
1	95	97,5
2	72,5	82,5
3	85	87,5
4	80	77,5
5	82,5	40
6	80	87,5
7	95	90
8	65	72,5
9	90	77,5
10	65	47,5
Mean	**81**	**76**
SD	**10,38**	**17,58**

According to Banglor et al. [5], the marginal range for an acceptable SUS rating lies between 50 and 70, as shown in Figure 6.18. Two scores for Prototype 1 lie in the highly marginal acceptable range with scores of 65 each. Prototype 2 had two scores of under 50, indicated unacceptable levels of usability. All other scores were within the acceptable range.

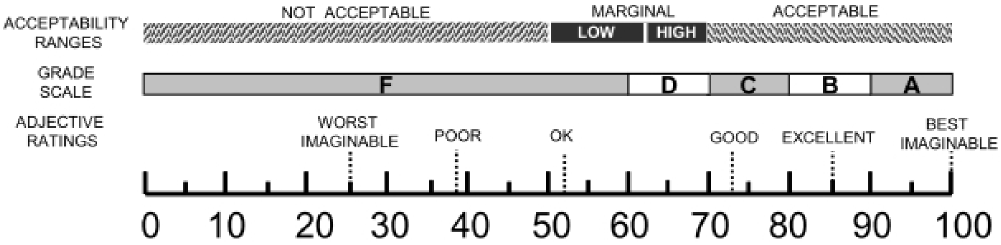


Figure 6.18: SUS score comparison with adjective ratings and acceptability ranges [5]

The lower results of Prototype 2 reflect test participants' views from interviews that Prototype 1 was preferred.

6.4 New Guidelines

A set of new and improved guidelines have been developed specifically for designing in-vehicle voice assistant interaction. These new guidelines are based on the problems and insights identified from the summative and prototype evaluations.

However, before discussing new or altered guidelines, it is important to note that many of the voice interaction guidelines were deemed sufficient, but were found to be often broken by both of the VAs examined by this project. These broken guidelines are:

- Assume users will reference anything presented onscreen by voice [1, 14]
- Provide signposts and suggestions to help users remember the different options to complete an intent or action [1, 14]
- Enable drivers to interrupt an interaction sequence, control the pace of the interaction, and resume an interaction at a logical point in time [30]
- Vary responses to make the dialog sound more natural, especially in repetitive tasks [1, 14]
- Anticipate and provide information and suggestions that are most important and contextually relevant to the user [1, 14]
- Avoid breaking trust with users by including offensive content, unsolicited content like advertisements, unmet expectations, or exposed technical or feature limitations [1]
- Personalize greetings or services with available data, such as the user's current location [1, 14]
- Announce items in a list to make it clear how much information is available [1]
- Give users a way to have the voice assistant read out more options [1]
- Handle errors by minimizing attention to an error and provide a way to get the dialog back on track [1, 14]

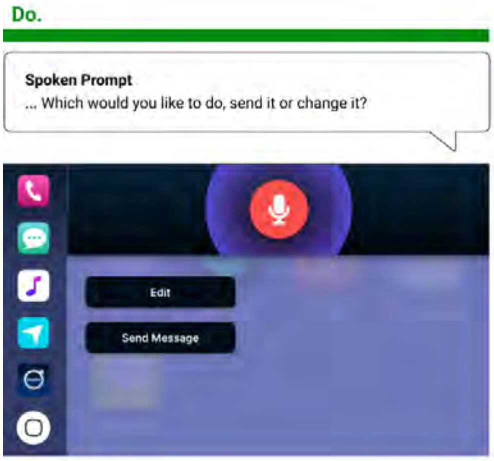
The new guidelines may also be described in terms of the same themes of the existing guidelines: designing car apps, voice and manual input, general voice responses, situation awareness, presenting choice, error handling, discoverability, display, and notifications. Major differences between the new guidelines and the existing guidelines with respect to those themes are described in the following subsections. There were no changes to the guidelines for the theme *designing car apps*.

6.4.1 Voice and Manual Input

These guidelines deal with how drivers are able to provide inputs to the VA while driving, either by voice or manual touch.

- Build multi-turn dialogs for beginners and one-shot commands for experts. Empower drivers to directly access what they want and reduce the amount of time to complete a task.

- Supplement spoken prompts with visual components such as suggestions, alternative actions, or non-critical information that may aid drivers in content selection.

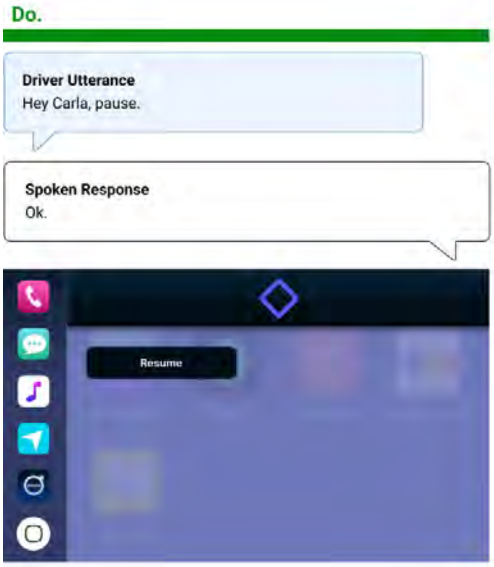


Do.

Spoken Prompt
... Which would you like to do, send it or change it?

E.g. Suggestion chips used above to serve as signposts and support the voice interaction for sending a text message.

- Enable drivers to interrupt an interaction sequence by both voice and manual input. Allow drivers to later resume the interaction as a later, logical point in time or return to a previous state of interaction if little effort is required to start the sequence again.



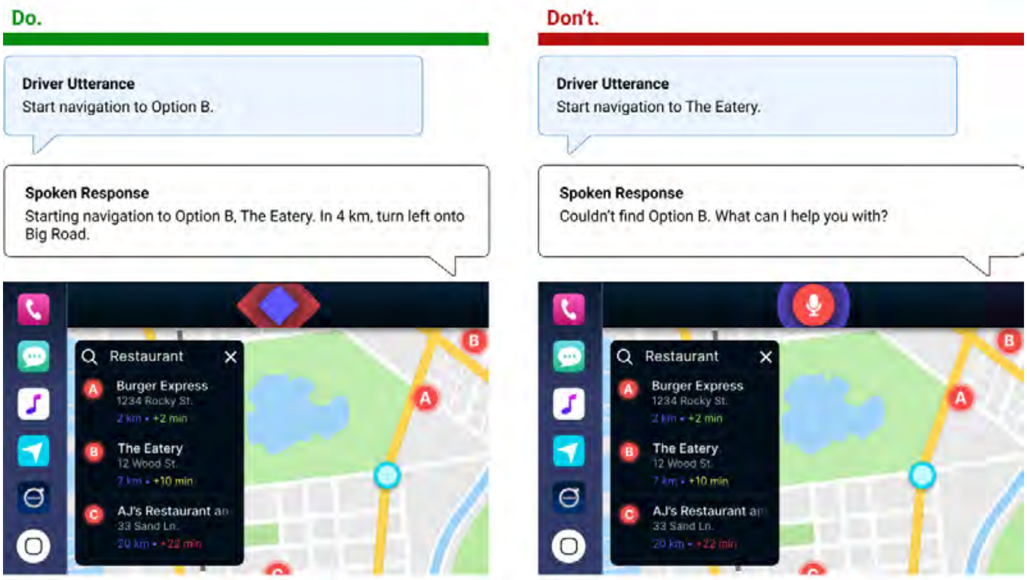
Do.

Driver Utterance
Hey Carla, pause.

Spoken Response
Ok.

E.g. Interrupted interaction via the "Pause" command. Alternatively, drivers *may cancel an interaction altogether.*

- Assume drivers will reference anything presented on the screen by voice. Allow drivers to reference on-screen items by both title, superlative, or generic reference.



Do.

Driver Utterance
Start navigation to Option B.

Spoken Response
Starting navigation to Option B, The Eatery. In 4 km, turn left onto Big Road.

Don't.

Driver Utterance
Start navigation to The Eatery.

Spoken Response
Couldn't find Option B. What can I help you with?

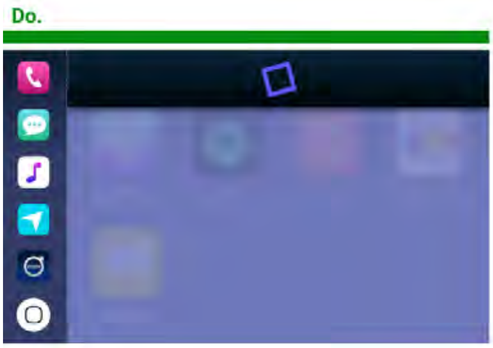
E.g. Driver can reference the items in a navigation list by name, label, or generic term.

- Allow drivers to trigger an action or intent by both manual touch and voice commands. This includes designing for multiple utterances for the same action or intent. Drivers can say "Start navigation" and "Take me to McDonald's", both of which start the intent for getting driving directions, the former which will require an additional turn in the dialog.

6.4.2 General Voice Responses

Existing guidelines for general voice responses were found to be overall sufficient when applied to the driving context, so only one new guideline is presented here.

- Provide responses quickly to minimize interaction between the driver and voice assistant. If the response time exceeds 2.00 seconds, the voice assistant to provide a clearly perceptible indication that the voice assistant is in the process of responding. The threshold of 2.00 seconds is based on NHTSA guidelines for traditional HMI input [30].

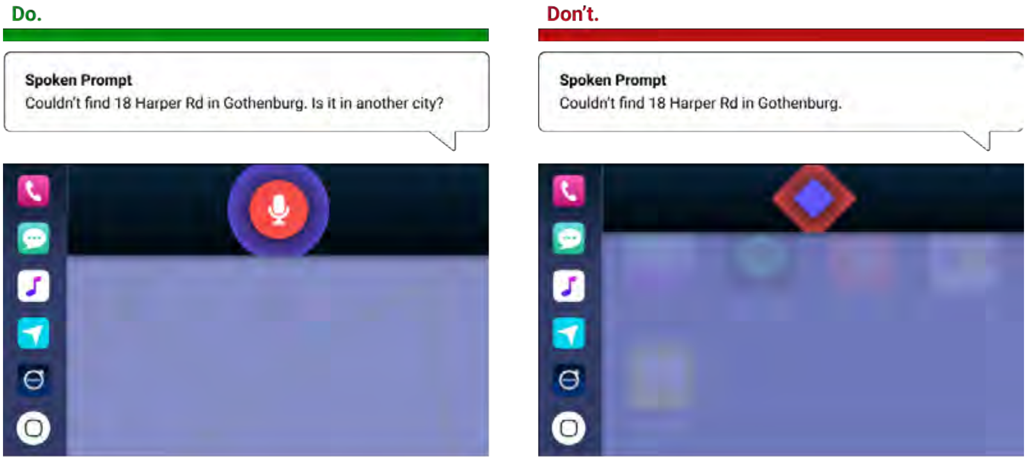


E.g. Loading symbol (rotating square) used to communicate the VA is processing. Sound may also be used in combination with visual elements.

6.4.3 Error Handling

These guidelines deal with handling errors and reiterate existing guidelines for voice assistants and how important these particular guidelines are with respect to driving.

- Prevent errors whenever possible to avoid increased driver attention to the screen. Provide suggestions to alternatives or use partial matches to driver utterances provide contextually relevant prompts.

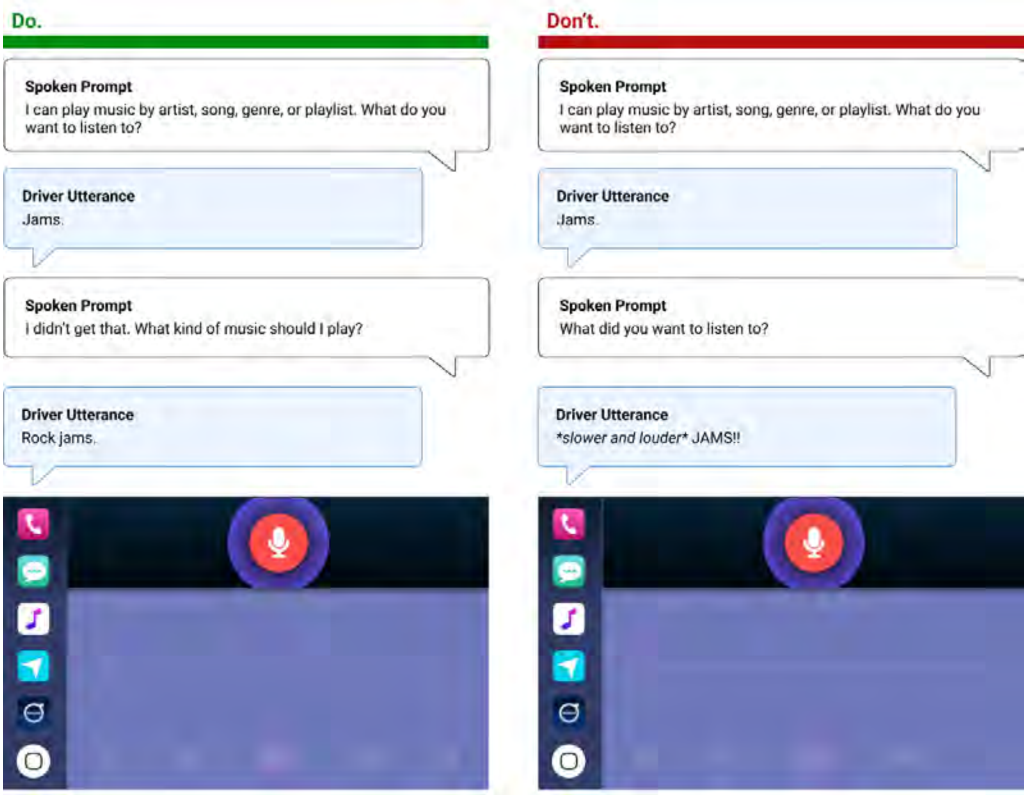


Spoken Prompt
Couldn't find 18 Harper Rd in Gothenburg. Is it in another city?

Spoken Prompt
Couldn't find 18 Harper Rd in Gothenburg.

E.g. VA uses an additional location prompt to help driver complete the desired action.

- Re-prompt drivers with a slight variation on the original prompt to provide additional clues for what kinds of inputs are appropriate to trigger the correct action or intent. When drivers don't understand what went wrong, they may repeat the same utterance slower and more clearly only to get stuck in the same error loop.



Spoken Prompt
I can play music by artist, song, genre, or playlist. What do you want to listen to?

Driver Utterance
Jams.

Spoken Prompt
I didn't get that. What kind of music should I play?

Driver Utterance
Rock jams.

Spoken Prompt
I can play music by artist, song, genre, or playlist. What do you want to listen to?

Driver Utterance
Jams.

Spoken Prompt
What did you want to listen to?

Driver Utterance
slower and louder JAMS!!

E.g. Vary a prompt for music input to help the driver better understand what they should say.

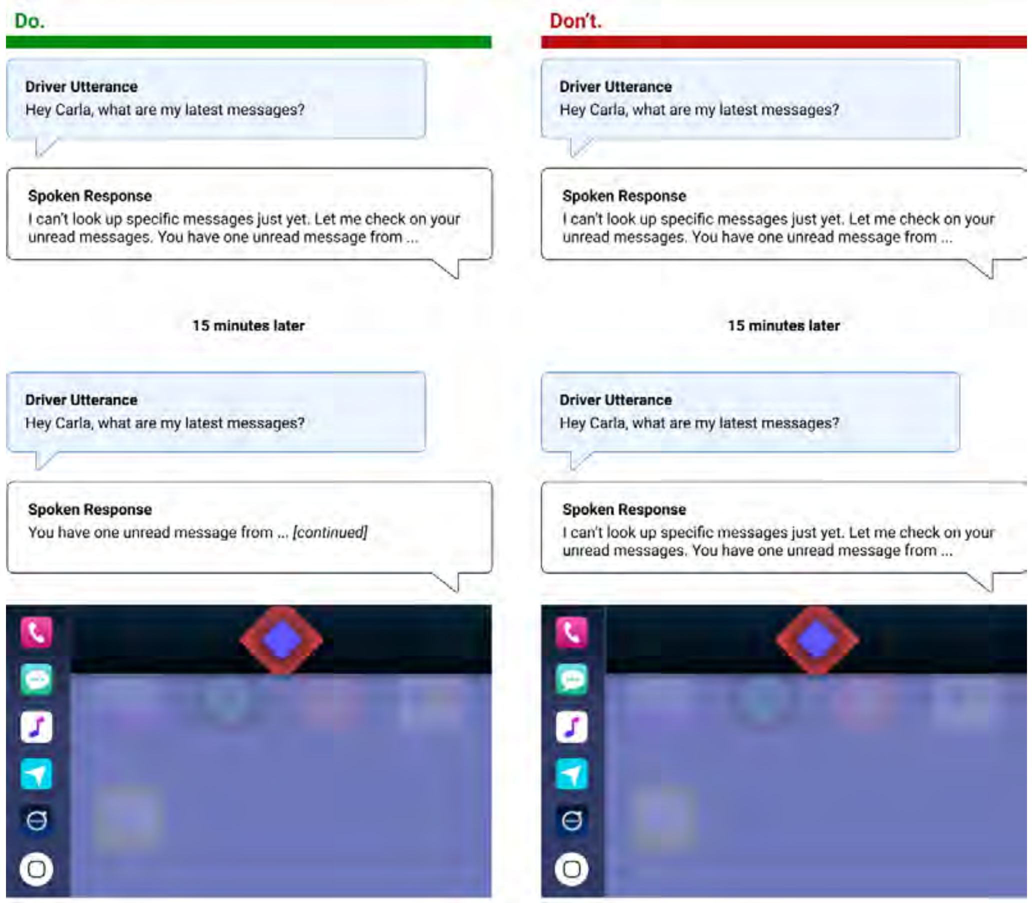
- Respond gracefully when data is unavailable and make the data connection status clear so drivers can know when they can attempt the action or intent again.

6.4.4 Situation Awareness

Guidelines for situation awareness deal with how the VA can build a context for the driver's current situation, as to prevent the driver from reaching their cognitive capacity.

- Keep track of the context of the dialog between the voice assistant and the driver to understand the use of pronouns and generic references and avoid repeating prompts or responses that may frustrate and distract the driver.

- Adapt to a driver's vocabulary for utterances and inputs. For example, a driver may have a preference for using the phrase “latest messages” to refer to “unread messages.”



Do.

Driver Utterance
Hey Carla, what are my latest messages?

Spoken Response
I can't look up specific messages just yet. Let me check on your unread messages. You have one unread message from ...

15 minutes later

Driver Utterance
Hey Carla, what are my latest messages?

Spoken Response
You have one unread message from ... *[continued]*

Don't.

Driver Utterance
Hey Carla, what are my latest messages?

Spoken Response
I can't look up specific messages just yet. Let me check on your unread messages. You have one unread message from ...

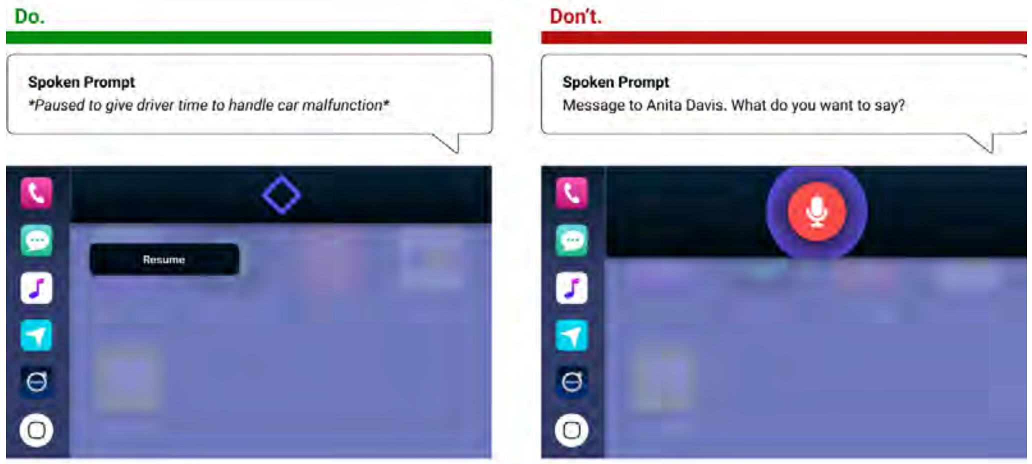
15 minutes later

Driver Utterance
Hey Carla, what are my latest messages?

Spoken Response
I can't look up specific messages just yet. Let me check on your unread messages. You have one unread message from ...

E.g. VA keeps track that an error has already been presented and adapts to the driver's personal vocabulary.

- Use the car's current context to avoid adding stress to a driver's situation. For example, if the driver encounters a car malfunction mid-interaction, do not create added stress to the situation by prompting or re-prompting the driver to complete the interaction sequence.



Do.

Spoken Prompt
Paused to give driver time to handle car malfunction

Don't.

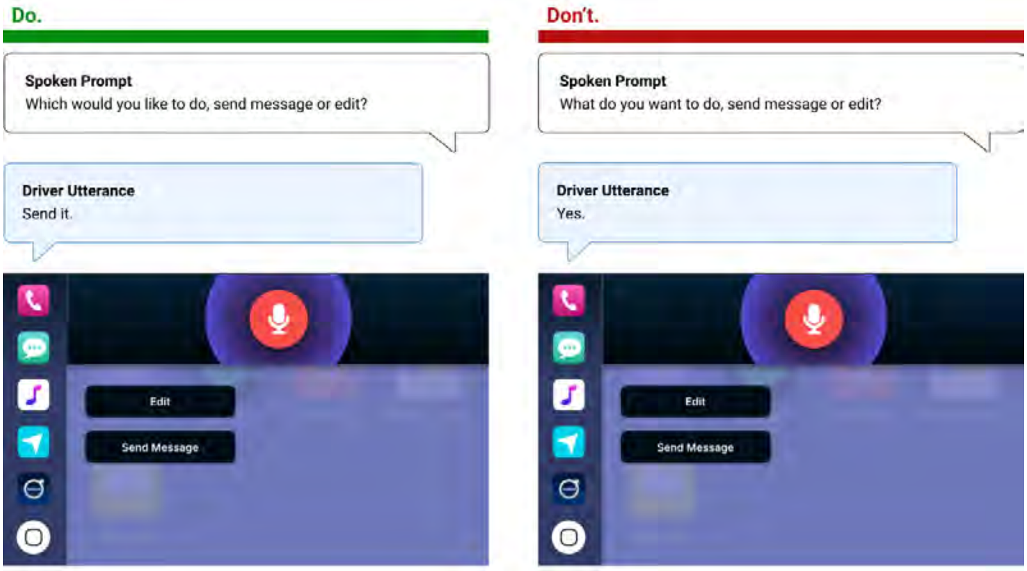
Spoken Prompt
Message to Anita Davis. What do you want to say?

E.g. VA uses information from the car to pause an interaction during a car malfunction.

6.4.5 Presenting Choice

These guidelines concern how VAs respond to drivers and present choice architectures.

- Present a clear, simple set of options for the driver to choose from. Avoid using open-ended questions for prompts which can confuse drivers or cause them to answer in unexpected ways.



Do.

Spoken Prompt
Which would you like to do, send message or edit?

Driver Utterance
Send it.

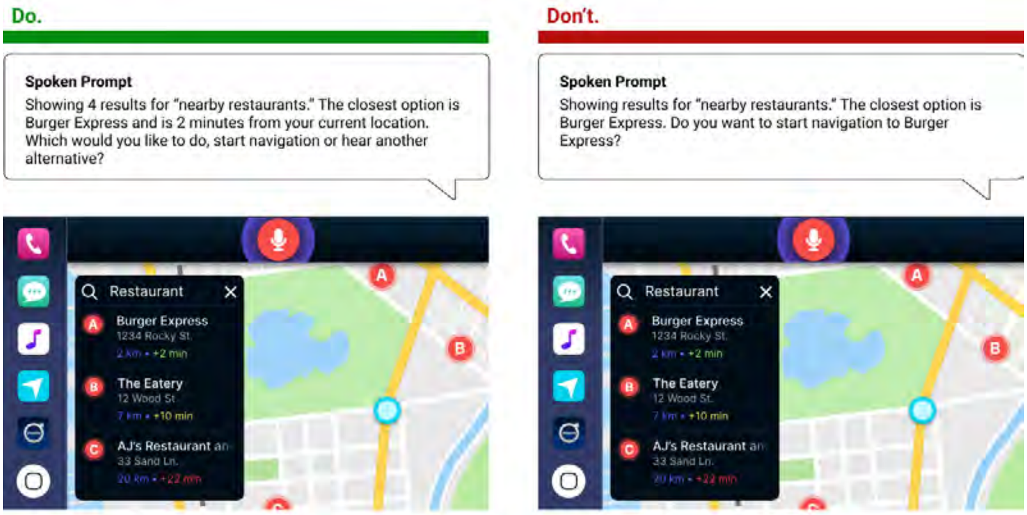
Don't.

Spoken Prompt
What do you want to do, send message or edit?

Driver Utterance
Yes.

E.g. When confirming a message, VA phrases the prompt to make it clear to the driver then they must select one of multiple options.

- Give drivers a brief overview when presenting a list, such as by noting how many items are in the list.
- Provide drivers with contextually relevant and differentiating information about items in a list to aid drivers in content selection without relying on a screen.



Do.

Spoken Prompt
Showing 4 results for "nearby restaurants." The closest option is Burger Express and is 2 minutes from your current location. Which would you like to do, start navigation or hear another alternative?

Don't.

Spoken Prompt
Showing results for "nearby restaurants." The closest option is Burger Express. Do you want to start navigation to Burger Express?

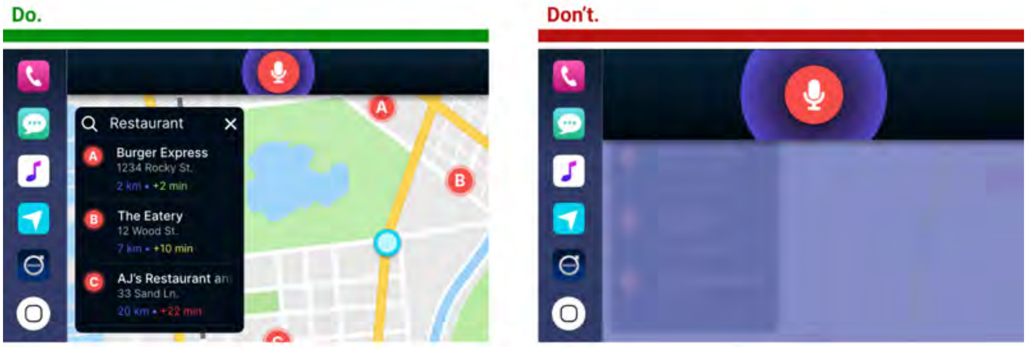
E.g. Providing information relevant to navigation, such as time from current location to destination.

- Avoid auto-selecting an option for the driver, unless done through a setting previous set by the driver. For example, drivers may have set a preference to always start navigation using the most convenient route, instead of having to choose form multiple options.

6.4.6 Display

These guidelines deal specifically with the IVI display and how it can support VA interaction and addresses issues uncovered in the summative evaluation.

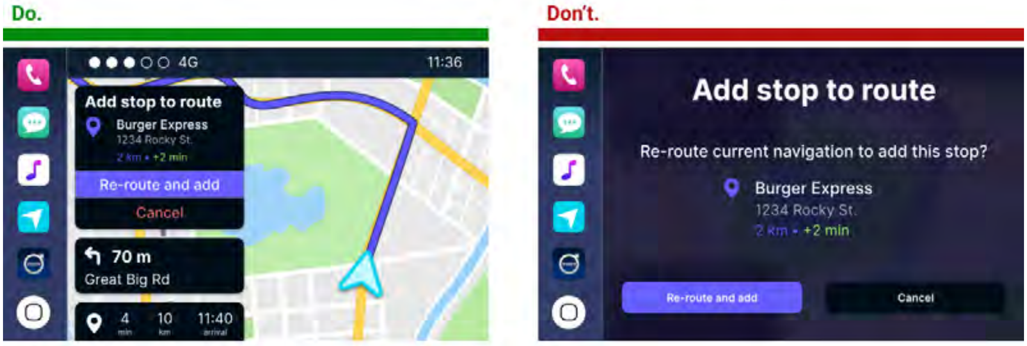
- Allow any selectable content on the screen to be visible to the driver even when the voice assistant is activated, so the driver can reference any onscreen content while giving voice input.



e.g. Driver is able to see and reference location names if necessary while driving, avoiding forced recall of an obscured label or title.

- Use contextually relevant suggestion chips to guide drivers to different task paths and provide a visual fallback in case the driver missed the accompanying voice output. Allow drivers to hide suggestion chips either as an intent or as a personal setting.

- Avoid using fullscreen alerts to display information to the driver, unless as part of an interrupted sequence initiated by the driver. Fullscreen alerts often obscure contextual visual information related to the alert.

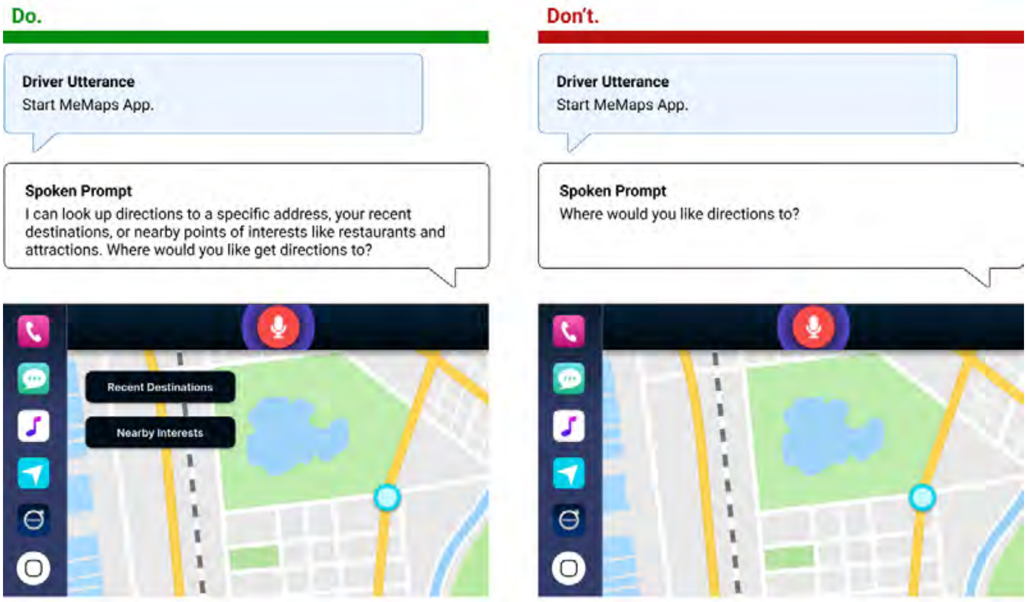


E.g. For manual interaction, allow driver to see context navigation decision rather than obscure that map with a fullscreen dialog.

6.4.7 Discoverability

A challenge with voice assistants is conveying useful features to the user by voice. These guidelines deal with improving discoverability of VA features by the driver.

- Provide hints or suggestions for difficult to discover or open-ended tasks. For example, playing music is a more discrete task as most drivers have previous experience with music plays where music selection is usually done by song title, artist, playlist, or music genre. In contrast, getting directions is more open-ended and varied.



Do.

Driver Utterance
Start MeMaps App.

Spoken Prompt
I can look up directions to a specific address, your recent destinations, or nearby points of interests like restaurants and attractions. Where would you like get directions to?

Don't.

Driver Utterance
Start MeMaps App.

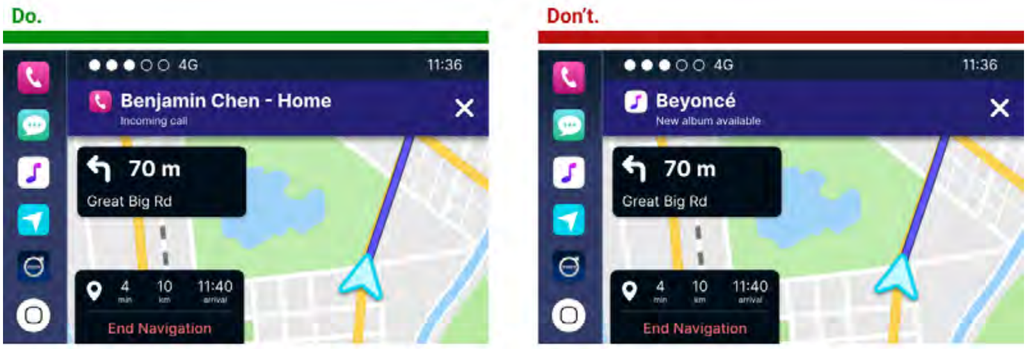
Spoken Prompt
Where would you like directions to?

E.g. VA suggests map features the driver may not have used yet and may find helpful at the start of an interaction.

6.4.8 Notifications

Guidelines for using push notifications.

- Reserve notifications with sound for information or tasks that require the driver's immediate attention. Notifications accompanied with sound draws the driver's visual attention to the screen.



E.g. Incoming call requires the driver's immediate attention, not news of a new music album.

7
Discussion

In this chapter the project findings, the process and the possibilities for future research will be discussed. The findings will be discussed in terms of validity and how they should be interpreted. The process discussion will focus on how the different methods was applied and how this influenced the outcomes of the methods and if it could have been done in other ways. In the last section of the discussion possible interesting areas for future related research are proposed and motivated.

7.1 Findings

The findings of this project are many and varied. In relation to the research questions some of the findings are more closely tied to either decreasing diverted attention or cognitive load.

The main contributor to increased diverted attention was errors. The VA failing to correctly understand the user ultimately resulted in longer interactions to complete the tasks, drawing more attention from the road. Working to minimize the amount of errors and to design error messages so that the impact of them is as small as possible is central to improving the caused distraction by VAs. This finding is somewhat similar to the findings of Mehler et al. where VA speech recognition errors were found to increase the driver's workload and frustration [28].

In relation to the cognitive workload, one of the main findings was that a more conversational approach enabled users to converse more naturally with the VA decreasing their cognitive load. A more voice focused VA would logically decrease the visual demand of a system but at the same time, it is probable that it would increase the auditory demand. According to Wickens' Attention Model, the cognitive processes of processing spatial and verbal information are separated, indicating an easier time dealing with inputs of these types simultaneously [48]. Voice interaction can be purely verbal, while screen interaction, which includes both text as well as graphical elements demands both verbal and spatial cognition. Driving is mostly spatial, at least when processing the on-road information, which would mean that voice interaction while driving has a greater separation of the types of cognitive processes which according to Wickens, would lead to a lower levels of effort for the user [48]. Another finding related to cognitive load, was that the user had very varied amounts of attention resources depending on the driving situations. During more intensive driving, like e.g. entering a roundabout, the driver often could not interact with the VA at all. This could possibly be explained by Kahneman's theory of intensity and selectivity of attention. A more demanding driving situation demands more intensive attention, drawing away from the driver's available attention resources [21]. A VA that continuously talks in these situations is bound to cause interference and situational stress, increasing the overall cognitive load.

Cognitive load may also be tied to how natural the conversation is between the driver and the VA. This This is one possible explanation for the results from the prototype evaluation which showed that in debrief interviews, participants found Prototype 1 to be easier to use, with its more conversational approach. Moreover, participants from the summative evaluation reported Siri as being easier to use, partially due to its more conversational responses compared to Google Assistant. Thus, VAs should leverage the structures of human-human conversation such as those described by Grice's Maxims [15].

The findings from the summative evaluation outlines many problems with currently available VAs when they are used in real driving scenarios. The qualitative results focuses on driver's behaviours and the driver's own experiences from using the VAs on roads. This data is limited in that it is focused on drivers driving alone in their car, which excludes many common situations when there are other passengers in the car. Due to the nature of the open roads used for the summative evaluations, the results may be skewed towards country road driving, missing to cover aspects of e.g. city and highway driving. Moreover, qualitative data from user interviews in both evaluations revealed the drivers are open to using VAs when they are alone in the car, but less so if they were travelling with a passenger. In fact, several test participants said they would have the passenger take over the role of the VA.

Results from the summative evaluation also suggested that PA, but more generally SAE Level 2 ADS, increases the number of off-road glances. However, these results may be influenced by the country roads used in the summative evaluation. The use of a Level 2 ADS is optimized for driving contexts such as highway driving and relies on clearly visible lane side marker lines on both sides of the current traffic lane. If suitable driving conditions are not met then Level 2 ADS my increase visual distraction rather than aid drivers to complete secondary tasks more safely. Additional research with respect to Level 2 ADS and VAs is needed to draw a definitive conclusion. These future evaluations of Level 2 ADS and its effect on secondary tasks should be conducted in driving contexts most suitable for Level 2 ADS, such as highway driving. Additionally, the results from the summative evaluation may be influenced by the test participants and the fact that only 2 had previous experience with PA, the Level 2 ADS used in the test. The effect of a Level 2 ADS on visual distraction and cognitive load for drivers experienced in using a

Level 2 ADS may be less and is an opportunity for future studies to examine.

The results may be affected by the common backgrounds of the users of the two user groups. The summative evaluation data is limited by having users who are more interested and invested in car development and generally in cars than an average person. The data for the prototype evaluation is limited by having users with more academic experiences than the average person, possibly skewing the data. Additionally, the results may be affected by the fact that the evaluations were held in English. Though not all participants were native English speakers, all used English as part of their daily work.

The quantitative data is limited by the small sample size of the tests. The data can be used together with the qualitative data in this report in order to highlight and indicate themes and patterns. Using the quantitative data on its own will make identification of patterns and themes and the analysis of the reason behind the findings more difficult.

7.2 Process

In this section the overall project process is discussed. This section is divided into subsections discussing the methods chosen and the execution of the tests.

7.2.1 Methods

In order to measure the cognitive load of the test participants the DALI method was chosen. It was chosen since it is a method specifically developed for measuring car driver's cognitive load [35]. Since the summative evaluations took place on a open road, the original plan, to fill out a DALI for each task, did not work out. Stopping the car at a safe place and filling out a DALI after each task would result in each user filling out a total of 36 DALIs which was deemed to by far too time consuming and tiring for the test participant. To tackle this problem, the use of the DALI differs from the standardized way of using it. The DALI was instead used to evaluate a group of tasks instead of the separate tasks. This adds a level of complexity for the test participants as they must mentally average their impression of the whole set of tasks when rating and weighting the dimensions of the DALI.

Other methods for cognitive load measuring could have been used, but all of them would have had to be adjusted to measure a whole set of tasks instead of just one, just like the DALI was used. The RSME and SWAT methods would have resulted in much less detailed data as the measure points are fewer, leaving less data to analyze and use for conclusions. The NASA-TLX would have resulted in the same amount of data. Some dimensions, e.g. physical demand, would however be irrelevant to the task of driving.

For assessing the test participant's subjective experience of the usability of the two prototypes, the SUS method was used. The more thorough and specifically speech interface focused SASSI method could have been used instead of SUS. SASSI would have provided more detailed information about their experience due to being a longer questionnaire, with its 32 items in comparison to SUS's 10 items. The SUS was chosen primarily due to its shorter length, allowing more time during the test to be spent on driving and interviewing. Qualitative results from observations and interviews were prioritized and due to experiencing difficulties getting test participants for the summative evaluation, extra effort to make the test shorter was made when planning the prototype testing.

The manual eye tracking method was chosen above automatic solutions of eye tracking. This choice was based upon internal consultation from experts within Volvo Cars. Automatic eye tracking solutions often has problems with the generated data which leads to a need to manually observe the data to make sure it is usable. It also varies in how well it can track the eyes of the subject depending on the subject's physical appearance as well as the light conditions. Based on this, manual eye tracking was deemed the most suitable method as it is quite reliable in regards to the above mentioned condition, all of which would not be possible to control for the test. Automated eye tracking methods and instruments could have possibly lead to more exact times for each eye glance. However, the glance data in this project was primarily used to support and contextualize findings from the collected qualitative data.

Using a WOz prototype for the VA in the prototype evaluation has a risk of limiting our findings. Since all possible interaction patterns had to be pre-made and the decisions of what information the VA is able to understand is decided by a person, the prototype run a risk of being able to understand the user too well, as several test participants did not realize the prototype was being controlled by a wizard. However, all test participants experienced errors during the evaluation to minimize this risk.

7.2.2 Tests

The relatively small test sample of both the summative evaluations and the prototype tests makes thorough analysis and interpretation of the numerical data more difficult. The qualitative findings have instead been the primary data utilized throughout the project. The quantitative data, in form of DALI results, glance data statistics, and SUS results have been used as secondary data, complementary to the qualitative findings.

Due to restrictions related to the car used for the summative evaluation, only Volvo employees were allowed to drive it. This vastly limited the possibilities for recruitment of test participants, resulting in a user group which has more experience and interest in cars than the average person. The average time for how long the test participants had had their driver's licenses was 12.6 years, more users with longer driving experience would have been desired to make for a more varied user group. For the prototype test the user group included mostly students with even shorter periods since they took their licenses. Students are likely to have more experience with VAs compared to the participants of the summative evaluation, further making it more difficult to determine of the test participant groups may have affected the result.

Conducting tests while driving on open roads makes for large amount of uncontrollable factors. Factors like weather and traffic has the potential to affect the driving experience drastically. The uncontrollable factors of open road driving have the potential to skew the results in unpredictable ways. These uncontrollable factors are at the same time present when VAs are used in real usage scenarios making testing with them desirable. Since the sample size was small, the numerical data from the open road tests probably varied more than it would have in a controlled environment. The qualitative data from the open road test has the possibility to enable new unforeseen findings and patterns to be discovered since many factors and traffic situations could have been missed in a totally controlled environment. In this aspect the open road testing allowed for more explorative research.

Changing from doing testing in a real car on open roads in the summative test phase to doing tests in a simulator during prototype testing makes it impossible to properly compare data in between the two tests. The conditions are so very different that it would be impossible to determine how the results relate to each other. The summative evaluation results shows the problems of currently existing VAs and how they relate to current VA guidelines. The prototype evaluation results in possible solutions to the problems discovered in the summative phase. To be able to prove that these solutions truly improves the driver's cognitive load and diverted attention, more testing would have to be conducted.

The WOz has been discussed in the method section. Errors, during the WOz tests, were predetermined and put into a schedule in order to mimic the limitations of real VAs but this also made it so that errors occurred in situations were a real VA might no have had any errors. It is difficult to determine how these forced error scenarios might have influenced the results, but since the errors was spread out to minimize any skewing of the results due to repeated errors at the same stage in the test, the negative effects of this way of doing it should be minimized.

The solutions tested during the prototype evaluations are limited to only one iteration of testing. Multiple iterations would have allowed more solutions to be tested which could have further improved the new guidelines coverage of the currently existing problem areas.

Conducting the prototype evaluations in a car simulator might affect and change the test participant's behaviour as the simulator environment is very different from that in a real car. There is never a real sense of danger in a simulator, creating a less serious mood when driving. This lower level stakes could lead to drivers paying less attention to the primary task of driving than they would have done in a real car.

7.3 New Guidelines

The guidelines produced as a result of this thesis work was done in response to the voice assistants that are available and commercially integrated in vehicles at the time of this writing. Thus, these new guidelines may be directly applied to those integrated voice assistants effective immediately. Since these guidelines were a result of evaluation under Level 2 and Level 0 driving conditions, these guidelines primarily seek to decrease visual distraction and cognitive load, since the driver is responsible and in control of the vehicle. Moreover, the guidelines have been written to be task independent, so if the category of apps allowed and available on IVIs expands in the near future, these guidelines will still be relevant. However, these guidelines are not exhaustive, and should ideally be used in combination with other guidelines for voice and in-vehicle interface interaction.

In the future where fully autonomous vehicles (Level 5) may make up the majority of cars on the road, turning today's drivers into passengers. As passengers, focus on decreasing visual distraction and cognitive load while performing secondary tasks may be less of a concern since passengers are not responsible for the safety of a vehicle's driving. However, new primary tasks may replace driving for these future passengers. It is possible that while riding in a Level 5 autonomous vehicle, passengers may shift their primary focus to working, exercising, or many of the other speculated activities. In these scenarios, reduced visual distraction and cognitive load while carrying out secondary tasks remains relevant in a fully autonomous future. However, if the primary tasks of fully autonomous vehicle passengers shifts to be entirely screen-based, then many of the guidelines produced here may become obsolete.

7.4 Future Work

The findings uncovered by this thesis help to address the pressing need for guidelines and understanding of voice assistant interaction in vehicles today. While this research work examined the integrated, holistic experience of voice assistants in vehicles and how voice and manual interaction work today, there is still a wealth of research left to be completed.

The guidelines developed by this work were primarily tested in a simulator setting. Thus, there is future work left to be done to examine how these guidelines may hold up in field tests, especially with added Level 2 ADS features turn on. Continued field tests and iteration of the prototype and guidelines presented here will improve and refine the guidelines. Once field tests are possible, these implemented guidelines can be fairly compared to the existing VAs that were evaluated in this project.

While this project examined the multimodal nature of voice assistant interaction in vehicles, it heavily focused on voice input and output. While visual output was a consideration through this work, it was not a primary focus which leaves room for future work to examine the perception of the visual components used by the existing voice assistants. Sound, as in audio output that are not worded responses or prompts, was also considered but not closely examined by this project. There is room in future work for deeper focus into these modalities outside of voice and how they work with and in influence visual distraction and cognitive load while driving.

8

Conclusion

This project explored vehicle integrated VAs and how the current implementation and designs of these were suited for safe driving. The aspects of cognitive load and diverted attention were focused throughout the project. The purpose of the research focus was to provide designers, car manufacturers and researchers with relevant guidelines to adhere to in order to minimize the negative effects of high levels of cognitive load and diverted attention when implementing and continuing development of car integrated VAs.

Research question 1a stated *"What existing design guidelines and patterns are implemented in voice assistant integrated infotainment systems?"*. The results indicated several similarities and differences in VA guidelines. Beneath is a brief summary:

- Car apps should be designed with voice as the primary mean of interaction. Strip apps of any functions and elements that cause distraction while not serving the apps main purpose.

- Design for varied and flexible use by many different types of users in terms of how they speak and interact with the VA.

- Minimize touch screen interaction and carefully consider what and how to display on screen to compliment whats presented through voice.

- Use error handling to minimize attention to error and guide user back to the right path for their desired intent.

Research question 1b stated *"What improvements to existing voice assistants can be made to minimize diverted attention from the primary task of driving"*. The result indicated that errors during interaction with the VA as the single largest factor increasing the diverted attention. Other design details raging from status indication to distribution of information over available channels were also find to improve the situation.

Research question 1c stated *"What improvements to existing voice assistants can be made to minimize cognitive load while executing a secondary task during the primary* *task of driving?"*. Findings indicate that making the conversation feel natural and the system's ability to adhere to the current driving context were important factors in lowering the cognitive load.

The answers to these sub-questions combined answer the overarching research question of this project: *What adjustments to existing NLP-based voice assistant design guidelines should be made for safer interaction while driving?* The adjustments found necessary are summarized in a new set of practical design guidelines which can be found in Chapter 6.4, New Guidelines. The new guidelines and findings of this report may guide future designers, manufacturers and researchers to creating better, safer and more enjoyable VAs in the exciting future of automotive development to come.

Bibliography

[1] Amazon. Alexa Design Guide. URL: `https://developer.amazon.com/docs/alexa-design/get-started.html` (visited on 05/03/2019).

[2] Apple. CarPlay - Human Interface Guidelines. URL: `https://developer.apple.com/design/human-interface-guidelines/carplay/overview/introduction/` (visited on 02/12/2019).

[3] Apple. SiriKit - Human Interface Guidelines. URL: `https://developer.apple.com/design/human-interface-guidelines/sirikit/overview/introduction/` (visited on 02/15/2019).

[4] K. M. Bach, G. Jaeger, M. B. Skov, and N. G. Thomassen. Interacting with In-Vehicle Systems: Understanding, Measuring, and Evaluating Attention. Technical report, 2009.

[5] A. Bangor, P. Kortum, and J. Miller. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[6] J. Brooke. SUS—a quick and dirty usability scale. 1996. *Usability evaluation in industry*, 189(194):4–7, 1996.

[7] W. Buxton. *Sketching User Experiences: getting the design right and the right design.* Morgan Kaufmann, San Fransisco, CA, 2007.

[8] J. De Winter and P. Happee. Advantages and disadvantages of driving simulators: a discussion. *Proceedings of Measuring Behavior 2012*:47–50, 2012.

[9] European Union. Statement of Principles on human-machine interface. *Official Journal of the European Union*, 2008.

[10] G. Fulton. Voice First, Screen Second: Designing for Voice in Adobe XD, 2018. URL: `https://blog.prototypr.io/voice-first-screen-second-designing-for-voice-in-adobe-xd-1a9e5efdca15` (visited on 02/06/2019).

[11] W. Gaver. What should we expect from research through design? In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, pages 937–946, New York, NY. ACM Press, 2012.

[12] J. Giacomin. What Is Human Centred Design? *The Design Journal*, 17(4):606–623, Dec. 2014.

[13] Google Design. Android Auto. URL: `https://designguidelines.withgoogle.com/android-auto/` (visited on 02/12/2019).

[14] Google Design. Conversation Design. URL: `https://designguidelines.withgoogle.com/conversation/conversation-design/welcome.html` (visited on 02/12/2019).

[15] P. Grice. Logic and Coversation, 1975.

[16] R. A. Harris. *Voice Interaction Design*. Elsevier, San Fransisco, CA, 2005.

[17] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Volume 52, pages 139–183. 1988.

[18] K. S. Hone and R. Graham. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, 6(3&4):287–303, 2002.

[19] IDEO.org. *The Field Guide to Human-Centered Design*. San Fransisco, CA, 1st. ed. Edition, 2015, page 189.

[20] JAMA. Guidelines for in-vehicle display systems. *SAE Technical Paper*, (2):1–15, 2008.

[21] D. Kahneman. *Attention and Effort*. Prentice Hall, New Jersey, NY, 1st. ed. Edition, 1973.

[22] L. Klein. *Design for Voice Interfaces*. O'Reilly Media, Sebastopol, CA, 1st. ed. Edition, 2015.

[23] S. R. Klemmer, A. K. Sinha, J. Chen, J. A. Landay, N. Aboobaker, and A. Wang. Suede: A Wizard of Oz Prototyping Tool for Speech User Interfaces. *Proceedings of the 13th annual ACM symposium on User interface software and technolog*, 2:1–10, 2000.

[24] J. R. Lewis. Standardized Questionnaires for Voice Interaction Design. *Voice Interaction Design*, 1(1):1–16, 2016.

[25] V. E.-w. Lo and P. A. Green. Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature. *International Journal of Vehicular Technology*, 2013:1–13, 2013.

[26] A. o. A. Manufacturers. Statement of Principles , Criteria and Verification Procedures on Driver Interactions with Advanced In- Vehicle Information and Communication Systems Including 2006 Updated Sections Driver Focus-Telematics Working Group, 2006.

[27] B. Martin and M. Hanington. *Universal Methods of Design*. Rockport Publishers, Bevery, MA, 2012.

[28] B. Mehler, D. Kidd, B. Reimer, I. Reagan, J. Dobres, and A. McCartt. Multimodal 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, Mar. 2016.

[29] NHTSA. Distracted Driving. URL: `https://www.nhtsa.gov/risky-driving/distracted-driving` (visited on 02/12/2019).

[30] NHTSA. Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. Technical report 81, NHTSA, 2013, pages 1–54.

[31] NHTSA. Visual-Manual NHTSA Driver Distraction Guidelines for Portable and Aftermarket Devices. Technical report 233, NHTSA, 2016, pages 87656–87683.

[32] D. C. Niehorster, T. H. W. Cornelissen, K. Holmqvist, I. T. C. Hooge, and R. S. Hessels. What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, 50(1):213–227, Feb. 2018.

[33] J. Nielsen. *Usability Engineering*. Elsevier Science & Technology Books, San Diego, CA, 1st. ed. Edition, 1994.

[34] Parks Associates. Penetration of Smart Speakers with Voice Assistants Will Reach 47% of U.S. Broadband Households By 2022. *Journal of Engineering; Atlanta*:1–2, 2018.

[35] A. Pauzié. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, 2(4):315–322, 2008.

[36] J. Preece, H. Sharp, and Y. Rogers. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, Hoboken, 2015.

[37] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. o. McNamara, and S. M. Williams. *Neuroscience: Third Edition*. Sinauer Associates, Inc., Sunderland, MA, 3rd. ed. Edition, 2000.

[38] G. B. Reid and T. E. Nygren. The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In pages 185–218. 1988.

[39] B. Reimer, B. Mehler, I. Reagan, D. Kidd, and J. Dobres. Multi-modal demands of a smartphone used to place calls and enter addresses during highway driving relative to two embedded systems. *Ergonomics*, 59(12):1565–1585, Dec. 2016.

[40] H. W. J. Rittel and M. M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, June 1973.

[41] SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical report, SAE International, 2018, pages 2–35.

[42] D. Strayer, J. Cooper, J. Turrill, J. Coleman, and R. J. Hopman. Measuring Cognitive Distraction in the Automobile III: A Comparison of Ten 2015 In-Vehicle Information Systems. (202):52, 2015.

[43] G. T. Taoka. Duration of drivers' glances at mirrors and displays. *ITE Journal (Institute of Transportation Engineers)*, 60(10):35–39, 1990.

[44] Vinnova. SEER - Sömlös, Effektiv och bEhaglig inteRaktion (SEER), 2016. URL: `https://www.vinnova.se/p/seer---somlos-effektiv-och-behaglig-interaktion-seer/` (visited on 02/12/2019).

[45] D. D. Waard. *The Measurement of Drivers' Mental Workload.* The Traffic Research Centre (now Centre for Environmental and Traffic Technology), University of Groningen, Haren, 1996.

[46] Y. Wadsworth. *Do It Yourself Social Research.* Left Coast Press, New York, NY, third edit edition, 2011.

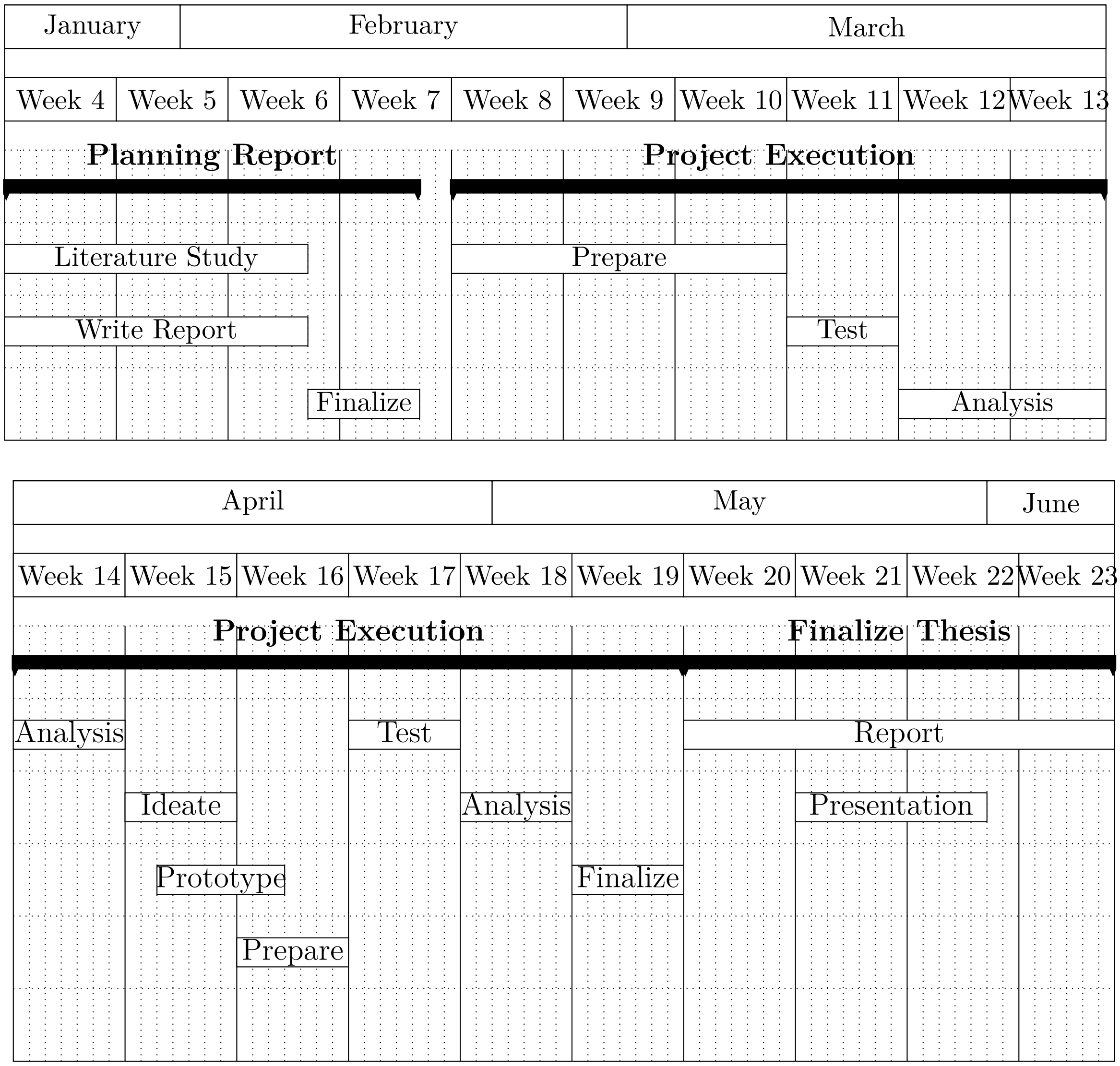
[47] K. Whitenton. Voice First: The Future of Interaction?, 2017. URL: `https://www.nngroup.com/articles/voice-first/` (visited on 02/06/2019).

[48] C. D. Wickens. Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455, June 2008.

[49] F. R. H. Zijlstra. Efficiency in Work Behavior: A Design Approach for Modern Tools. *PhD diss., Technical University Delft. Delft: Delft University Press*, (January 1993), 1993.

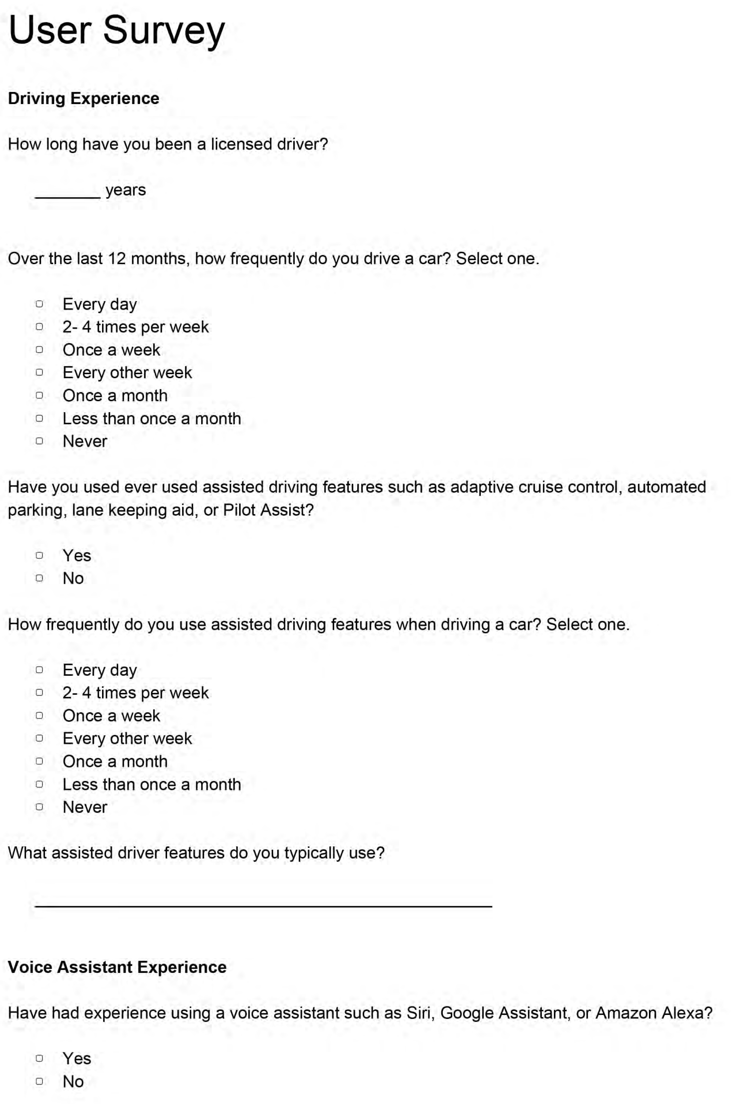
A

Project Plan GANTT Chart



B

Summative Evaluation Survey



User Survey

Driving Experience

How long have you been a licensed driver?

_______ years

Over the last 12 months, how frequently do you drive a car? Select one.

- □ Every day
- □ 2- 4 times per week
- □ Once a week
- □ Every other week
- □ Once a month
- □ Less than once a month
- □ Never

Have you used ever used assisted driving features such as adaptive cruise control, automated parking, lane keeping aid, or Pilot Assist?

- □ Yes
- □ No

How frequently do you use assisted driving features when driving a car? Select one.

- □ Every day
- □ 2- 4 times per week
- □ Once a week
- □ Every other week
- □ Once a month
- □ Less than once a month
- □ Never

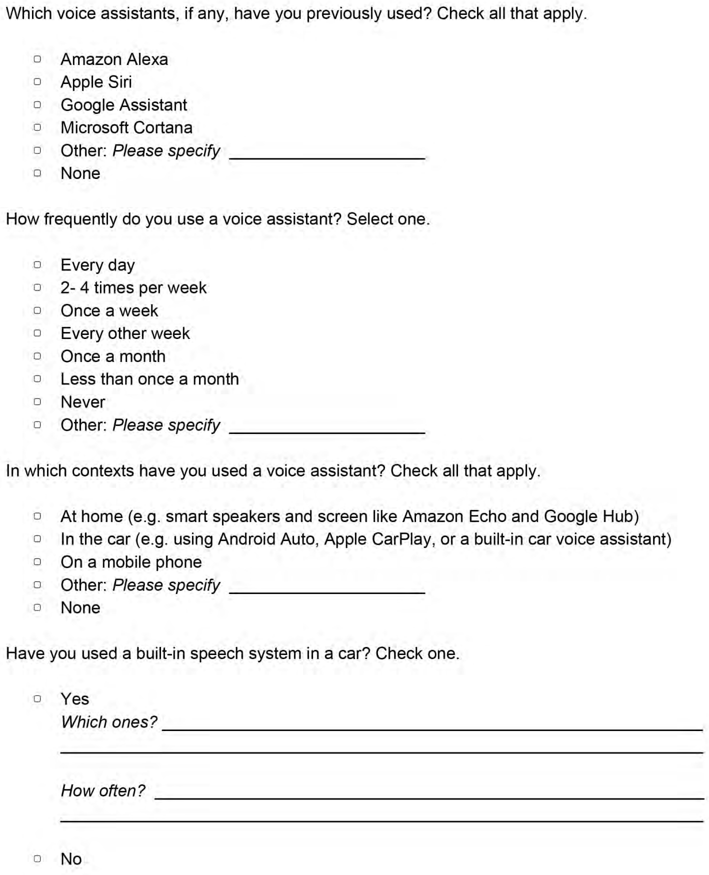
What assisted driver features do you typically use?

____

Voice Assistant Experience

Have had experience using a voice assistant such as Siri, Google Assistant, or Amazon Alexa?

- □ Yes
- □ No



Which voice assistants, if any, have you previously used? Check all that apply.

- □ Amazon Alexa
- □ Apple Siri
- □ Google Assistant
- □ Microsoft Cortana
- □ Other: *Please specify* ____________________
- □ None

How frequently do you use a voice assistant? Select one.

- □ Every day
- □ 2- 4 times per week
- □ Once a week
- □ Every other week
- □ Once a month
- □ Less than once a month
- □ Never
- □ Other: *Please specify* ____________________

In which contexts have you used a voice assistant? Check all that apply.

- □ At home (e.g. smart speakers and screen like Amazon Echo and Google Hub)
- □ In the car (e.g. using Android Auto, Apple CarPlay, or a built-in car voice assistant)
- □ On a mobile phone
- □ Other: *Please specify* ____________________
- □ None

Have you used a built-in speech system in a car? Check one.

- □ Yes

  Which ones? ____

  ____

  How often? ____

  ____

- □ No

C

Summative Evaluation Test Protocol

Description

Test participants will be asked to perform a set of secondary tasks under for conditions:

	Manual Drive	**Assisted Drive**
Google Assistant	Google Assistant & Manual Drive	Google Assistant & Assisted Drive
Siri	Siri & Manual Drive	Siri & Assisted Drive

The evaluation is organized so that test participants will first be randomly assigned one of the two voice assistants. The test participant will then be trained on that assistant and asked to perform voice interactions using the voice assistant while driving. The test participant will perform the tasks in using both manual and assisted drive modes. The order of which drive mode is used first is randomly assigned. The test participant will then be trained on the second voice assistant and asked to use the second voice assistant for the second half of the evaluation. Test participants will then perform tasks using both manual and assisted drive in a randomly assigned order.

This evaluation is to assess how drivers interact with voice assistants while driving and to compare the two voice assistants against each other. This evaluation also aims to compare differences in interactions between the two driving modes.

See the *Evaluation Schedule* section for more details about the permutations of the evaluation. Each test participant will be assigned one of eight possible permutations.

Test participants: 8 - 12 licensed drivers

Script

Introduction	Hello, we are two master thesis students exploring voice interaction in vehicles. Today, we want to learn more about the existing voice assistants, like Google Assistant and Siri, and how they are used in cars. If you prefer to have the instructions be done in Swedish instead, that is possible, but the actual test itself should be done in English. The test today will start with a questionnaire. You will be asked to perform a series of tasks using a voice assistant we have set up in the car. You will receive training for the different voice assistants you will be trying as well as some opportunity to get acquainted with the drive conditions of the car. Throughout the test, you will be driving on public roads while asked to perform a series of tasks. After each test drive, you will be asked to fill out a short survey. When all of the test drives have been completed, we will have a short debrief to ask you some questions about your experience. The test should take about 2 hours to complete, We would like to emphasize that we are testing the system and not you. Additionally, we want to emphasize safety as being the top priority in this test, so if you feel that you cannot safely complete a task, you can let us know and we can pause or stop the test. You can stop the test at any point. We will be recording data from this test but any data that may be published from this test will be anonymized. We will be using this data to evaluate the system and its effect on driving. Just like the test, you can stop the collection of data at any point and contact us if you want the data to be removed. I you have any questions, please let us know. Otherwise, would you like to continue? To, continue, please fill out this consent and waiver form from Volvo.
Questionnaire & Training	Great! First we would like you to fill out this questionnaire about your driving habits and your previous experience with voice assistants. Now we will go over the basics of this car. We will also go over how to activate pilot assist which you will be asked to use for a portion of the test. (Drive training is only necessary on the first sessions)

Questionnaire & Training	• Accelerator and brakes • Gear shift • Emergency brake • Turn indicators • Push-to-talk button • Pilot Assist Do you have any questions now about the car? If not, you can start the car and we will drive to the first location where you will receive training of the voice assistant we'll be using. On the drive there, please try activating the Pilot Assist so you can get acquainted with it. After parking in an area that will serve as the start of the test drive route. Now we will go over training of the voice assistant you will be using. You will start by using <Google Assistant through Android Auto/Siri through Apple CarPlay>. The voice assistant has already been connected to the car for you, so we will cover how you can interact with the system through manual input and voice command. The voice assistant is set up in English so you should use commands in English. We would also like to encourage you to try some commands yourself as you think of them. To start, let's go over the basics of the visual interface. Demonstrate manual input of the system, e.g. what gestures are available, where menus are, how to return to home screen. Now let's move on to the voice command. You can start a voice command using the push-to-talk button on the steering wheel. You will hear a tone after which you can say a command. You can also start a voice command using a hotword or trigger phrase. For this system, the hotword is "<Ok Google,Hey Siri>". I'll demonstrate a few example commands. You can structure commands as a question or a command. For example, "<Ok Google, Hey Siri>, what's the forcast today?" another example is "<Ok Google/Hey Siri>, tell me what's on my calendar for today." Do you have any questions so far?

Questionnaire & Training	Let's move onto the kinds of features and tasks you will be using during the test. Coach the tester if they have difficulty completing a task using voice. **Communications** You can use your voice assistant to read your messages and then reply to them. Try to read and reply to a text message using the voice assistant. Voice assistants are also capable of making calls for you. Try to make a call to <Jessica Jones> using the voice assistant. **Media** Voice assistants are capable of playing music from audio apps on the phone. Try using the voice assistant to play music. Let's also try playing music using different levels of specificity. For example, you can play music by a genre, song, artist or playlist name. **Navigation** You can use the voice assistant to get directions to a place or address. Try using the voice assistant to get directions to a nearby restaurant. You can also use the voice assistant to stop any active commands. Try using your voice to cancel the navigation directions we just started. You can also interrupt a voice command using a long press on the push-to-talk button. Now that you are familiar with the voice assistant and the car, we can begin the test if you don't have any questions.
Tasks	Now let's start the test. You will be driving along a public road in a route we have designated for the test. As you drive, you will be given the necessary driving directions for the route. You will also be given a task to complete using the voice assistant. After each scenario, you will be asked to complete a short survey about the system. As a reminder safety is the highest priority, so you should focus on driving safely rather than the task. Are you ready to begin?

Tasks	You will complete a set of tasks first while using <first drive condition> and then again using <second drive condition> Test participant may start driving along designated route. Allow three attemps if there is a failure, after which the test should move on to the next task. **Messaging** Facilitator sends incoming message. Can you show me how you would open the message you just received using the voice assistant? Don't reply. Can you show me how you send a new message to your contact <Anita Davis> to let her know you are going to be late? **Calling** ¨ Can you show me how you would call your colleague Jessica Jones? Can you show me how you would call your friend, <Benjamin Chen> on his home phone? **Music** Can you show me how to play a genre of music using the voice assistant? Can you show me how you would play a specific song using the voice assistant? **Addresses** Can you show me how you would get directions and start navigation to <18 Harper Road, London; 6 Abbey Road, London> using the voice assistant? **POIs** Can you show me how you would get directions for a nearby cafe on your route? End navigation from previous task so that next task is "fresh". Can you show me how you get directions and start navigation to the nearest McDonald's?

Tasks	-Stop car at safe spot- Now, we would like to ask you to complete this survey about the overall experience of the system you have just used while in <first drive condition> Great, now we try completing the same tasks while using <second drive condition>. Repeat tasks using the second assigned drive condition. Great work. Next we will try using < second voice assistant>.
Optional Break	Would you like a break before we continue to the next part of the test?
Training (2)	No training on the conditions is necessary. This training session should be shorter than the first with an abbreviated task introduction and special emphasis should be given to the navigation tasks. **Communications** Let's try sending a message to a contact named <Anita Davies>. **Navigation** You can use the voice assistant to get directions to a place or address. Try to use the voice assistant to get directions to a nearby restaurant.
Tasks (2)	See previous Tasks section for script.
Debrief	How was your overall experience using voice assistants while driving? Was there anything positive about the either voice assistant that stood out to you? Was there anything negative about either voice assistant that stood out to you? How useful did you find the voice assistant features? What other voice commands or features would you like to have while driving? Would you use voice assistants again in your own car?

Debrief	Would you recommend them to anyone else? Why? How did the two voice assistants compare? Ease of use? Accurate responses? Do you have any more thoughts about using the voice system while driving? What about the different driving conditions (manual vs. pilot assist)? How did they compare to each other when using the voice assistants? While using Google Assistant? While using Siri? Did you feel that the assisted drive affected your overall driving and performance of the tasks? Do you have any questions for us or about the test or the voice assistants you tested today? Is there anything else you would like to add about your experience today? Thank you so much for your time. If you have any questions, you can contact us.

D

Summative Evaluation Test Schedule

Evaluation Schedule

Testers will perform tasks using voice assistants and drive modes assigned in a random order. Below is a chart of the generalized schedule and possible permutations if the schedule. Each tester will be assigned one of the eight possible permutations.

Generalized Scedule:

Introduction
Training for both drive conditions and Voice Assistant 1
Voice Assistant 1 & Drive Condition 1
Voice Assistant 1 & Drive Condition 2
Optional Break
Training for Voice Assistant 2
Voice Assistant 2 & Drive Condition 1
Voice Assistant 2 & Drive Condition 2
Debrief

Testing Permutations

GA = Google Assistant	S = Siri
MD = Manual Drive	AD = Assisted Drive

Permutations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Introduction								
Training for both conditions and Voice Assistant 1	GA MD AD	GA MD AD	GA MD AD	GA MD AD	S MD AD	S MD AD	S MD AD	S MD AD
Voice Assistant 1 & Drive Confition 1	GA MD	GA AD	GA MD	GA AD	S MD	S AD	S MD	S AD
Voice Assistant 1 & Drive Confition 2	GA AD	GA MD	GA AD	GA MD	S AD	S MD	S AD	S MD
Optional Break								
Training for Voice Assistant 2	S	S	S	S	GA	GA	GA	GA
Voice Assistant 2 & Drive Condition 1	S MD	S MD	S AD	S AD	GA MD	GA MD	GA AD	GA AD
Voice Assistant 2 & Drive Condition 2	S AD	S AD	S MD	S MD	GA AD	GA AD	GA MD	GA MD
Debrief								

E

Prototype Evaluation Test Protocol

Description

Testers will be asked to perform a set of secondary tasks on two different VA prototypes while driving.

Prototype 1	**Prototype 2**

The evaluation is organized so that testers will first be randomly assigned one of the two voice assistants. The tester will then be trained on that assistant and asked to perform voice interactions using the voice assistant while driving. The tester will then be trained on the second voice assistant and asked to use the second voice assistant for the second half of the evaluation.

This evaluation is to assess how drivers interact with voice assistants while driving and to compare the two voice assistants against each other.

Introduction	Hello, we are two master thesis students exploring voice interaction in vehicles. Today, we want to learn more about two voice assistant prototypes that has been developed and how they are used in this driving simulator. The test today will start with a short questionnaire. You will be asked to perform a series of tasks using a voice system we have set up in the simulator. You will receive training for the different voice assistants you will be trying as well as some an opportunity to get acquainted with the driving simulator. After each set of tasks, you will be asked to fill out a short survey. When all of the tasks have been completed, we will have a short debrief to ask you some questions about your experience. The test should take about 90 minutes to complete. We would like to emphasize that we are testing the system and not you. If you feel uncomfortable or like you cannot complete a task for any reason, you can let us know and we can pause or stop the test. You can stop the test at any point. We will be recording data from this test but any data that may be published from this test will be anonymized. We will be using this data to evaluate the system and its effect on driving. Just like the test, you can stop the collection of data at any point and contact us if you want the data to be removed. If you have any questions, please let us know. Otherwise, would you like to continue? To continue, please fill out this waiver form from Volvo.
Questionnaire & Training	Great! First we would like you fill out this questionnaire about your driving habits and your previous experience with voice assistants. After completing questionnaire, proceed to training using the driving simulator. Go through: - Gas/brake - Gear shifting - Steering wheel - Blinkers - Speedometer - Driving directions

Questionnaire & Training	After completing drive training, proceed to training using first VA prototype. Now we will go over the training of the voice assistant you will be using. You will start by using <Prototype 1/Prototype 2>. The voice assistant is connected to the car simulator, so we will cover how you can interact with the system through manual input and voice command. The voice assistant is set up in English so you should use commands in English. To start, let's go over the basics of the visual interface. Show screen interface. If you feel you need to interact with the screen, it is a touch interface so you can do so. However, not all element are active, since this is still just a prototype and not a fully developed application. When you are doing the tasks the screen will work for all commands related to the task. Now let's move on to the voice command . You can start a voice command using a hotword or trigger phrase. For this system, the hotword is <"Hey Carla/Ok Carla">. I'll demonstrate a few example commands. You can structure commands as a question or a command. For example, <"Hey Carla/Ok Carla">, what's my latest message?" Another example is, <"Hey Carla/Ok Carla">, read me my latest message" Do you have any questions so far? Let's move onto the kinds of features and tasks you will be using during the test. Coach the tester if they have difficulty completing a task using voice. Communications Let's try by sending a message to a contact named <Anita Davis>. Navigation You can use the voice assistant to get directions to a place or address. Try to ask it for some nearby restaurant.

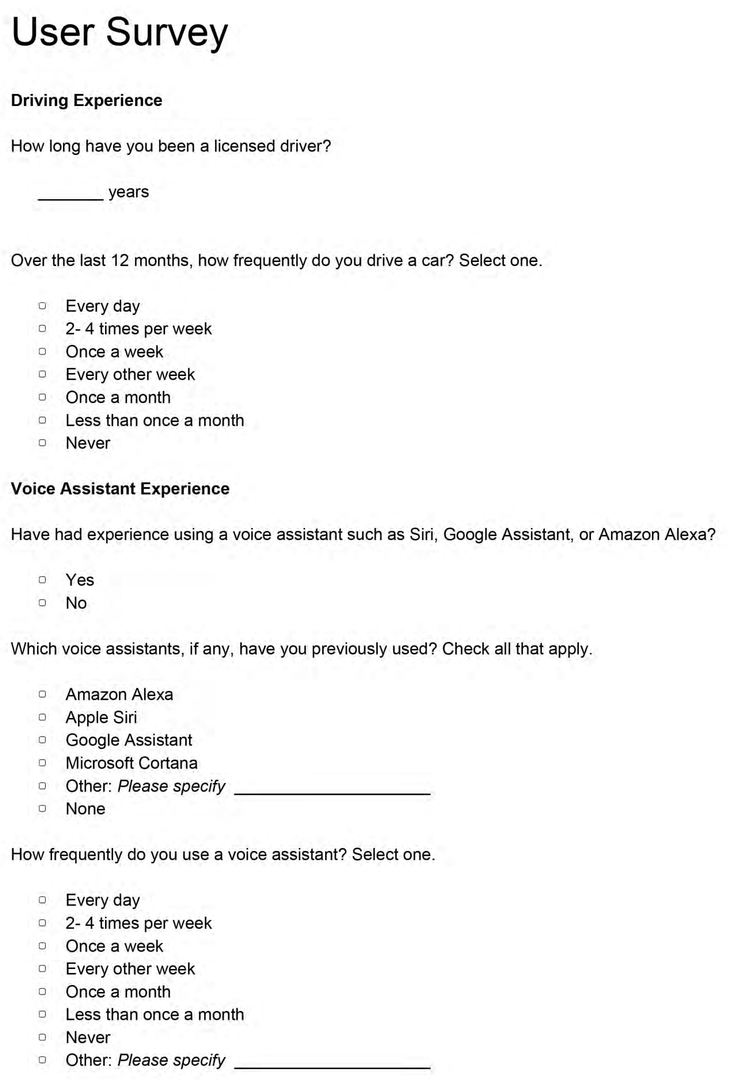
Questionnaire & Training	You can also use the voice assistant to stop any active commands. For <Prototype 1> Show standard way of doing it For <Prototype 2> Show interjections & standard way Reset screen - ask for restaurants Show interjection symbol It can be used to cancel or change a search/command Cancel - Reset screen Try getting it to show you nearby restaurants and then change it show cafes
Tasks	Now let's start the test. As you drive, just follow the directions on screen. You will also be given tasks to complete using the voice assistant. Are you ready to begin? You will complete a set of tasks first while using <first prototype> and then again using <second prototype>. **Messaging** Can you show me how you would open the message you just received using the voice assistant? Don't reply. Can you show me how you send a new message to your contact Anita Davis to let her know you are going to be late? **Calling** Can you show me how you would call your friend, Benjamin Chen on his home phone? **Music** Can you show me how to play Jazz using the voice assistant? **Addresses** Can you show me how you would get directions to <18 Harper Road, London> using the voice assistant? **POIs** Can you show me how you would get directions for a nearby cafe on your route? Can you show me how you get directions to the nearest McDonald's? That was the first part of the test, good job. DALI and SUS

Optional Break	Would you like a <X> minute break before we continue to the test?
Training (2)	This training session should be shorter than the first with an abbreviated task introduction and special emphasis should be given to the relevant tasks. **Communications** Let's try by sending a message to a contact named <Anita Davis>. **Navigation** You can use the voice assistant to get directions to a place or address. Try to ask it for some nearby restaurant. You can also use the voice assistant to stop any active commands. For <Prototype 1> Show standard way of doing it For <Prototype 2> Show interjections & standard way Reset screen - ask for restaurants Show interjection symbol It can be used to cancel or change a search/command Cancel - Reset screen Try getting it to show you nearby restaurants and then change it show cafes
Tasks (2)	See previous Tasks section for script.
Debrief	How was your overall experience using voice assistants while driving? Was there anything positive about the either voice assistant that stood out to you? Was there anything negative about the either voice assistant that stood out to you? How useful did you find the voice assistant features? What other voice commands or features would you like to have while driving? Would you use voice assistants again or in your own car? Would you recommend them to anyone else? Why?

Debrief	How did the two voice assistants compare? Ease of use? The way they responded? Do you have any more thoughts about using the voice system while driving? **Workshop** Each question can include follow-up questions such as, what are your suggestions for alternatives? How do you think they could be improved? What did you like about this feature? What did you didn't like about it? What are your thoughts on the processing sounds in Prototype 1? What are your thoughts on the suggestion chips in Prototype 2? Did you notice them or use them? What are your thoughts on the interjection feature in Prototype 2? Did you use notice it or use it? How can it be improved? Let's look at encountering the bad connection error. Which prototype handled this error better? What do you think is the ideal way to handle it? What about if the connection is lost for 30 seconds versus 20 minutes? What are your thoughts on selecting an item from a list in Prototype 1 vs. Prototype 2? What are your thoughts on selection a route from a list in Prototype 1 vs. Prototype 2? What are your thoughts on the way the system handle errors dealing with speech recognition or its own feature limitations? Thank you so much for your time. If you have any questions, you can contact us at <email address>.

F

Prototype Evaluation Survey



User Survey

Driving Experience

How long have you been a licensed driver?

_______ years

Over the last 12 months, how frequently do you drive a car? Select one.

- □ Every day
- □ 2- 4 times per week
- □ Once a week
- □ Every other week
- □ Once a month
- □ Less than once a month
- □ Never

Voice Assistant Experience

Have had experience using a voice assistant such as Siri, Google Assistant, or Amazon Alexa?

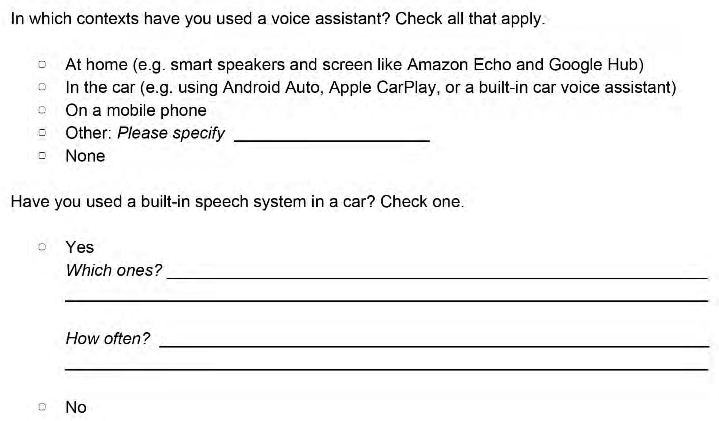
- □ Yes
- □ No

Which voice assistants, if any, have you previously used? Check all that apply.

- □ Amazon Alexa
- □ Apple Siri
- □ Google Assistant
- □ Microsoft Cortana
- □ Other: *Please specify* ____________________
- □ None

How frequently do you use a voice assistant? Select one.

- □ Every day
- □ 2- 4 times per week
- □ Once a week
- □ Every other week
- □ Once a month
- □ Less than once a month
- □ Never
- □ Other: *Please specify* ____________________



In which contexts have you used a voice assistant? Check all that apply.

- □ At home (e.g. smart speakers and screen like Amazon Echo and Google Hub)
- □ In the car (e.g. using Android Auto, Apple CarPlay, or a built-in car voice assistant)
- □ On a mobile phone
- □ Other: *Please specify* ____________________
- □ None

Have you used a built-in speech system in a car? Check one.

- □ Yes
  Which ones? ____
  ____

  How often? ____
  ____

- □ No

G

Interaction Paths of VA Prototypes

Task 1 - Read Message

- Success: single shot
- Error: no recognition
- Error: no function
- Error: bad connection

Task 2 - Write Message

- Success: single shot
- Success: multi-step
- Error: speech recognition
- Error: wrong message

Task 4 - Call Ben

- Success: single shot
- Success: multi-step
- Error: wrong number
- Error: bad connection
- Error: no recognition

- Error: no function

Task 5 - Play Genre

- Success: single shot
- Error: wrong genre
- Error: bad connection
- Error: no recognition
- Error: no function

Task 7 - Navigate to Adress

- Success: single shot
- Success: multi-step
- Success: 2 routes
- Error: wrong adress
- Error: bad connection
- Error: no recognition
- Error: no function
- Error: wrong city

Task 8 - Add Café

- Success: multi-step
- Error: wrong café
- Error: no function

Task 9 - Navigate to McDonald's

- Success: single shot
- Success: multi-step list of options

- Success: multi-step 2 routes
- Error: bad connection
- Error: no recognition
- Error: no function
- Error: wrong city

H

Prototype Evaluation Schedule

Schedule 1

	Prototype 1	**Prototype 2**
1	No Recognition	
2		Speech Recognition
4	No Function	
5	Processing + Bad Connection	Processing
7		Processing + Bad Connection
8	Processing	
9		No Function

Schedule 2

	Prototype 1	**Prototype 2**
1	Processing + Bad Connection	No function
2		
4	Wrong Number	
5	No Recognition	Wrong Genre
7		Processing + Bad Connection
8	Processing	
9		Processing

Schedule 3

	Prototype 1	**Prototype 2**
1	Processing	
2	Wrong Message	Processing + Bad Connection
4	No Recognition	
5		Wrong Genre
7	Processing + Bad Connection	Processing + Two Routes
8		No Function
9	Two Routes	

Schedule 4

	Prototype 1	**Prototype 2**
1	No Function	Processing
2	Wrong Message	
4		No Function
5	Processing + Bad Connection	No Recognition
7		Two Routes
8	Processing	
9		Processing + Bad Connection

Schedule 5

	Prototype 1	**Prototype 2**
1		No Recognition
2		
4	Processing + Bad Connection	Processing
5	No Function	Processing + Bad Connection
7	Wrong City	
8		No Function
9	Processing	Two Routes

Schedule 6

	Prototype 1	**Prototype 2**
1		
2	Processing	Processing
4		No Recognition
5	No Recognition	Processing + Bad Connection
7	Two Routes	Wrong Address
8	No Function	
9	Processing + Bad Connection	Two Routes

Schedule 7

	Prototype 1	**Prototype 2**
1	No Recognition	
2		Speech Recognition
4	No Function	Processing
5	Processing + Bad Connection	
7		Processing + Bad Connection
8	Processing	
9		No Function

Schedule 8

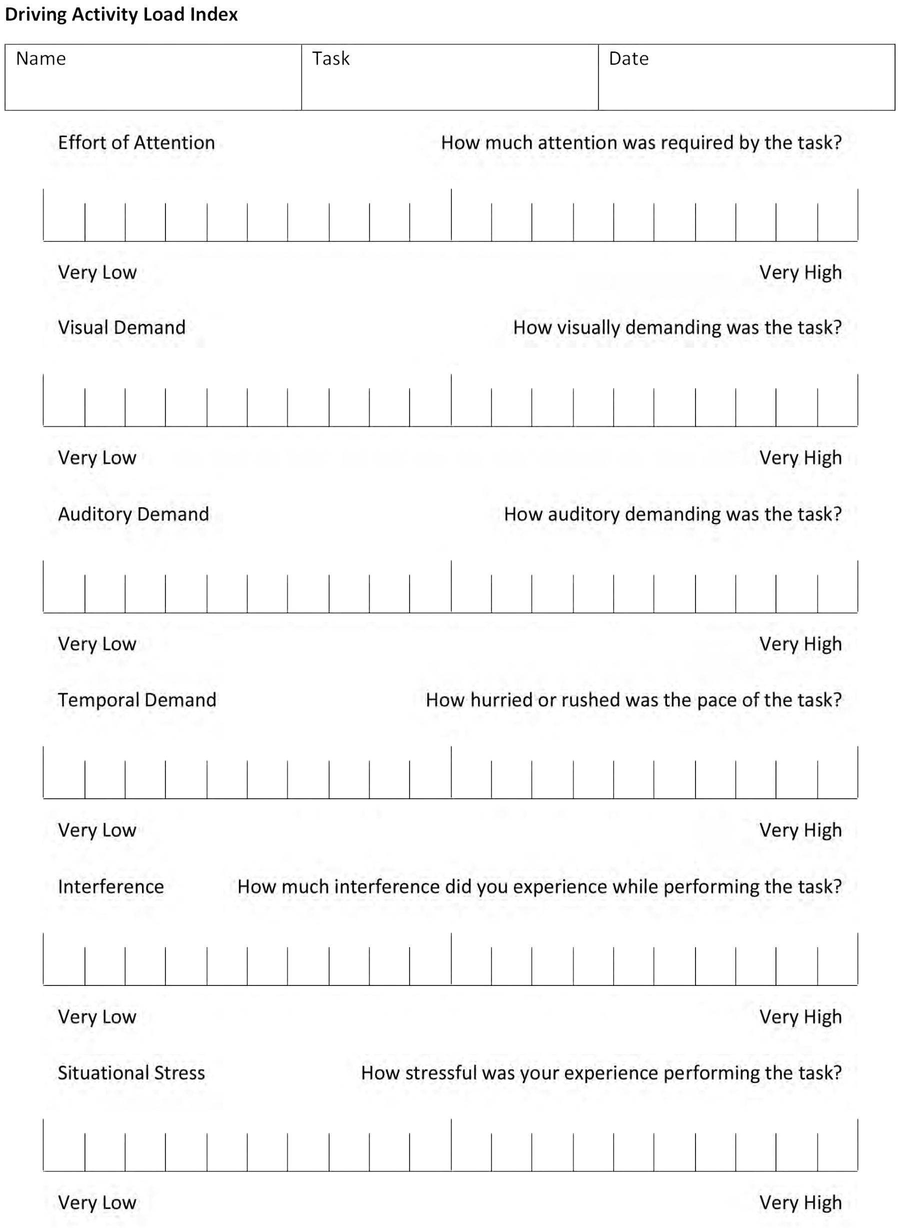
	Prototype 1	**Prototype 2**
1	Processing + Bad Connection	No Function
2		
4	Wrong Number	Processing
5	No Recognition	Wrong Genre
7		Processing + Bad Connection
8	Processing	
9		

Schedule 9

	Prototype 1	**Prototype 2**
1	Processing	
2	Wrong Message	Processing + Bad Connection
4	No Recognition	
5		Wrong Genre
7	Processing + Bad Connection	Processing + Two Routes
8		No Function
9	Two Routes	

I

DALI Survey



Driving Activity Load Index

Name	Task	Date

Effort of Attention — How much attention was required by the task?

Very Low — Very High

Visual Demand — How visually demanding was the task?

Very Low — Very High

Auditory Demand — How auditory demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

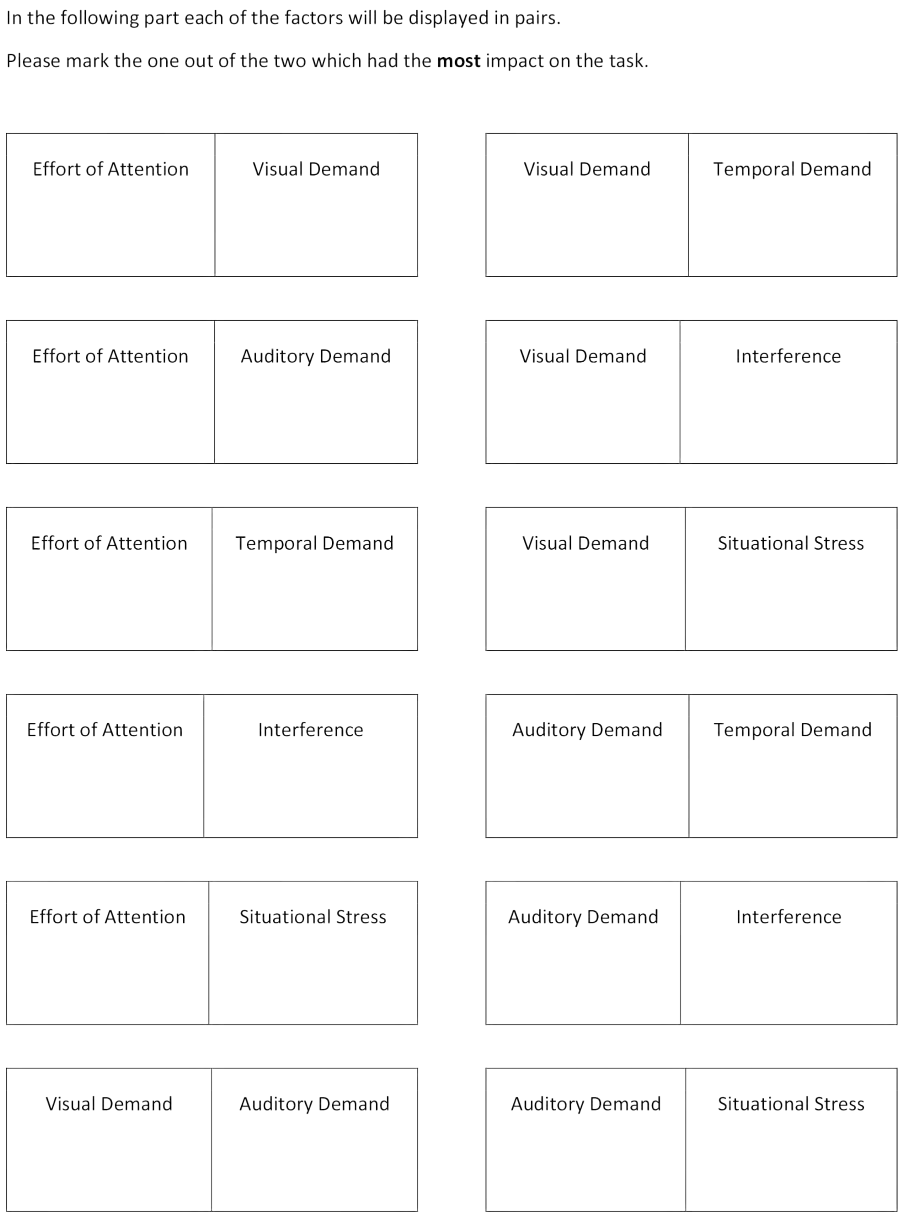
Very Low — Very High

Interference — How much interference did you experience while performing the task?

Very Low — Very High

Situational Stress — How stressful was your experience performing the task?

Very Low — Very High



In the following part each of the factors will be displayed in pairs.

Please mark the one out of the two which had the **most** impact on the task.

Effort of Attention	Visual Demand
Effort of Attention	Auditory Demand
Effort of Attention	Temporal Demand
Effort of Attention	Interference
Effort of Attention	Situational Stress
Visual Demand	Auditory Demand
Visual Demand	Temporal Demand
Visual Demand	Interference
Visual Demand	Situational Stress
Auditory Demand	Temporal Demand
Auditory Demand	Interference
Auditory Demand	Situational Stress

Temporal Demand	Interference

Temporal Demand	Situational stress

Interference	Situational Stress

J

SUS Survey

Please enter your participant number: _____

System Usability Scale (SUS)

This is a standard questionnaire that measures the overall usability of a system. Please select the answer that best expresses how you feel about each statement after using the website today.

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1.	I think I would like to use this tool frequently.	□	□	□	□	□
2.	I found the tool unnecessarily complex.	□	□	□	□	□
3.	I thought the tool was easy to use.	□	□	□	□	□
4.	I think that I would need the support of a technical person to be able to use this system.	□	□	□	□	□
5.	I found the various functions in this tool were well integrated.	□	□	□	□	□
6.	I thought there was too much inconsistency in this tool.	□	□	□	□	□
7.	I would imagine that most people would learn to use this tool very quickly.	□	□	□	□	□
8.	I found the tool very cumbersome to use.	□	□	□	□	□
9.	I felt very confident using the tool.	□	□	□	□	□
10.	I needed to learn a lot of things before I could get going with this tool.	□	□	□	□	□

K

Summative Evaluation KJ Results

- Reduce the amount of errors to reduce the visual distraction (Glances)
  - Skepticism diminishes with extended successful user, BUT trust is easily lost when an error occurs (VAs while driving)
  - Drivers want to know the error. If it is a user or system error and if it is something they can solve on their own (VAs while driving)
- Use multiple channels for output, always voice first. Never use only 1 channel. (Errors - No Match)
  - Match visual and voice output to action, e.g. adding a cafe to a route (Unclear next steps)
  - Voice assistant output and screen must match (Unclear next steps)
- Audio tones draws attention to CSD (Glances)
- Use context of car to avoid extra stress (Driving, example: don't expect user to focus on V.A. if the P.A. currently is signalling an error)
  - Consider the context (Errors - No Match, example: if map is active, expect user to give commands related to it)
- Repeating error messages gets annoying - memorize the user's preferences (Errors - No Match)
- Remind user when the connection is good again, consider pausing and auto resuming tasks (Errors - Connection/Data)
- Voice assistant doesn't communicate that it is processing well enough (Status of VA)
  - Google visual elements as an inspiration for conveying voice assistant

status (Siri vs Google - Google)

- Non-sequitur responses causes negative emotions and distrust (Errors - Non-Sequitur)
- Make recovery from speech recognition errors possible (Errors - Wrong, Speech Recognition)
- Option to choose between multi-step and single-shot interactions (Flow)
- Guiding users through multi-step interactions (Flow)
- Use memory for most likely option if more than one matches (Flow)
- Account for users providing extra info [Breaks guidelines] (Disregards info, Stops listening)
- Adjust response to geographical context (Irrelevant, Language support)
- If voice command is unclear, drivers will use hands to complete an action on the touchscreen which is bad (Unclear next steps)
- Use large and clear visual elements (Unclear next steps)
- CSD as a guide for voice input (Adjusting voice input)
- Users change their way of speaking as they lose confidence in the voice assistant (Adjusting voice input)
- Visual to help convey voice assistant status (Status of VA - Processing)
- Sound as an option to indicate status (Status of VA - Processing)
- Screen state should match required input from user (Status of VA - Processing)
- Responses are easy to understand and non-repetitive (Siri vs Google - Siri)
- Users feel defeated when voice assistants require extra work to complete a task (Emotion)
- Users treat voice assistants like a person, using existing speech patterns (Emotion)
- Guide user to be effective with their voice commands (Bonus)
- Overly verbose voice responses which can be made more concise (“Correct”)

- Persona for a typical user of voice assistants while driving would be a heavy phone users (Useful? Would recommend?)

L

Prototype Evaluation KJ Results

Choosing From a List

- More audio communication from VA generally helpful to the user - makes the user feel more in control
    - Smart auto-selection/start might work well if pre-determined, but automatically takes away some control the user (could be a setting)
    - Give contextual information that is relevant, e.g. time to destination, prices, etc
- Responsive - The VA should understand if the user already knows what she/he wants and in those cases give the correct result as quickly as possible, or if he she/he is exploring and in that case present more options
- When presenting only one option without the option to hear more, users may feel obligated to select that first options
- Back and forth glancing when VA presents different options for McDonald's. Present what distinguishes them

Navigation Glances

- Presenting one out of two routes and autostart causes confusion and visual distraction
- Seem to only be in Prototype 2 where user only has one option read out.

Processing

- Current solution not clear enough
- Either solution causes a lot of eyes of road time

- When an action takes longer than expected, drivers check CSD to see what is happening
- Sound during processing seems to draw attention to the CSD on par with a no sound prototype
- Processing state is unclear. Drivers will retry a command or think they should in order to complete the request.

Interjections

- Users should be able to cancel/stop VA at any time
- Current solution not very useful when it comes to changing/tweaking command
- Drivers like the interjection feature in theory, but didn't really use it during the test
- Interjections interrupt the flow of the conversation and can sometimes add stress to a situation
- Current implementation of interjections can be improved; a bit messy to do
- Drivers most like the ability to cancel or quiet the system
- Wasn't always clear when to use the interjection

Processing Sound

- Current sound is bad
- Unclear whether sound or no sound is best solution
- Most people were confused by the sound and some confused it with making a call
- One person believes that once you get used to processing flow, you won't need a sound

Phrasing Nav

- Allow varied input through voice and touch
  - Allow user to interact with items on screen
- Mostly answered in the affirmative when asked a yes/no question

- Drivers started navigation using option label or location name

Conversation vs Visual

- Aim for conversational voice interaction, allowing the user full control of their actions
- Don't automatically choose options for user if he/she hasn't specifically stated that he/she wants a certain types of choices (i.e. always choose fastest route)
- Important nav. info like which route is the fastest one should be presented through audio
- Drivers preferred a more conversational approach, made it easier to hear multiple options
- Drivers wanted to hear what distinguishes one options from another
- Auto-start between routes made one driver feel like he missed out on making a choice
- Drivers liked hearing the contextual information (distance from current location)

Errors

- Current solution seems somewhat adequate
  - If parts of the user's input was understood but other parts wasn't, ask user to fill in the blanks
- Drivers want to know the cause of an error and what to say differently for it to work
- When drivers don't know what went wrong, they usually repeat the same command either in a smaller part, more slowly, or more clearly
- Errors with a clear prompt for action went largely unnoticed by drivers

Confusion on Bad Connection

- VA's connection status needs to be clearly communicated to user
- Clearly communicate through voice, what the VA is doing and will do
- In theory, a potentially good features, but most were confused on how it worked

- The lack of status of whether or not the voice assistant was seeking re-connection added to the confusion
- Perhaps easiest to end the interaction and ask the user to try again later
- Can the voice assistant predict when it will travel through areas of low coverage?

Bad Connection Retries

- See above
- Due to confusion about how the bad connection feature worked, most drivers just re-tried the command directly

Bad Connection Ask

- Users prefer that the VA asks before retrying an action
- Asking when the connection is restored is at most times necessary
- Asking when the connection first goes bad is not as important
- Generally, drivers prefer it if the voice assistant asked before completing an action after re-establishing a connection
- If the connection disconnect is very short, then it doesn't matter

Thoughts About VAs

- It should adapt to the user over time. Personalized to the user like a real personal assistant

Visual Integration

- Screen was most important during navigation, of the tasks tested
- In other tasks, glances where directed to the VA status

Turn Taking

- Expect the user to be quick with their replies
- VA status sound mostly good and useful (processing exception)
- Most drivers understood when it was their turn (exceptions, processing, bad

connection)

- Some drivers wanted to give a command with the trigger word in a single shot

Suggestion chips

- Suggestion chips useful in specific scenarios (back-up and learning) but should be designed so that they do not cause distraction during normal interaction
- Half of tester felt there was a benefit to the chips, either as a backup or a guide for beginners
- One driver felt it took caused them to focus on the screen when driving

Features

- Currently existing functions good, missing calendar and car functionality
- Mostly would use a voice assistant for music and navigation.
- Would like access to other functions, such as calendar, email, car information, notes

Situations for Recommendation

- For people who want to or normally use their phones while driving
- For people who are used to using a voice assistant, since it requires a different approach to speech

Phrasing

- VA should ask questions so that the user know which his/her options for answering is
  - If the user fails to answer, further clarify "Which did you want me me to do ... or ... "
- People sometimes answered in the affirmative when given two alter natives because the prompt started with a verb. E.g. Do you want to send it or change it?
- Hard to know how to phrase requests and if they could be dismissed
- Testers had to change and think more about how to compose their requests

VA Activation

- Hey Carla annoying when performed many times - allow customization
- Saying "Ok Carla" or "Hey Carla" feels tedious and awkward after a while

Beginner

- Low discoverability of the system
- Easier to rephrase commands that are more closed (e.g. music) than open-ended ones (navigation)
- Discoverability of features and commands was poor; Unsure if they would have discovered it on their own if not asked to perform it in the test

M

Summarized Existing Guidelines

The guidelines reviewed and summarized here varied context application. Most of the suggestions prescribed by the guidelines are general in nature and are distinguished in the following summary by the use of the word “user” in the guideline. The word “driver” is used in place of “user” to distinguish between generalized guidelines and those expressly directed for designing in-vehicle user experiences.

Designing Car Apps

- Use a voice-first approach when designing voice interactions across multiple devices, in order to create a consistent experience for users [1, 14].

- Do not include all features of the mobile app or website in the car-version of the app. Focus on the app’s most prominent features and information to make them easier to use and minimize distraction for the driver [13].

- Disable features that might be distracting to the driver, such as manual texting [13, 30].

- Consider adding car-specific functional to the car version of the app if it is relevant to the context of driving, even if it does not exist in the original mobile app or website [13].

Voice and Manual Input

- Build multi-turn dialogs for beginners and one-shot commands for experts, empowering users to directly get what they want and reduce the amount of time to complete a task.

- Identify variables that a user may include in an utterance to allow for customization and cover a wide variety of use cases [1, 14].

- Expect users to provide more information that is required at a given turn and use the extra information to perform an action or respond faster [1, 14].

- Avoid asking users for complex input, such as alphanumeric passwords [1].
- Provide signposts and suggestions to help users remember the different options to complete an intent or action [1, 14].
- Assume users will reference anything presented visually by voice [1, 14].
- Give users a way to have the voice assistant read out more information [1].
- Support the “Repeat” command in case the user missed a response or prompt. Determine what part of the messages is most relevant and should be repeated by the voice assistant [1].
- Handle multiple ways to trigger an action or intent, including manual touch and varied voice commands [1, 14].
- Research explicit confirmations for high-risk situations to protect against high-consequence failures. Otherwise use implicit or contextual confirmations to convey to users that they were understood [1, 14].
- Allow drivers to complete all tasks with at least one hand on the steering wheel with only brief glances to the screen [13, 30].
- Enable drivers to interrupt an interaction sequence, control the pace of the interaction, and resume an interaction at a logical point in time [13, 30].
- Minimize the required number of touchscreen interactions for drivers [2, 13].
- Make accommodations for low-fidelity touchscreens so that touch gestures are not required to complete basic tasks such as scrolling [13].

General Responses

- Use informal, conversational language the is easily understood by the user [1, 14].
- Vary responses to make the dialog sound more natural, especially in repetitive tasks [1, 14].
- Respond with the top three best matches from an input [1].
- Keep responses informative, relevant, but concise [1, 14].
- Longer responses are typically more difficult to follow and remember [1].
- Use discourse markers to provide context and organize the dialog into chunks

[1, 14].

- Provide responses quickly to minimize interaction between the driver and the system [2, 30].

- Keep responses within one breath at a normal speaking pace as to not overload the user with too much information [1].

- Anticipate and provide information and suggestions that are most important and contextually relevant to the user [1, 14].

- Use prompts to clearly mark when it is the user's turn to speak. People generally answer questions right away, so prompts or rhetorical questions should not be included in the middle of a response [1, 14].

- Provide shorter prompts as users use an intent or action more frequently [3, 1, 14].

- Avoid providing users a correction of their input and respond with the most contextually relevant answer [1].

- Avoid breaking trust with users by including offensive content, unsolicited content like advertisements, unmet expectations, or exposed technical or feature limitations [1].

Situation Awareness

- Keep track of the context of a dialog to understand the use of pronouns and generic references [1, 14].

- Personalize greetings or services with available data, such as the user's current location [1, 14].

- Use interactions with an action or intent to guide subsequent interactions with the action or intent [1].

Presenting Choice

- Present a clear, simple set of options for the user to choose from. Open-ended questions can confuse users or cause them to answer in unexpected ways [1, 14].

- Provide users with less information and few options to help them develop a clear mental map of the action or intent [1, 14].

- Use a narrow-focus question to set user expectations of what they can respond with. For example, “The fan speed is set with a number between one and ten. What speed do you want to set the fan?” [14]

- Use lists to enable users to select one out of many items where those items are most easily differentiated by their title which should be unique and conversation friendly. [14]
- Announce items in a list to make it clear how much information is available [1].
- Keep lists brief and only have essential content about each list item read out [1, 14].
- Give users a way to have the voice assistant read out more options [1].
- Use an either/or question if the list only has two items and avoid presenting a list of one [1, 14].

Error Handling

- Handle errors by minimizing attention to an error and provide a way to get the dialog back on track [1, 14].
- Provide context specific prompts for No Input and No Match errors for every turn in a dialog [14].
- Prompt users to clarify partial or additional information to trigger the correct action or intent [1, 14].
- Re-prompt users with a slight variation on the original prompt if the user does not give a response, adding detail in case the user did not understand the original prompt [1].
- Respond gracefully when data is unavailable, making sure connection problems are handled non-intrusively [2].
- Present errors to driver in the car, not on the connected mobile device [2].

Discoverability

- Provide hints about what a voice assistant can do for a user by leading with benefits. “To get x, do y” [1, 14].
- Suggest alternatives of what a voice assistant can help a user to accomplish [1, 14].

- Provide users with signposts and paths to follow [1, 14].

Display

- Do not display automatically scrolling text, video, or animated images to the driver [13, 30].
- Do not display strings of text longer than 120 character to drivers. Drivers should not be reading long texts when driving [13, 30].
- Avoid visual designs that display information unrelated to the driver's current task. Visual components should not be used for entertainment [2, 13, 30].
- Use visual components in conjunction with voice output to present non-critical, detailed information [1, 14].
- Display most items on screen without the need for the driver to scroll [13].
- Use a common navigational structure across a system, so drivers only need to learn one navigational model [2, 13].
- Minimize the number of touchscreen interactions and screens required to teach each action [2, 13].
- Provide content on the display to drivers, even when the data is unavailable [2].
- Use the visual display as a supplement to voice prompts which should carry the bulk of the interaction with the user. Visual components can be used to provide suggestions to or pivot the conversation [1, 14].
- Use animation only when it would improve a driver's understanding of the system and allow drivers to disable them. When used, animations should be used consistently [2, 13, 30].

Notifications

- Provide succinct and timely information about relevant events from the system or specific apps to the driver [2, 13].
- Prioritize transactional notifications which enable drivers to engage in human-human interaction, function better in daily life, or control or resolve transient device states [13].
- Limit notifications shown to the driver to keep the list length manageable and remove notifications to make room for newer and more relevant content [13].

- Use non-intrusive status messages over notifications to convey low-priority information to the driver [2].

- Make notifications available to the driver on the phone after the drive [13].