The effects of noise-robustness of in-car voice-controlled systems on user perceptions and driving behaviour

By Neil Sokol

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science

Graduate Department of Mechanical and Industrial Engineering

University of Toronto

© Copyright by Neil Sokol 2017

The effects of noise-robustness of in-car voice-controlled systems on user perceptions and driving behaviour

Neil Sokol
Master of Applied Science
Department of Mechanical and Industrial Engineering
University of Toronto
2017

Abstract

In-car voice-controlled infotainment systems are becoming increasingly common in automobiles, but the effects on users when their accuracy degrades in the presence of background noise has not been examined. This thesis compares use of both noise-sensitive and noise-robust simulated voice-controlled infotainment systems under three background noise conditions. It was found that the sensitive system was perceived to be less useful and satisfying even when it performed identically to the robust system. No differences were observed between the systems in several driving performance metrics (with the exception of brake response), but the use of either system impaired driving performance compared to baseline. Glances and subjective workload demonstrated advantages to the robust system. Increased heart rate was observed with the robust system. The results demonstrate that noise robustness is a key factor in user acceptance, and may mitigate visual distraction generated by voice-controlled systems; however, the effects on driving performance are inconclusive.

Acknowledgements

First, I would like to thank my supervisor Prof. Birsen Donmez for her assistance and guidance with this thesis. I would also like to express my thanks to Dr. Winnie Chen for her assistance with developing this study, and her helpful comments on this document. I would also like to thank both Mitacs and Qualcomm for funding and technical assistance on this thesis.

I would like to thank my committee members, Dr. Greg Jamieson and Dr. Olivier St-Cyr for their comments and help in developing this work.

I would like to express my gratitude to my colleagues in the Human Factors and Applied Statistics lab for their support throughout my masters. I could never have done this without you folks. I'd like to give a special thanks to my summer undergraduate research assistant Nima Mirjalali for his invaluable assistance with developing the experiment, and recruiting and running participants. The background noise samples in particular would not have been what they were without his expertise.

Finally I would like to thank my family and friends for their unwavering support throughout the last two years. To my loving girlfriend Ally who has been so understanding throughout all the late nights writing this thesis. Thanks to my parents and grandparents whose unceasing love and encouragement has helped me become the first person in the family to get a masters. I'd like to thank my roommate Natalie for allowing me to take over half the apartment with multiple laptops and textbooks while I did data analysis, as well as the rest of my friends for putting up with me constantly complaining about driving simulators, eye tracking systems, and weird statistical problems for the past two years.

Table of Contents

List of Tables

List of Figures

Chapter 1

1.0 Introduction

Distracted driving is a major factor in automobile crashes, and has been estimated to be a contributing factor in at least 68% of the 905 injury/property-damage crashes observed in the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) (Dingus et al., 2016). Driver distraction has been defined in a number of ways, but one concise definition is the misallocation of attention from driving to a non-driving task or source of information (Smiley, 2005). Use of various ancillary technologies while driving can be a key distractor for drivers, but not all non-driving activities have equally detrimental effects on driving.

According to the Multiple Resource Theory (Wickens, 1984), a person has multiple ‘pools’ of information processing that can be accessed simultaneously while completing a task (e.g., visual, auditory). Different tasks do not necessarily use the same resource pools or might have to compete for resources if they do. Further, increased workload does not necessarily result in decreased performance, nor does reduced workload always result in increased performance, such as boredom resulting in a loss of attention (Hart, 2010). However, task performance can decrease in the face of a shortage of one or more of these resources.

Visual-manual secondary tasks in particular have been found to increase crash risk significantly as these tasks claim resources that are central to the driving task itself. This risk has been demonstrated in analyses of recent naturalistic driving studies such as the aforementioned analysis reported by Dingus et al. (2016) on the data from SHRP2 NDS. A naturalistic driving study involves outfitting a large number of participant vehicles with unobtrusive sensors and cameras, and then recording all of their day to day drives (European Naturalistic Driving Study, n.d.). This type of study allows for real life insights into driver behaviour that cannot easily be observed in a controlled laboratory experiment. SHRP2 NDS is the largest naturalistic driving study completed to date, capturing naturalistic driving data from over 3,500 participants in six locations across the United States over the course of a three year period. Data was collected for every trip taken in a participant's vehicle from the time the key was turned to start the vehicle to the time the key turned to stop the vehicle. This study has provided a wealth of information on what behaviours and interactions influence driver distraction.

For example, Dingus et al. (2016) found that texting on a cell phone increased crash risk by a factor of 6.1. Auditory-vocal tasks on the other hand, while still increasing crash risk appear to do so to a lesser degree; Dingus et al. (2016) found that talking on a handheld cell phone instead of texting increased crash risk by a factor of 2.2. As mentioned earlier visual-manual secondary tasks compete for resources central to the driving task as driving itself is primarily a visual-manual task. An additional visual-manual task, such as texting on a cell phone, would infringe on the same resources that the task of driving requires. Although auditory-vocal interactions such as talking on the cell phone would also claim attentional resources, they would largely tap into the separate auditory-vocal resources, and not the same primary input-output modalities used by the driving task.

Carried-in devices, such as cell phones, are not the only form of technology-related distraction in cars. Automobile manufacturers have been clamouring to include a variety of new and interesting technologies in their automobiles in order to differentiate themselves from the competition, and one such example is in-vehicle infotainment systems. These systems allow users to perform functions unrelated to the main task of vehicle control, mainly navigation (e.g., finding directions using built-in GPS) and entertainment (e.g., selecting music to be played on the car's stereo). The analysis of Dingus et al. (2016) revealed that the use of in-vehicle information systems (a superset of infotainment systems that also include driver assist), while at a relatively low prevalence of use during baseline (or "model") driving (occurred at least once during 0.83% of such driving segments), resulted in a risk of crash 4.6 times higher than during baseline driving. While not explicitly stated in the study, it is implied that the systems involved required touch-screen interaction and therefore significant levels of visual-manual interaction. The use of voice-command is expected to alleviate the potentially dangerous levels of demand placed on drivers by these visual-manual interactions in these infotainment systems.

It should however be noted that voice-command does not entirely eliminate visual demand as most of the production level systems also provide a visual interface to the drivers. While commands can be given vocally, these systems often require drivers to glance to a dashboard-mounted display as information and feedback are often provided visually. Studies have shown that this visual demand is still substantial even when attempts are made to streamline tasks by reducing the number of interactions required (Reimer et al., 2014). In addition, drivers have been observed to make "orienting responses" when engaging with these production-level systems (Reimer et al., 2013). An orienting response is when the driver looks or turns towards the system's display during a command as if the voice-controlled system is located inside that display. These orienting responses have been found to be more common among older participants and when the system required longer interactions. Thus, visual engagement may occur not only when users receive information from voice-controlled infotainment systems, but also when the driver issues commands.

While prevalence of use may be relatively low now, infotainment systems in general and voice-controlled systems in particular are becoming increasingly common in modern automobiles, with one industry estimate suggesting voice-controlled systems will be present in at least half of new cars by 2019 (IHS Technology, 2013). It has also been suggested that infotainment systems, especially voice-controlled ones, are becoming increasingly liked and viewed as useful by drivers (Consumer Reports, 2016). With their increasing adoption in mind, the effect these systems have on drivers is of increasing interest. Several studies have compared voice-controlled systems to traditional manual interaction (Maciej & Vollrath, 2009; Ranney, Harbluk, & Noy, 2005; Shutko, Mayer, Laansoo, & Tijerina, 2009). While these studies all found advantages compared to manual interaction, especially in terms of improved lane-keeping measures and lowered driver distraction, voice-controlled infotainment systems must contend with a number of environmental conditions in on-road use which can degrade their performance. One prominent factor is noise. Interfering noise can affect the ability of the system to accurately recognize the voice commands of the user. Ambient road noise, noisy passenger conversations, music playing on the stereo, engine noise; all of these can degrade the performance of voice-controlled systems. Studies have been carried out which demonstrate that accuracy reduction of voice-controlled systems can affect driving performance (Gellatly & Dingus, 1998; Kun, Paek, & Medenica, 2007; McCallum, Campbell, & Richman, 2004). Thus, the advantages conferred by voice-controlled systems may be reduced by the real-world environment in which they must operate.

In addition to the effects on driving performance, how user perception is affected by changes in the accuracy of voice-controlled systems is also of interest. With regards to automation in general, previous research by Lee & See (2004) has proposed that when automation fails due to environmental factors beyond its control (e.g., environmental noise), the failure can still result in users distrusting and disusing the system. The development of a voice-controlled system that is disused by users would be a waste of resources for automobile manufacturers. More importantly, as various voice-controlled in-car infotainment systems have the ability to also be operated via manual interaction, user distrust in voice-controlled systems may cause those users to switch to manual control, rendering the potential safety advantages offered by voice-control ineffective.

Driving can already be a mentally demanding task on its own, and increased mental load due to additional tasks has been shown to affect driving performance (Patten, Kircher, Östlund, & Nilsson, 2004; Waugh et al., 2000). As mentioned previously, the task of driving is primarily visual-manual. However, voice-control can still claim cognitive resources that may interfere with the performance of the driving task, and accuracy degradation may further exacerbate this effect. Therefore, in addition to visual demand, user perceptions, and driving performance, how the use of in-car voice-controlled infotainment systems affects changes in mental workload is also an area of interest.

Presently, new technologies are being developed which may help reduce or eliminate the recognition accuracy variability within production level voice-controlled systems. One such technology aims to eliminate accuracy issues that may arise from background noise. While the effects of in-car voice-controlled accuracy have been studied before (discussed in more detail in chapter 2), there are still many gaps in the literature. For example, earlier studies manipulated accuracy in a vacuum, i.e., no external environmental condition changed which the participant could attribute the accuracy change to (e.g., noise) (Kun et al., 2007; McCallum et al., 2004). Whether driver experience is affected by accuracy changes being linked to environmental changes is still an open question. Studies on acceptance of in-vehicle voice-controlled infotainment systems are rare, and of the ones that exist none to our knowledge examined the effects of voice-controlled recognition accuracy on acceptance.

An analysis of the effects a voice-controlled in-vehicle systems has on workload has been performed in previous studies. However, changes in system accuracy either were not included as with Reimer et al. (2013), or included but not used as a factor in the workload analysis as was the case in Kun et al. (2007). While studies such as Reimer et al. (2013) have looked at the effects voice-controlled infotainment systems have on visual demand and glance behaviour, no changes in accuracy occurred, and the aforementioned studies that did modify system recognition accuracy did not examine visual demand or driver's glances. Finally, to the best of our knowledge, no study has evaluated how a noise-robust system, that is to say a system whose accuracy is unaffected by background noise changes, compares to a noise-sensitive one, that is to say a system whose accuracy is affected by background noise changes (in terms of driving performance, driver distraction, cognitive workload, and system acceptance) in any of the aforementioned measures. To investigate this apparent research gap, a literature review was undertaken to determine the current state of knowledge, and then based on this review an experiment was developed to investigate several gaps discovered in the knowledge. The remainder of this thesis outlines this review, the experiment, and the results and implications of the experiment.

Chapter 2

2.0 Literature Review

This chapter outlines a review of the relevant literature. This review was undertaken to investigate the current state of knowledge on the effects of in-vehicle infotainment systems on driving performance, visual distraction, user acceptance, and cognitive workload. The effects of changing system accuracy on these outcomes were also reviewed.

2.1 The effects of voice-controlled in-vehicle infotainment systems on driving performance and visual distraction

A number of prior studies have evaluated the effects of in-vehicle voice-controlled systems, including infotainment systems, on driving performance and visual distraction. These evaluations were largely performed by comparing voice-controlled systems to manually-controlled systems with equivalent functions. In one closed-track on-road study, participants were asked to use a voice-controlled system as well as a manually-controlled one to perform both a simple (i.e., open message list and make a voice memo) and a complex task (i.e., use address book and phone system to retrieve info, make voice memo) while following a lead vehicle (Ranney et al., 2005). The voice-controlled system provided auditory feedback, and the manually-controlled system provided visual feedback. Concurrently, participants were asked to complete a peripheral detection task in which they were asked to respond when LEDs arrayed in the lower left periphery of their line of sight activated. The detection task was included to simulate the visual search requirements of normal driving. Two groups of participants were used: group 1 drivers had no special driving skills, while group 2 drivers were all staff engineers at the National Highway Traffic Safety Administration’s Vehicle Research and Test Center. The authors aimed to determine if voice interactions affected the amount of peripheral and attentional interference, as defined by Strayer and Johnston (2001), caused by the secondary tasks. Peripheral interference suggests a conflict between the visual-manual demands of the task and the visual-manual demands of driving, while the attentional interference hypothesis attributes performance issues to “attentional demand generative components of processing” (Strayer & Johnston, 2001, p.466).

It was found that participants had a higher detection rate when using the voice interface (M = 0.72) than the manual interface (M = 0.62) and a faster response time using the voice interface (M= 0.80 s) than the manual interface (M = 0.80 s) when performing the peripheral detection task. In addition, their lane-keeping (i.e., standard deviation of lane position) was better during the use of the voice interface (M = 20.2 cm) compared to manual interaction (M = 22.3 cm). Lead vehicle following distance and steering reversal rate were not significantly different between the two interfaces. Further, use of either system resulted in a worse detection rate and a higher response time in the peripheral detection task compared to a baseline drive with no secondary task (M = 0.88 and 0.55 s, respectively). In addition, both interfaces resulted in a longer lead vehicle following distance and a higher steering reversal rate compared to the baseline. Complexity of the task did not have a significant impact on any of the above measures.

The group 1 (less experienced) drivers experienced more pronounced difficulties in lane-keeping than group 2 (expert) drivers. The authors note that in general, group 1 drivers had a more difficult time in managing the combination of car following, target detection, and secondary tasks. However, they note that the worse performance results could also be due to group 2’s generally higher experience with technology in addition to their driving experience. In summary, the authors concluded that a voice interface did result in less peripheral impairment compared to a manual interface. However, that same voice interface resulted in no reduction to attentional impairment, as defined by Strayer and Johnston (2001), when compared with manual interaction.

A simulator study carried out by Shutko et al. (2009) found similar advantages as Ranney et al. (2005) to voice interaction through the comparison of a voice-controlled infotainment system to manual cell phone use. Participants were asked to undertake a series of tasks while driving using both hands-free voice-controlled interaction with Ford’s SYNC infotainment system, and hand-held manual interaction with the same cellphone. Feedback from the voice-controlled system consisted of auditory responses and limited visual feedback. Task performance was evaluated through task completion time. Driving performance was examined through standard deviation of lane position, percentage of drives with at least one lane exceedance, and difference in vehicle speed during task performance (maximum speed – minimum speed over the period tasks were performed). Driver attention was assessed through total eyes off-road time and response to a pedestrian detection task in which a simulated pedestrian appeared on the highway shoulder ahead of the driver after a secondary task had started. Similar to the result of Ranney et al. (2005), standard deviation of lane position was found to decrease with use of the voice-controlled system compared to manual interaction; the reduction found in this study (estimated difference = 0.124 m less for voice-controlled) was an order of magnitude greater than the effect observed by Ranney et al. (2005). The difference might be due to the visual feedback provided by the voice-controlled system, or to the differences between real-world and simulator driving.

Use of the voice-controlled system also resulted in less eyes-off road time per task (estimated difference = 13.7s less for voice-controlled), with the exception of receiving an incoming call (estimated difference = 1.0 s less for manual) as well as a shorter total task completion time in all cases (estimated difference = 15.9 s less for voice-controlled), except dialing a phone number (estimated difference = 17.7 s less for manual). Manual interaction also resulted in a higher pedestrian detection response time for the phone dial (median difference = 0.25 s higher for manual), artist search (median difference = 0.25 s higher for manual), and text message review tasks (median difference = 0.35 s higher for manual). Thus, while advantages were identified in using a voice-controlled system with regards to reduced visual distraction, not all tasks benefitted equally. The tasks that had the strongest visual-manual component: song search, artist search, contact search, and text message review/reply benefitted the most from the voice-controlled interactions.

Similar to the above two studies, Maeicj and Vollrath (2009) found that voice-controlled systems held advantages with regards to lane-keeping and visual distraction compared to manually-controlled systems. The authors undertook a simulator study evaluating the effects of four kinds of voice-controlled infotainment systems on gaze behaviour and perceived distraction, as well as driving performance. These voice-controlled systems provided information through a mix of auditory and visual responses. Driving performance was assessed by mean deviation and standard deviation of lane position, as well as the participant’s reaction time to the lane change task. In the lane change task, the simulation periodically displayed road signs instructing the participant to change to a specific lane and they had to do so as soon as they noticed the sign. Their reaction time was measured from when the sign became legible to when they commenced the lane change. The voice-controlled systems were compared to both manual interaction and baseline driving with no secondary task. It was found that voice-control resulted in lower standard deviation of lane position and lower lane change reaction time compared to manual interaction. Off-road gaze duration was significantly reduced during voice-control, and participants' subjective distraction ratings were lower for voice-control compared to manual. However, compared to baseline driving, voice-control lane-keeping and reaction time were still inferior. As with Ranney et al. (2005), while advantages were found for voice-controlled interaction compared to manual control, use of voice-controlled systems was still found to result in inferior driving performance and significant attentional impairment compared to driving with no task.

A pattern can be seen across the three studies detailed in this section as well as the additional studies that will be further discussed in the following sections, i.e., Reimer et al. (2013) and McCalum et al. (2004). The use of voice-control resulted in advantages over manual control, particularly in the ability of drivers to maintain their lane positions and in reducing their off-road glances. However, even with the use of a voice-controlled system, driving performance still is impaired compared to driving with no secondary task. The reason could be the distraction still being present, and this distraction being more pronounced with higher levels of visual feedback required to complete the secondary task. It should be noted that none of the three studies discussed the accuracy or effect of accuracy of the voice-controlled systems they used. However, as all three studies used off the shelf systems and none discuss fixing their system's accuracy, it is a good possibility that these voice-controlled systems did not have

100% voice-recognition accuracy. If effects due to accuracy changes were present, they were not explicitly examined in the aforementioned papers.

2.2 User acceptance of voice-controlled in-vehicle infotainment systems

A user's perceptions of a system, specifically acceptance, can be valuable predictors of the intention to use. There are several methods of measuring technology acceptance, such as Davis' Technology Acceptance Model (TAM), Lowry et al.'s (2012) Hedonic-Motivation System Adoption Model, or Scherer's (1986) Matching Person Technology Model. The latter two have been adopted in more specific domains, e.g., 'hedonistic activities' such as online games and for assistive technologies for the disabled, whereas TAM has been adopted for a wider range of applications. The original Technology Acceptance Model (TAM) proposed by Davis (1989), describes two main characteristics that drive user acceptance. The first is perceived usefulness, which is the extent that users believe a technology will enable them to perform their job better. The second is perceived ease of use, which is the extent that users believe a technology is free from effort. Later attempts have been made to extend this model, such as by breaking down perceived usefulness into social factors and cognitive instrumental processes (Venkatesh & Davis, 2000). However, these extensions still stem from the same two core constructs of the original model. Davis' TAM is not without its critics, with arguments that it does not consider emotional, group, or cultural factors in technology use, as well as the role of self-regulatory processes (Bagozzi, 2007). It has also been suggested that existing studies showing the usefulness of TAM have been narrowly focused on office automation (Legris, Ingham, & Collerette, 2003). Similar models, while not explicitly extended from TAM, exist for the driving domain, which utilize similar constructs. With a focus on in-car technologies specifically, Van Der Laan & De Waard (1997) developed a simple model by selecting a number of existing scales which allowed for evaluating two dimensions of acceptance: perceived usefulness (as with Davis) and how satisfying users find an in-vehicle technology.

User acceptance of voice-controlled in-vehicle infotainment systems has rarely been a focus of study. A handful of acceptance studies relating to infotainment or systems of a similar function were found. In order to identify unique factors that might influence the acceptance of in-vehicle navigational systems, Park and Kim (2014) attempted to extend TAM through a survey study. The authors investigated perceived locational accuracy of the system, system service and display quality, perceived processing speed, attitude, and satisfaction. These factors were selected based on interviews conducted with two groups: expert drivers, and engineers and marketers from the car navigation industry. The authors found that perceived processing speed and perceived locational accuracy played key roles in user's satisfaction, which itself was significantly related to the intention to use the system. The authors concluded that the influence of technical issues, such as locational accuracy, on user acceptance requires further investigation.

In a survey study, Keuntje and Poormohammadroohafza (2014) examined how app-based infotainment systems (such as Apple CarPlay or Android Auto) are perceived by drivers and what factors influence their acceptance. Participants were first described the functions and capabilities of a typical app-based infotainment system and then were shown a demonstration video of an Apple CarPlay based infotainment system. They were then asked a number of questions on a five-point Likert scale designed to assess their acceptance of app-based infotainment systems in six categories: perceived usefulness, perceived ease of use, perceived task-technology-fit, perceived risk, perceived costs, and behavioural intention to use. Participants were also afforded the opportunity to give open-ended feedback on their views on voice-controlled infotainment systems.

The close-ended survey results showed that users largely perceived in-vehicle infotainment systems in general to be useful and easy to use, but that participants also strongly perceived risks to using them while driving, as well as being concerned about the costs of such systems. Perceived usefulness and perceived task-technology-fit were found to positively influence acceptance, while perceived risk and perceived cost negatively influenced acceptance. Behavioral intention to use responses indicated that participants were uncertain if they would actually use app-based infotainment systems. A further breakdown of the intention to use questions showed that participants would be more likely to experiment with such a system if they happened to have one in their vehicle, but that they were hesitant to spend any additional money on one when purchasing a vehicle. Open-ended responses by participants were largely concerned with risk. The statements on risk revealed that participants were split on whether they felt the systems were a risk to driver safety or whether they believed the systems offered a potential safety improvement compared to cell phone use. The authors concluded that participants placed a high value on, and had high expectations with regards to, usefulness. Participants also perceived a high level of ease of use with current infotainment systems, and the authors concluded that infotainment system designs that further improve ease of use may reduce the perceived risk associated with the use of these systems. A limitation with this study was that 73% of the respondents were male, and thus the results may not have been fully representative of the opinions of the general population.

While there appears to be no other studies specifically on in-vehicle infotainment system acceptance, studies have been performed on acceptance of other forms of in-vehicle systems (Donmez, Boyle, & Lee, 2007; van Driel, Hoedemaeker, & van Arem, 2007; Waard & Brookhuis, 1997). Many of these studies focus on various forms of driver assistance systems, and are not necessarily voice-controlled. For example, Donmez et al. (2007) investigated different distraction mitigation features for in-vehicle information systems. Participants interacted with an in-vehicle system with adaptive distraction mitigation features that informed them via visual feedback if during interaction they were glancing away from the road for an unsafe amount of time. User acceptance was assessed through the system acceptance questionnaire developed by Van der Laan et al. (1997) mentioned earlier.

While studies on acceptance of other forms of in-vehicle information systems have some relevance, there are key differences that may arise between user interactions with driver assistance systems and user interactions with infotainment systems, which are the focus of this thesis. Infotainment interaction is largely voluntary, with the user deciding when, where, and how to initiate an interaction. A driver may temporarily put off their interaction or structure it around the driving task. An alert from a driver assistance system, while also an in-vehicle information system, is not so voluntary. An alert arising from a distraction mitigation system or a lane-departure warning system is at the discretion of the system, not the user.

2.3 Voice-controlled in-vehicle infotainment system use and cognitive workload

Increased cognitive workload caused by the performance of a task in addition to the driving task has been shown to have an effect on driver behaviour and performance. Reimer et al. (2013) examined the effect that in-vehicle infotainment systems have on cognitive workload.

This report detailed an on-road study in which participants were asked to undertake a series of tasks with Ford's SYNC Infotainment system using both manual and voice input while driving in highway conditions. The SYNC system was a more advanced version of the system used by Shutko et al. (2009), and in this case, the visual responses given were displayed on a larger, more complex LCD touchscreen display. Subjective and physiological workload, glance behaviour, and driving performance were examined. The tasks consisted of two manual radio tuning tasks (easy and hard), two voice-controlled radio tuning tasks (easy and hard), voice-controlled navigation entry, voice-controlled contact dialing, and voice-controlled song selection. One song selection task was designed to be impossible so that participants would believe the system's voice recognizer did not recognize their request. N-back tasks (0-, 1-, and 2-back) were used as a reference to help scale workload (Kirchner, 1958). For the n-back tasks, the participants were read a series of numbers, and had to verbally respond with the number that was presented to them n-steps prior. The higher the n-value, the higher the load on working memory and therefore the higher the workload. Heart rate assessed via electrocardiogram (ECG), skin conductance, and subjective workload assessed on a scale of 0-10 were collected for all tasks.

Voice tasks resulted in a significantly lower major steering wheel reversal rate compared to manual tasks. A reduction in average velocity compared to the baseline was observed for all tasks and was interpreted as compensatory behaviours by drivers undertaken in order to reduce their workload. Voice tasks were found to result in a lower change in heart rate from the baseline heart rate (Mean change of 2.05%) than the difficult manual radio task (Mean change of 3.60%), as well as the 1-back (Mean change of 4.54%) and 2-back (Mean change of 8.87%) reference tasks. For subjective workload, the only voice task in which participants considered workload to be higher than the difficult manual radio task (mean score of 2.48) was the song search designed to be impossible to complete (mean score of 6.56). Otherwise, in terms of perceived workload, voice tasks fell between the easy manual radio task (mean score of 1.89) and the difficult manual radio task. The authors noted that heart rate aligned reasonably well with subjective workload. As levels of workload measured through physiological signals for all tasks were below that of the 1-back task, the authors concluded that the level of physiological arousal measured for the 1-back task may be a useful reference point for future studies.

Another investigation of the subjective effects of workload on voice-controlled infotainment systems was the study Wu et al. (2015). In this experiment, participants drove their own vehicles and interacted with their own in-vehicle voice-controlled infotainment systems (which gave responses in a mix of auditory and visual feedback) while being asked contextual questions about their experience. Subjective workload was assessed using the NASA-TLX scale (Hart & Staveland, 1988). While driving, participants were asked to undertake four different kinds of tasks with the voice-controlled systems: communication, information, entertainment, and navigation. Participants' voice-controlled systems were categorized into three: smartphones with voice-controlled abilities mounted somewhere in the vehicle, OEM-installed systems integrated into the vehicle by the manufacturer, or a hybrid system in which the interactions were with the vehicle system but a smartphone was providing the functionality in the background. Workload analysis was conducted on navigation tasks, which were the type of tasks deemed most difficult by the authors. The participants found the OEM installed systems to not have the same level of sophistication as smartphone based systems, and the associated increase in task complexity resulted in increased workload, effort, and frustration.

Participants also felt that voice-controlled tasks were impacting their workload even though the systems did not require them to use their hands. An additional interesting finding was that when participants had difficulties interacting with their voice-controlled systems, they did not necessarily feel the voice-controlled system was at fault for any errors or long interactions that resulted.

Both heart rate and Galvanic Skin Response (GSR) have been shown to be effective in measuring changes in cognitive workload in driving studies. In a simulator study, Mehler et al. (2009) examined changes in these measures when participants were required to complete n-back tasks while driving. As mentioned earlier, the n-back task is an established method of increasing cognitive workload (Reimer et al., 2013). Significant main effects were found for both heart rate and GSR. However, it was noted that there was a threshold to these measures responding to workload increases at the higher workload levels. Specifically, between the 1-back and 2-back tasks, heart rate increased by only an average of 1.1 beats per minute. For GSR, the change between 1-back and 2-back was so small as to not be statistically significant.

The Reimer et al. (2013) study also noted some interesting non-workload related negatives to voice-controlled interactions. It was found that one voice task, voice-controlled navigation entry, resulted in off-road glance times (mean = 32.7s) longer than the National Highway Transport Safety Administration distraction guidelines on in-vehicle electronic devices in which secondary non-driving visual manual tasks should not take more than a cumulative 12 seconds to complete (National Highway Transport Safety Administration, 2012). However, all voice-controlled tasks resulted in a much longer task completion time than any manual task. For example, for the hard radio tuning task, using the voice-controlled system took 23 seconds longer than manual control. In summary, voice-controlled systems have been found to have potential advantages over manual-interaction in terms of workload. Both subjective and physiological measures of workload have been found to be lower during voice-controlled system use. However, these measures show that workload during voice-interaction is still significant, and that interviewed users recognize that these systems still have appreciable workload.

2.4 Accuracy effects of voice-controlled in-vehicle infotainment systems

Very few studies have been conducted on the effects of voice-controlled system accuracy on driving performance, user acceptance, cognitive workload, or visual distraction. In a review of the literature, only two simulator studies and a single on-road study could be found. While all three studies were on voice-controlled in-vehicle systems, not all of these systems were infotainment systems.

Kun et al. (2007) undertook a simulator study to investigate the effects of reduced voice recognition accuracy on driving performance. Participants performed a series of secondary tasks with a voice-controlled system while following a lead vehicle. The system provided auditory responses. The secondary tasks involved participants using commands to patch police messages between simulated police vehicles and a simulated headquarters. For example, they might have received a request to retransmit a message on channel A in city B, and then they had to do so using a command to select the channel, another to select the city, and a third to confirm the patch of the message. The voice-controlled system had four conditions which participants experienced in one continuous drive: high (89%) and low (44%) accuracy, combined with triggering the interaction with either voice prompts or a push-to-talk button.

The push-to-talk button was placed on the center console and necessitated removing a hand from the steering wheel to operate. Voice-controlled system accuracy was set by having participants use pre-selected phrases to interact with the system, and then having the system respond correctly the desired percentage of times.

In terms of driving performance, it was found that reductions in recognition accuracy slightly increased steering wheel variance, but did not affect average speed or lane keeping. There was one exception. When the system was in low accuracy mode, push-to-talk activation of the system resulted in higher lane position variance compared to vocal activation. A possible explanation provided by the authors was that the frustration with the system's low accuracy resulted in participants repeatedly and forcefully depressing the push-to-talk button. The authors measured frustration using a five-point Likert scale question asking participants whether they were frustrated with the speech interaction, and found that most responded "Somewhat" for low accuracy and "Not at all" for high accuracy. While this study was a useful examination of the effects of accuracy of in-vehicle systems, the secondary task was not representative of infotainment system tasks.

McCallum et al. (2004) investigated the effects of varying voice-control accuracy, and voice interaction vs. manual interaction on driving performance and cognitive workload through a Wizard of Oz simulator study. Wizard of Oz refers to a paradigm in which users believe they are interacting with a computer system which understands their speech, but in fact all system responses are being manipulated by the experimenter, the proverbial "man behind the curtain" (Kelley, 1984). While driving, participants interacted through voice-control with a personal digital assistant (PDA) with three accuracy levels: low (58%), medium (78%) and high

(100%). They also interacted with the PDA manually, and also performed a baseline drive with no secondary tasks. Feedback from the voice-controlled system was visual in nature. The exact tasks participants were asked to perform were not detailed in the paper, but the authors state that the PDA was set up to accommodate making phone calls through a contact list, reading emails, finding transit schedules and traffic conditions, and displaying entertainment in the form of movies. These tasks would be consistent with those of in-vehicle infotainment systems. During each drive, an emergency event was presented to the participants in the form of a car or pedestrian pulling out in front of the participant vehicle. Driving performance was measured by participants' reaction time to these events, as well as the number of collisions that occurred. Cognitive workload was measured through a modified NASA TLX questionnaire (S. G. Hart & Staveland, 1988). The exact nature of the authors' modifications was not specified in the study.

The voice-controlled system resulted in significantly higher collision frequency under the low accuracy condition (5 collisions) than medium (2 collisions) and high accuracy (2 collisions) conditions. Voice-control resulted in a lower collision frequency (9 collisions) than manual control (12 collisions), at a rate that was comparable to the no secondary task condition (9 collisions). Manual control was found to result in a significantly higher reaction time (1.96 s) than the no secondary task condition (1.52 s). Voice control resulted in a reaction time (1.73 s) that fell between manual control and no secondary task, but was not statistically significantly different than either. Similarly, in terms of cognitive workload, manual control resulted in higher subjective workload (35.8) than no secondary tasks (25.1), and while voice control resulted in a workload rating (28.9) that fell between the other two conditions, there were no statistically significant differences. The study was limited in that it was a proof-of-concept study and not all collected data was analyzed with all conditions. For example, the effect of changing accuracy was only investigated in terms of what effect accuracy had on collision frequency. No comparisons were made between different accuracy levels in terms of perceived workload and reaction time.

As mentioned earlier, to the best of our knowledge, only a single on-road study investigated the effects of in-vehicle voice-controlled system accuracy. Gellatly and Dingus (1998) undertook an on-road driving study to determine the effects of varying voice-controlled system accuracy levels on driver performance and secondary task performance. The study compared four voice-controlled system accuracy levels of 100%, 90%, 75%, and 60%. In addition, a control condition of manual input was also included. The voice-controlled conditions were created using a Wizard of Oz methodology. The voice-controlled system returned feedback in visual form. While not all secondary tasks could be considered infotainment tasks (e.g., unlocking the car door with a voice command), several tasks could be (e.g., dialing a stored phone number or searching radio stations). Two types of voice-recognition errors were also compared: rejection errors in which the system does not understand the input at all, and substitution errors in which it recognizes input but not the correct phrase. Statistically significant differences in peak lateral acceleration between voice-recognition and manual interaction were found only at the 60% recognition accuracy level. However, these differences were too small to be considered as cues for unsafe driving behaviours. The authors concluded that since driving performance wasn't affected until the 60% recognition accuracy level, imperfect voice-recognition performance is acceptable with regards to driving safety. In terms of participants' task performance, when voice-controlled system failures appeared in the form of substitution errors, there was a significant increase in task completion time whenever system accuracy decreased. This increase did not exist when the errors used were rejection errors; in those cases, there were similar task completion times across all accuracy levels.

Aside from these three studies, we were not able to find any other studies focusing on the effects of voice-controlled in-vehicle system accuracy on cognitive workload, driver distraction, or driving performance. Further, no prior studies could be found on the effects of voice-controlled in-vehicle infotainment system accuracy on user acceptance. However, there has been research in other domains on how voice-controlled system accuracy affects user perceptions. In one study, it was found that speech recognition accuracy had a significant effect on whether users found a voice-controlled word processor acceptable to use (Casali, Williges, & Dryden, 1990). Even when methods to correct voice recognition errors were easy to use, interrupting the task to use them was considered "disturbing". Rebman, Aiken & Cegielski (2003) carried out a study comparing voice-controlled text input to keyboard input in terms of user perceptions. Participants were asked to input text in two different ways using both voice entry and keyboard entry. The authors required participants to, (1) input text as fast as possible without correcting errors, and (2) correct errors as they typed. Regardless of typing skill, participants rated the speech recognition system worse than typing in terms of ease of use and efficiency. The speech recognition system's poor accuracy was believed to be the main reason for the worse rating, as it caused user frustration.

2.5 Research gaps & experimental objective

A number of patterns appeared across the prior literature. Voice-controlled infotainment systems hold advantages in terms of improved driving performance and reduced visual distraction compared to manual input, but these advantages do not bring performance up to the level of baseline driving with no secondary task. Voice-controlled infotainment systems also demonstrate advantages in terms of cognitive workload. Despite these informative findings, there are still important gaps in the literature. While a limited number of studies have looked at the effects of accuracy in in-car voice-controlled systems on driving performance, no context accompanied these accuracy changes. An example of such context would be a change in background noise conditions. How voice-controlled accuracy affected cognitive workload was also not investigated in any of the prior work reviewed. In addition, how accuracy affects user acceptance of voice-controlled systems does not appear to be investigated at all up to this point.

The impact of context for automation failures on performance has precedence in prior literature. Bagheri and Jamieson (2004) compared performance in a multi-task flight simulation with an automated system-monitoring task component in two conditions: where participants were given context about the automation's reliability and how it may fail, and when they were not given any information on the automation's reliability. The authors found that the addition of context significantly improved task performance, and it was believed this was due to context allowing participants to better allocate their attention across the multiple tasks.

As mentioned in the introduction, new systems are being developed aimed at significantly improving the noise-robustness, and therefore the recognition accuracy, of in-car voice-controlled systems. Given that no prior work investigated the effectiveness of such noise-robust systems, this thesis focused on comparing a noise-robust voice-controlled system to a noise-sensitive one in a driving simulator study. Through the noise sensitivity/robustness lens, this study improved on prior studies on accuracy that did not provide any environmental context that could explain to the users the degradations in system accuracy. It was hoped that this context would allow users to better calibrate their attentional allocation as seen in prior work (Bagheri & Jamieson, 2004). It must be noted that some differences are present between this work and Bagheri and Jamieson. The nature of the context provided by background noise requires users to know that voice-controlled systems can fail in the presence of loud disruptive noise, compared to the context provided by Bagheri and Jamieson in which users were explicitly told the automation may fail. However, it was hoped that user's prior experiences with voice-controlled technologies would allow them to implicitly make this connection through the context provided.

In addition to driving performance and visual distraction, particular interest was placed on user acceptance of the systems, as the review of the literature showed that user acceptance of in-car voice-controlled systems has not been extensively studied to this point. Further, as mentioned previously, understanding the factors underlying the acceptance of in-car voice-controlled systems is key to designing systems that would support the adoption of voice-control instead of the use of the less safe manual control that is also an option. Cognitive workload was also of interest, as while prior studies, such a McCallum et al. (2004), have shown voice control to reduce workload compared to manual interaction, little could be found on how different accuracy levels affect workload while using a voice-controlled system.

The objective of this thesis was to carry out an experiment to understand how a noise-robust voice-controlled infotainment system compares to a noise-sensitive one in terms of user perception, workload, and driving performance. A baseline condition with no system was also included as a reference for driving performance with no secondary task. The hypotheses were:

H1: A noise-robust system would result in higher acceptance by users than a noise-sensitive system as more consistent accuracy will reduce frustration and reduce required error correction.

H2: A noise-robust system would result in lower cognitive workload than a noise-sensitive system due to less error correction and less concentration being required to use the more reliable system.

H3: A noise-robust system would result in a lower level of visual distraction than a noise-sensitive system, as less visual interaction will be needed to confirm if the system was operating correctly when using a robust system.

H4: A noise-robust system would result in safer driving performance than a noise-sensitive system, but performance would still be worse than baseline driving as suggested by Shutko et al. (2009), Ranney et al. (2005), and Maciej and Volrath (2009).

Chapter 3

3.0 Methods

A driving simulator study provides the ideal environment for studying the hypotheses outlined in Section 2.5. A simulator environment allows for tightly controlled noise conditions. Given that we were curious about the influence of background noise, the simulator setting allowed us to set the exact volume and nature of ambient noise without the occurrence of the typical random environmental noise one would see in a real-world driving environment. We also could ensure that the driving environment itself was controlled and safe; e.g., there were no adverse weather conditions and we could repeat different traffic events (e.g., lead vehicle braking) consistently from one participant to the next. Finally, controlled lighting conditions allowed for more accurate eye tracking using the FaceLab eye tracking system possessed by the laboratory. Driving simulator studies do have some potential drawbacks, and these are discussed in the Discussion section of this thesis (specifically Section 5.2).

In the experiment, participants were asked to drive in a simulated urban environment while performing various infotainment tasks with an attached voice-controlled system. The purpose of this experiment was to determine the effect of in-car infotainment voice-controlled system accuracy on driving performance, as well as if accuracy (degraded by background noise) was a major factor in the user acceptance of these systems. As accuracy levels needed to be fixed, this study was a Wizard of Oz study. The method used is referred to as Wizard of Oz as our participants were told that they were interacting with a computer system which understood their speech, but in fact all system responses were being manipulated by the experimenter, the proverbial “man behind the curtain” (Kelley, 1984). It would have been difficult, if not impossible to develop a voice-controlled system with a pre-set, guaranteed accuracy level from the ground up. By having the experimenter directly manipulate the system responses, set accuracy from the point of view of the participant was achieved.

Two voice-controlled systems implemented through Wizard of Oz (as well as a Baseline condition of no voice-controlled system) were tested. The first system had high recognition accuracy regardless of the background noise present, while the second system had an increasing degradation in accuracy as the level of noise increased. Driving performance was evaluated through various driving measures extracted from the simulator, such as average speed, brake reaction time, lane deviation, and steering wheel angle. Physiological measures including Galvanic Skin Response and ECG (heart rate) were collected to evaluate mental workload. Both are established measures of measuring cognitive workload in driving simulator studies (Mehler et al., 2009). In addition, subjective workload was captured with a questionnaire. An eye tracking system was used to track glances to the display.

3.1 Participants

36 participants (18 male and 18 female) between the ages of 25 and 40 were recruited for this study (mean age=29.8, standard deviation=4.1; males: mean age=30.5, standard deviation=4.1; females: mean age=29.1, standard deviation=4.0). The average number of years of driving experience across participants was 12.6 years, and the standard deviation was 4.1. On a self-report of how many moving violations participants’ have been cited by a police officer in the previous five years, 23 reported zero, 8 reported one, 3 reported two, 1 reported four, and 1 reported 7.

Participants were asked a series of questions about their technology use. These questions were adopted from Reimer et al. (2013). Participants were asked about their self-reported level of experience with technology on a scale of 0-10 (mean = 8.9, standard deviation = 1.3). Similarly, participants were asked about how much they considered themselves "early adopters" of new technology on a scale of 0-10 (mean = 7.4, standard deviation = 1.4). Thus, our recruited participants considered themselves to be familiar with new technologies and considered themselves wholehearted adopters of new technology. However, when asked whether they owned or regularly operated a vehicle with in-vehicle voice-controlled infotainment systems, only 10 participants said they did, indicating a low level of experience with in-vehicle voice-controlled systems specifically.

The desired sample size of 36 was set based on the number of different experimental conditions that were run in order to counterbalance the conditions and prevent learning effects. To achieve counterbalancing, 3!x3! unique condition orders were required. Recruitment criteria included self-report of having good hearing, having either uncorrected vision or being able to wear contact lenses, and having driven a minimum of 1600 km in the last 12 months. The participants were also required to have had a full G license or equivalent for at least three years, and participants were required to present their license for confirmation prior to commencement of the study.

Participants were recruited through online advertisement boards and social media. Potential recruits were asked to fill out an online questionnaire (Appendix 1) for screening purposes in order to assess propensity to simulator sickness as well as their driving qualifications. The questionnaire was accessed through an online survey system hosted on a Canadian server by

JitsuTech to meet University of Toronto Research Ethics Board requirements. Participants were compensated $15/hour, and all participants who completed the experiment were afforded a $5 “voice-controlled task performance bonus” which they were told prior to the experiment depended on their interactions with the voice-controlled systems. This bonus was included so that participants would take their interactions with the voice-controlled systems seriously, and potentially amplify frustration if the systems did not work successfully. All participants were provided the full bonus amount as task performance was the same for all participants due to the Wizard of Oz nature of this study.

3.2 Experimental design

This experiment was a 3x3 within-subjects design, with the first factor being the system and the second being the type of background noise presented. The three levels of the system factor were: Baseline (no voice-controlled system), Noise Sensitive Voice-Controlled System (i.e., voice recognition accuracy is “affected” by the level of background noise), and Noise Robust Voice-Controlled System (i.e., constant accuracy regardless of noise). Each of the two voice-controlled systems had distinct accuracy levels. The noise-sensitive system had high accuracy with no background noise, medium accuracy with music, and low accuracy with the louder noise of children talking. This accuracy change was to represent a degradation of capability in the presence of disruptive noise. The noise-robust system had high accuracy under all three noise conditions. Exact accuracy levels can be seen in Table 1. These accuracy levels were suggested to us by our industry partner based on their experience with noise-sensitive voice-controlled systems. The three levels of disruptive noise were:

- “None”: No additional background noise.

- “Music”: The song “Billie Jean” by Michael Jackson, presented at 60 dB.
- “Children Talking”: A series of segments of background noise of a class of kindergarten children conversing with each other and of children arguing, combined and looped at 70 dB.

All participants undertook 9 experimental of drives utilizing all combinations of these factors (3 system levels x 3 noise levels). The order of presentation was blocked on system. Each block consisted of one system level experienced over three consecutive drives (one per noise level). Thus, although there was no interaction with a system during the baseline block, the participants still experienced all three noise levels. The order of the blocks as well as the order of noise levels within each block were fully-counterbalanced leading to 36 unique orders (3! x 3!). Examples of these 36 orders can be seen in Appendix 2. There was blocking on system so that participants could experience each system holistically; counterbalancing was performed to reduce the possibility of learning effects.

Table 1: Accuracy levels of the voice-controlled systems under different noise conditions

System	**Noise**	**Accuracy Level**
Noise-Robust	None	90%
	Music	90%
	Children Talking	90%
Noise-Sensitive	None	90%
	Music	70%
	Children Talking	30%

3.3 Apparatus

A NADS quarter-cab MiniSim™ Driving Simulator (Figure 1) was used, which consists of three 42” widescreen displays, with a 130° horizontal and 24° vertical field of view at a 48” viewing distance. The experiment was developed using the MiniSim Software Suite. The driving scenarios were authored using the Interactive Scenario Authoring Tool. The simulator collects driving measures at 60 Hz. Gaze data was also collected at 60 Hz, using a faceLAB 5.1 Eye Tracking System, equipped with two dashboard mounted cameras. The range of gaze tracking is ±22º in the vertical and ±45º in the horizontal.

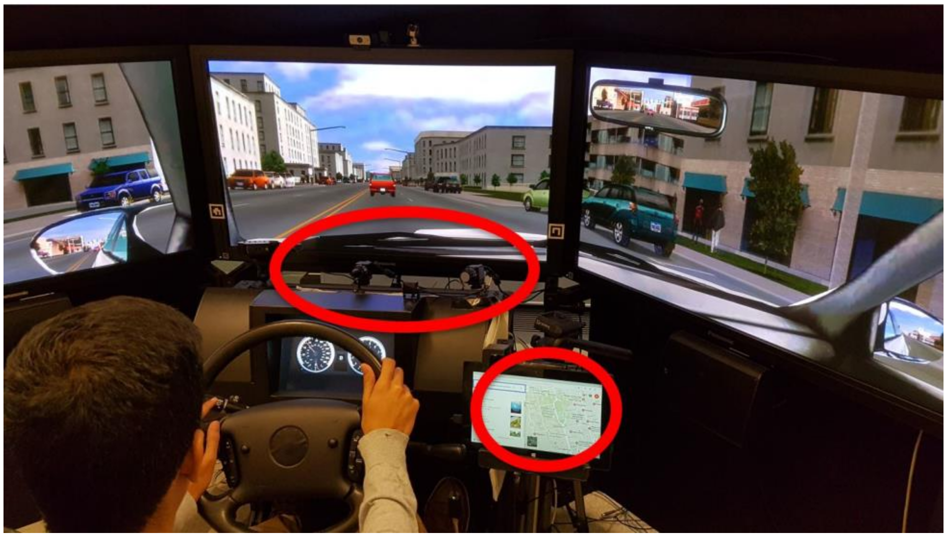


Figure 1: Simulator setup with eye tracker and secondary display highlighted

A Microsoft Surface Pro 2 was used to present the voice-controlled systems and associated tasks to participants. This tablet was positioned to the right of the simulator’s dashboard. In order to provide a realistic Wizard of Oz simulation of a voice-controlled system, two specialized programs were developed using Python. The first was a display program running on the Microsoft Surface Pro 2. This program was capable of displaying images and playing an alert chime and simulated the voice-controlled system itself. The second program was a control panel enabling the experimenter to act as the Wizard of Oz. This second program ran on a desktop computer (the DLab computer mentioned below) that was connected to the

Surface through an internal network. The control panel was also connected to the MiniSim computer, so that each time a control panel function was triggered by the experimenter, the control panel retrieved the simulator frame number. These recorded frame numbers were used during analysis to determine the period of time during which the voice-controlled task interactions occurred.

The background noises were played through the simulator speakers, initiated by triggers set in MiniSim to engage after participants drove 1812 ft (552 m) in a drive. The background noise sound samples were obtained from two different sources. The child noise sample was created by combining multiple sound samples obtained from open source directory of binaural background noise maintained by the European Telecommunications Standards Institute. The music sample was purchased through Apple’s iTunes.

Electrocardiogram (ECG) and Galvanic Skin Response (GSR) sensors and amplifiers developed by Becker-Meditec were used to record participant’s physiological state at a rate of 240 Hz. The ECG system utilizes electrodes placed on the participant’s chest in order to measure electrical activity in the heart. These adhesive electrodes are approximately 1” in diameter. The placement of the three ECG electrodes can be seen in Figure 2. By measuring changes in skin conductivity caused by perspiration, the GSR sensor system can measure emotional and sympathetic responses of the driver. These physiological signals are also responsive to certain emotional states such as stress, frustration and anger.   The GSR system uses the same model of electrode previously described for the ECG system. The two GSR electrodes were placed on the arch of the bottom of the participant’s non-driving foot approximately 1 inch apart. Physiological data was collected and recorded using the DLab experimental recording software, synced to simulator events through frame data collected via a network link between the DLab computer and MiniSim. Electrocardiogram (ECG) and Galvanic Skin Response have been shown to detect changes in mental workload (Meshkati, 1988; Nourbakhsh, Wang, Chen, & Calvo, 2012; Reimer et al., 2013; Ryu & Myung, 2005). Such measures have been shown to be effective in determining the cognitive workload of in-vehicle infotainment systems (Reimer et al., 2013).

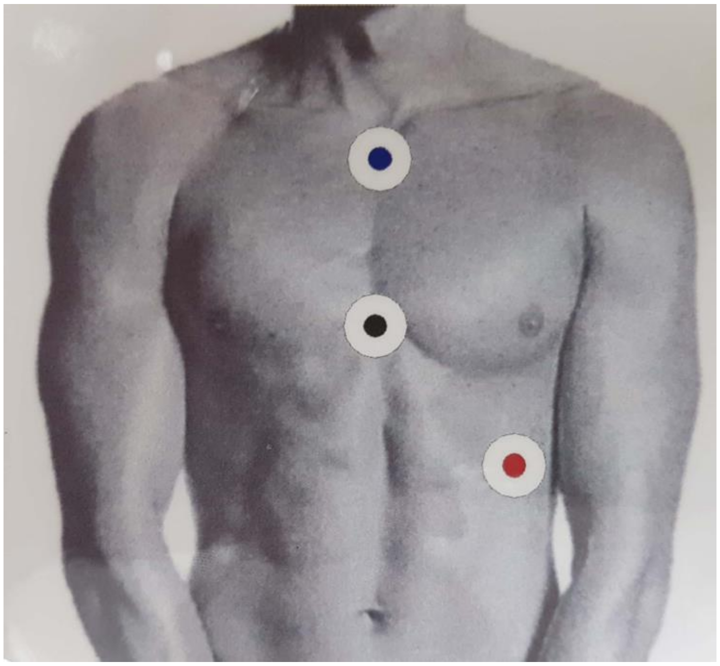


Figure 2: ECG sensor placement recommended by manufacturer

3.4 Driving task

All nine experimental drives used the exact same road, which consisted of an approximately 22912 ft section (6983 m) of a 45ft (13.7 m) wide, undivided four-lane road in an urban environment with a speed limit of 40 mph (64.37 kph). Lane width was 11.25 ft (3.4m). Light oncoming traffic of approximately 8 cars per minute was also present. While the road curved, there were no turns at intersections (participants proceeded straight through intersections). Additional visual clutter was present, consisting of parked cars, as well as stationary and mobile pedestrians.

Participants were informed that their primary task was to drive as safely as possible through this environment while following a lead vehicle and maintaining the speed limit of 40 mph. Each drive took approximately 6-7 minutes depending on the speed the participant maintained. After participants had completed all the tasks they were presented on the voice-controlled system, the final braking event had occurred, and an additional approximately 30 seconds had passed, participants were prompted by the experimenter to pull over to the side of the road at the first safe location they could locate. Once participants had brought the car to a complete stop the drive ended.

The lead vehicle that the participants were asked to follow, was configured to maintain a 2s headway between itself and the participant's vehicle to keep pace with the participant's speed (while the participant maintained a range of 30-50 mph). Four lead-vehicle braking events occurred in each drive, where the lead vehicle braked for 5 s at a rate of 0.61 g (gravitational acceleration). The point at which the braking events occurred in the drive is presented in Table 2. For the two voice-controlled system conditions, braking events 1-3 corresponded to the period that voice-controlled tasks occurred; further details are presented in Section 2.4. It should be noted that the fourth braking event was included as an indicator to the experimenter where to stop the drive. The response to the fourth braking event was not analyzed.

Table 2: Location of braking events in the experimental drive

Braking Event	Location of event from drive start
1	~ 3687 ft (1124 m)
2	~ 6937 ft (2114 m)
3	~ 9143 ft (2787 m)
4	~ 20411 ft (6221 m)

3.5 Voice-controlled infotainment tasks and systems

Aside from the primary driving task, participants undertook a series of secondary tasks while interacting with the simulated voice-controlled systems. In order to prevent biases in their evaluation of the systems, the participants were told that these systems were not developed by the experimenters.

Tasks were chosen to mimic the kinds of voice-controlled infotainment tasks users might perform in a car. The tasks were broadly separated into three categories: music tasks (e.g., find a song by a certain artist), contact tasks (e.g., find the phone number of a certain person), and map tasks (e.g., find the location of nearby pizza restaurants, an example of which can be seen in Figure 1). The complete list of tasks is presented in Appendix 3. Overall, 27 tasks were created in order for there to be no task repetition in any single system block (maximum number of tasks that a participant could complete at 90% accuracy was nine per noise level, thus for three noise levels a total of 27 unique tasks were needed). Further, within each noise level each task type (i.e., music, contact, and map) was included at least once. Aside from this stipulation, the order that the tasks were presented was random. An example task order can be seen in Appendix 4.

Participants were prompted by the experimenter to start these voice-controlled tasks. For example, prior to the example in Figure 4, the participant would be prompted by the experimenter with “Please use the voice-controlled system to find pizza restaurants around the University of Toronto”. Participants were instructed early in the experiment to use the voice command that they felt was most appropriate. The experimenter did not prescribe what the exact voice commands should have been.

In each system drive participants completed 10 interactions with the system. The result of the interaction could either be the system performing a task correctly, or failing at the task. Failures, i.e., recognition errors, were made to appear as substitution errors instead of rejection errors as prior work has shown substitution errors to have a stronger effect on driving behaviour (Gellatly & Dingus, 1998). With a failure, the participant had to repeat the same task. This repeat was considered as another interaction. Thus, a drive with 90% accuracy could have a string of interactions as follows: *Task 1 Success, Task 2 Success, Task 3 Success, Task 4 Success, Task 5 Success, Task 6 Failure, Task 6 Success, Task 7 Success, Task 8 Success, and Task 9 Success.* In the lowest accuracy condition, there were multiple failures on a single task before a success. A drive with 30% accuracy could have a string of interactions as follows: *Task 1 Failure, Task 1 Failure, Task 1 Success, Task 2 Failure, Task 2 Failure, Task 2 Failure, Task 2 Success, Task 3 Failure, Task 3 Failure, and Task 3 Success.*

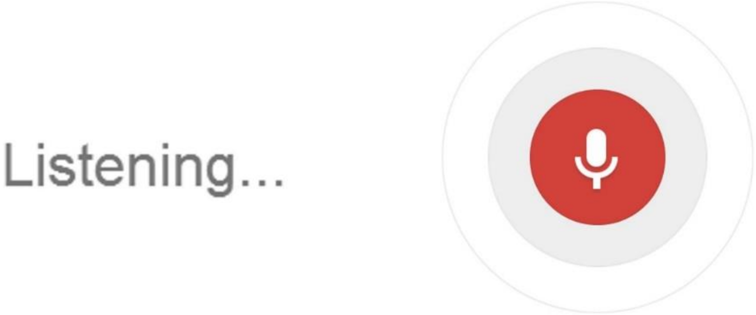
The overall procedure of the voice-controlled tasks was as follows:

1. The experimenter referred to pre-determined task orders to identify if the next task was planned to be a success or failure and input this choice in the control panel that was described in the Apparatus Section.

2. The experimenter prompted the participant to interact with the voice-controlled system, e.g., “Please use the voice-controlled system to find bars near the University of Toronto”.
3. The participant started the interaction with giving the vocal prompt “Hey VC”. This prompt was chosen over established ones (e.g., “Okay Google”) to prevent biases and avoid expectations based on how established systems behave.
4. The experimenter then pressed the button marked “Play Sound” on the control panel, which caused the MS Surface to play a chime and display the word “Listening” along with the image of a microphone (Figure 3). The listening image was constructed by resizing elements of a picture of an Android listening screen obtained from an online source (Jones, 2014).
5. The participant gave their command verbally in a naturalistic manner.
6. The experimenter pressed the “Send Image” button on the control panel to display the voice-controlled system response. See Figure 4 for a successful response and Figure 5 for a failed response. The control panel was designed to also enable the experimenter to display a generic “error” message, in case the participant failed to speak after giving the “Hey VC” command, or if their command was entirely off that displaying the existing canned messages could possibly expose the Wizard of Oz deception. The former case happened a few times during the experiment and these cases were not counted as a failed interaction when implementing system accuracy.
7. If the response was a failure, the same task was repeated from Steps 2 to 6, until the number of failures specified was met. The experimenter then switched the control

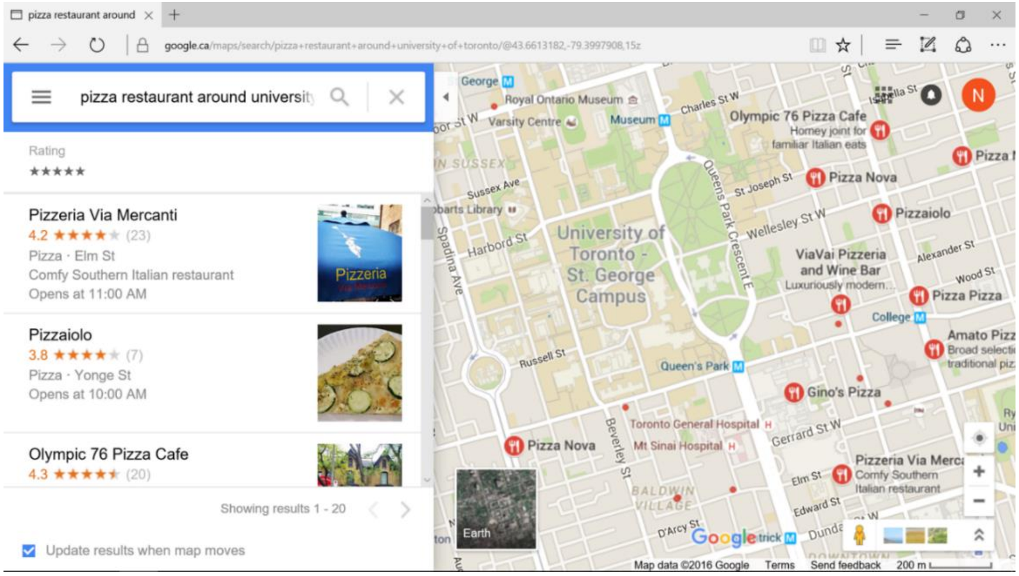
panel from “Failure” to “Success” before the final interaction that was to result in a successful system response.

8. Following 15-20 seconds after a successful response, the next task was presented to the participant starting from Step 1.



Listening...

Figure 3: Listening screen displayed to participants after the participants verbally prompted the system to initiate interaction



pizza restaurant around ×

google.ca/maps/search/pizza+restaurant+around+university+of+toronto/@43.6613182,-79.3997908,15z

pizza restaurant around universit

Rating
★★★★★

Pizzeria Via Mercanti
4.2 ★★★★★ (23)
Pizza · Elm St
Comfy Southern Italian restaurant
Opens at 11:00 AM

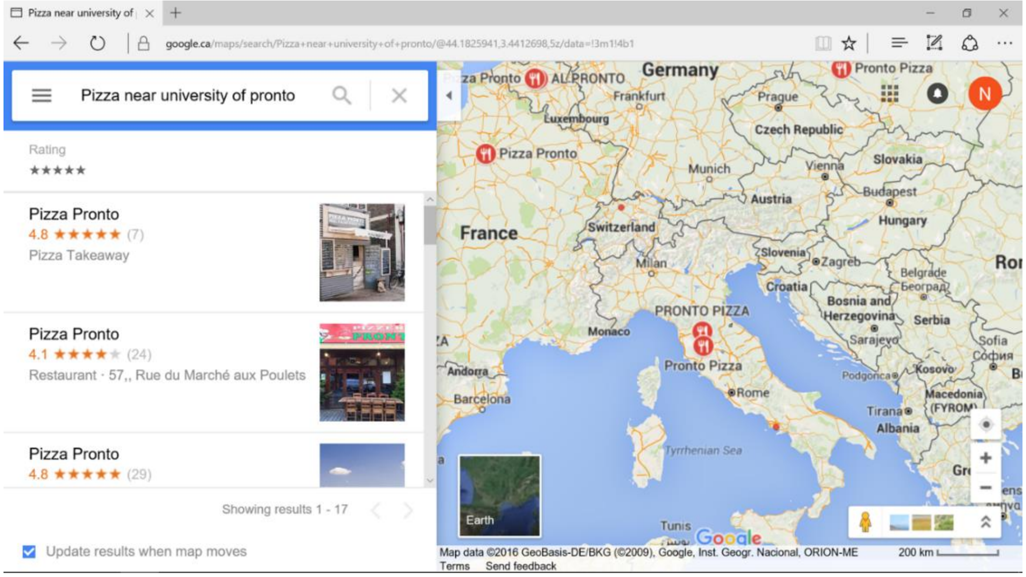
Pizzaiolo
3.8 ★★★★★ (7)
Pizza · Yonge St
Opens at 10:00 AM

Olympic 76 Pizza Cafe
4.3 ★★★★★ (20)

Showing results 1 - 20

☑ Update results when map moves

Figure 4: Screen displayed to participants when "Search for Pizza Restaurant" command was recognized by the system successfully



Pizza near university of pronto

Rating ★★★★★

Pizza Pronto
4.8 ★★★★★ (7)
Pizza Takeaway

Pizza Pronto
4.1 ★★★★★ (24)
Restaurant · 57,, Rue du Marché aux Poulets

Pizza Pronto
4.8 ★★★★★ (29)

Showing results 1 - 17

☑ Update results when map moves

Figure 5: Screen displayed to participants when "Search for Pizza Restaurant" command was not recognized by the system successfully

3.6 Procedure

At the start of the session, participants were provided with written and verbal information regarding the experiment and its procedures. Participants were required to complete and sign a consent form prior to the start of the experiment (Appendix 9). Once the participant’s consent was obtained, the experiment proceeded. Participants then completed a questionnaire on their demographic information, driving history, and overall technology use while driving including voice-controlled systems (Appendix 6). The entire process of participant arrival, informed consent, and the pre-experimental questionnaire took approximately 15 minutes.

Next, the Facelab eye tracking system was calibrated. Calibration was repeated until at least 75% accuracy could be maintained. Following this calibration, ECG and GSR sensors were applied to participants. A silicon-based conductive gel was applied to the ECG electrodes prior to adhesion in order to increase surface conduction. As the ECG sensors were in potentially sensitive areas, participants were instructed on the proper area to apply each sensor pad and lead and then left in private to attempt to apply them themselves. If they had difficulty doing so, or if upon inspection of the waveforms there appeared to be abnormalities, an experimenter of the same gender as the participant assisted them with the application. If either Facelab could not be properly calibrated or a proper ECG and GSR signal could not be achieved, the experiment was stopped and the participant was compensated pro-rated based on the time they spent up to that point.

After successful system calibration, participants became accustomed to the simulated environment through an approximately 5-minute-long practice drive. Before this drive, participants were familiarized with the signs of simulator sickness and were instructed to inform the experimenter if they felt they were experiencing them. If it became clear that a participant was experiencing simulator sickness, the experiment was immediately terminated and the participant was compensated up to that point in order to preserve both their well-being and data integrity. During the practice drive, participants also practiced interacting with the voice-controlled system with three tasks (T1, T2, and T5 in Appendix 4) at 100% recognition accuracy. The practice drive ended once participants reported that they felt comfortable with operating the simulated vehicle and that they were comfortable with how to interact with the voice-controlled system.

Following the practice drive, participants completed the nine experimental drives blocked into three system conditions (i.e., Baseline, Noise Sensitive, and Noise Robust). Before the two blocks that corresponded to a voice-controlled system, the participants were informed that we were asking them to evaluate two new voice-controlled systems developed by a group outside the lab. This instruction was given to prevent any biases towards the systems stemming from participants believing that the experimenters developed the systems themselves. The participants were not informed whether the systems would be noise sensitive or noise robust; however, they were told that the systems were different. Participants were offered a 5 minute rest after each block of three drives. At the end of each drive, in which a voice-controlled system was used, the participants completed the Van Der Laan System Acceptance Questionnaire (Van Der Laan et al., 1997) (Appendix 7) and the Rating Scale Mental Effort Questionnaire (Zijlstra, 1993).

After completing the nine drives, participants removed the sensors themselves and were asked to complete another battery of questionnaires, which took the participants about 10 minutes to complete (Appendix 8). This battery included the Susceptibility to Driver Distraction Questionnaire (Feng, Marulanda, & Donmez, 2014), Cognitive Failures Questionnaire (Broadbent, Cooper, FitzGerald, & Parkes, 1982), Eyesneck Impulsiveness Questionnaire (Eysenck, Pearson, Easting, & Allsopp, 1985), Arnett Inventory of Sensation Seeking Questionnaire (Arnett, 1996), and the Manchester Driver Behaviour Questionnaire (Lajunen, Parker, & Stradling, 1998; Parker, Reason, Manstead, & Stradling, 1995; Reason, Manstead, Stradling, Baxter, & Campbell, 1990). These questionnaires were collected to provide further insights into participant behaviours but will not be reported in this thesis. The analyses are left for future work. The Susceptibility to Distracted Driving Questionnaire is a tool developed to determine how participants allocate attention while driving. This questionnaire may assist in understanding participant engagement in voice-controlled tasks. The Cognitive Failures Questionnaire (CFQ) measures a person’s likelihood of committing an error in everyday tasks due to inattention. CFQ may be useful in assessing participant ability to suppress distracting information, such as the disruptive background noise used in this experiment. The impulsiveness (i.e., inability to withhold impulsive behaviour) and venturesomeness (i.e., risk aversion) items from the Eysenck Impulsiveness Questionnaire were included as these have shown (and would allow further investigation of) to be correlated with distraction engagement frequency and unsafe driving behaviours (Feng et al., 2014; Owsley, McGwin, & McNeal, 2003). The Arnett Inventory of Sensation Seeking and the Manchester Driver Behaviour may enable the investigation of correlations between self-reported risky behaviours and observations made in the simulator.

Once the final questionnaire was completed, the participants were debriefed on Wizard of Oz nature of the study and the deception surrounding the task performance bonus. An informal verbal debrief was then given and participants were afforded the opportunity to withdraw their consent were these deceptions considered disagreeable to them. As per ethics protocol guidelines, a formal email debriefing also was provided to all participants following the completion of all data collection, in which another opportunity to withdraw from the study was given. No participant withdrew his or her data. Three participants stated they had realized the deception part of the way through the study. However, their data was not removed as during statistical analysis it did not appear that this knowledge resulted in outlier results.

3.7 Measures and data analysis

Analyses were conducted on the data that was collected on the segment of the drive where the task was available. For drives involving a system, one of two criteria was used. First, if the frame numbers recorded by the voice-controlled system were intact (there were no network errors which disrupted recording) then all frames from between when “Play Sound” was clicked on the first task to when the final response screen of the last task was displayed were used. Otherwise, the frames from 45 seconds into the drive (approximately when background noise, and therefore, the tasks would start) through five minutes later (approximately the time it took to complete a task series) were used. Out of the 216 total drives using either voice-controlled system, 25 had network error necessitating the use of the second method. The second methodology was also used for baseline drives as there were no tasks in the baseline drive by definition. The resulting range of frames is referred to as the task period. A quality test was performed comparing the results of one measure (average speed) using both methods on 30 random drives that had intact frame files. It was found that the average difference of calculated average speed between the two methods was 2%, and this was deemed a small enough difference for the two methods to be comparable.

A summary of all dependent measures can be seen in Table 3; these measures are defined in detail in the following sections. For user acceptance, ratings of usefulness and satisfaction were calculated. For workload, subjective workload based on the Rating Scale Mental Effort score, as well as the physiological signals of heart rate and galvanic skin response were obtained. For driving performance, average speed, standard deviation of speed, and standard deviation of lane position, accelerator release time, transition time, brake response time, maximum deceleration, and minimum time to collision (TTC) were calculated. For glance behaviour, average glance duration, rate of glances longer than 2 seconds (per minute), rate of glances longer than 1.6 seconds (per minute), and rate of glances shorter than 1.6 seconds (per minute) were obtained.

Table 3: Summary of dependent measures analyzed

Construct	**Dependent measures**
Acceptance	Usefulness Satisfaction
Driving Performance	Avg. Speed SD Speed SD Lane Position Avg. Accel. Release Time Avg. Transition Time Avg. Brake Response Time Avg. Max. Deceleration Avg. Min. TTC
Workload	RSME Score GSR Heart Rate
Glances to the Display	Average Glance Duration Rate of Glances ≥ 2s, ≥ 1.6s, and < 1.6s

3.7.1 Acceptance

Data from the system acceptance questionnaire was processed as described in the original paper (Van Der Laan et al., 1997). Items 1,3,5,7, and 9 were summed and divided by 5 to get the satisfaction score, and items 2,3,6,8 were summed and divided by 4 to create the usefulness score, both on a scale of -2 to 2 with -2 corresponding to very unsatisfying/not useful, and 2 corresponding to very satisfying/useful. There were no missing responses to any of the items.

3.7.2 Driving Performance

Average speed was calculated by taking the average velocity (MiniSim's 'VDS_Veh_Speed' variable, in miles per hour) throughout the task period. Speed variability was calculated by applying MatLab's standard deviation function to the same data. Standard deviation of lane position was found by taking the standard deviation of MiniSim's 'SCC_Lane_Deviation' (in feet from center of lane) variable throughout the task period. However, data around the braking events was excluded for all three measures, starting from the beginning of the braking event (lead vehicle brake lights coming on) to ten seconds after the lead vehicle starts to accelerate again.

Accelerator release time was measured according to SAE Guideline J2944, in which release time is measured from the time the braking vehicle starts to brake to when the accelerator is completely non-pressed. Brake response time was also determined according to SAE Guideline J2944. The guideline stipulates that the brake event response time should, in the case of the driver’s foot not already being in contact with the brake, be measured from the time that the braking vehicle starts to brake, to when 1% or greater brake pressure is applied. Transition time is the difference between brake response and accelerator release times. Transition time was determined by taking the frame difference between the frame when the accelerator was released and the frame when the brake was applied, divided by 60 (as data was captured at 60 Hz).

Maximum deceleration was determined by taking the minimum value of the first column of Minsim’s ‘VDS_Chassis_CG_Accel’ variable (in feet per second squared) from the onset of the lead vehicle braking to 10 seconds after. This variable corresponds to the acceleration of the participant’s vehicle. Minimum time to collision (TTC) was calculated by taking the minimum of the fifth column of Minisim’s ‘SCC_Follow_Info’ variable (in seconds) from the onset of the lead vehicle braking to 10 seconds after. This variable stores the time to collision between the participant’s vehicle and the designated lead vehicle. Time to collision is the period time it would take for the two vehicles to collide assuming they both maintain their current velocities.

Certain drives were dropped if either of the following conditions were met. First, a drive was dropped from the dataset if the participant response to all three braking events that occurred during the task period could not be detected. Given the criteria, the lack of detected brake response could have been because the participant anticipated the braking event and slowed enough prior to the event that they did not have to depress the brake once the lead vehicle did. If the calculated accelerator release or brake response times resulted in a zero (e.g., the participant was not pressing the accelerator pedal when the lead vehicle started braking), or the time was too high as to be a clear error or lack of response to the braking event (one braking event was indicated at ~42 seconds), the drive was also dropped. As suggested by SAE Guideline J2944, any minimum TTC higher than 15 seconds was also dropped. These filters resulted in a loss of 104 data points out of 285. The remaining data points were still highly evenly distributed across all system and noise conditions. All of these brake response measures were averaged across the three braking events that occurred across the task period.

3.7.3 Workload

Data from the Rating Scale Mental Effort Questionnaire was cleaned by removing any survey records with empty RSME scores (due to the participant missing it on the second page of the survey). One participant's RSME scores were dropped as they reported during the second to last survey that they had previously misunderstood what the scaling of RSME represented.

Galvanic Skin Response data was analysed in the following way. First, a ten second period was taken at the beginning of each drive (before task start) and the average GSR was calculated as a baseline level of GSR. Next, the section of frames during the task period was segmented out and the average GSR in that time was calculated. The ten-second average at the start of the drive minus the average over the task period was the final GSR value analyzed. The change in GSR was analyzed instead of a direct comparison of GSR values in order to account for GSR being higher or lower at the start of some drives than others (e.g., if a participant had previously completed one of the more stressful conditions they might have a higher starting GSR, or if they had previously taken a break they might have a lower starting GSR).

Collected ECG data was used to determine heart rate instead of heart rate variability for analysis, as this measure is still a potential measure for workload but does not require the same longer timespan of data as heart rate variability, where timespans less than five minutes are not recommended (European Society of Cardiology, 1996). ECG data collected was first detrended to normalize the signal and remove any drifts in signal strength over time. Next, peaks (i.e., R waves) were identified using the following criteria. Matlab's *findpeaks* function was used, with a minimum distance between peaks of 40 data points (corresponding to 0.167s) and a minimum height of 60% of the normalized waveform. Heart rate was then determined by taking the number of peaks within the task period, and dividing by the length of task period to get a Beats Per Minute (BPM) measure.

3.9.4 Glance to the Display

Glance data was exported from Facelab and converted to a non-proprietary text file format by Facelab's built in conversion tool. The data then had to be reformatted in order to be loaded into Matlab, as rows in which eye tracking was lost had an inconsistent number of columns of data. A custom Python script was used to pad out these rows with zeros, after which the data was loaded into Matlab. Next, the number of glances and the duration of glances to the Surface were determined in the following manner with the guidance of ISO 15007 (International

Standards Organization, 2013). If either the left or right eye was indicated to be looking at the Surface after looking at another plane, it was considered the start of the glance. The end of the glance was when both the left and the right eye were no longer looking at the Surface.

Glance durations shorter than 100ms were dropped as fixations are rarely less than this period (Salvucci & Goldberg, 2000) and therefore were considered tracking errors/noise. Furthermore, there were instances in which a glance to the Surface had started, but was followed by a short section of “Nothing” (i.e., loss of tracking), and then another period of looking at the Surface. It was believed that in some cases these sections could be one continuous glance, with the eye tracking system having difficulty maintaining tracking. Therefore, if these periods were 100ms or less (the minimum time of a new fixation) then the two Surface glances were combined into a single glance. In addition to glance duration, the following glance rates were analyzed: glances longer than 2 seconds, glances longer than 1.6 seconds, and glances shorter than 1.6 seconds. For in-vehicle information systems, glance times longer than 1.6 seconds have been shown to affect driving performance compared to baseline driving without an in-vehicle task, while glances longer than 2 seconds have been shown to even more significantly affect driving performance compared to baseline driving with no in-vehicle task (Hankey et al., 2000)

3.9.5 Statistical models

For all measures except the glance rates, linear mixed models were constructed in SAS 9.3 using the PROC MIXED function. System and noise were fixed effects and participant was a random effect. Interaction between system and noise was included initially, but dropped when not significant in the model. In the case of accelerator release, transition, and brake response times, average inverse headway distance at the start of the braking event (when the lead vehicle brake lights were triggered and the lead vehicle started to decelerate) was included as a covariate. Headway was included as it has been shown that headway time taken at the time of a lead vehicle's onset of braking may have an effect on drivers' brake response time (Winsum & Heino, 1996). Headway was captured through into inverse headway distance as visual angle of a lead object such as a lead vehicle is inversely proportional to the distance to the object (i.e., headway distance) (Hosking et al., 2013). For glance counts, PROC GENMOD was used to create a Poisson model with the same fixed effects while taking into account the repeated nature of observations through Generalized Estimating Equations. The logarithm of task period duration was the offset variable.

Chapter 4

4.0 Results

This section reports the results on user acceptance, driving performance, subjective and physiological workload, and glance patterns.

4.1 Acceptance

Figure 6 presents the acceptance scores for the two systems evaluated. A significant interaction was found between voice-controlled system and noise (Table 4). Significant main effects were also found for both voice-controlled system and noise. The noise-robust system scored higher than the noise-sensitive system in terms of both usefulness and satisfaction in the presence of background noise: music (usefulness: Δ = 0.54, *t*(181) = 5.77, *p* < .0001; satisfying: Δ = 0.62, *t*(181) = 5.55, *p* < .0001) and child noise (usefulness: Δ = 0.87, *t*(180) = 9.15, *p* < .0001; satisfying Δ = 1.18, *t*(180) = 10.44, *p* < .0001). Post-hoc contrasts showed that the difference was much more pronounced under the child noise condition than the music condition in both usefulness (Δ = 0.49 *t*(180) = 5.20, *p* < .0001) and satisfaction (Δ = 0.55, *t*(180) = 4.95, *p* < .0001). Most interestingly, there was a small but significant difference between the systems under the no noise condition (usefulness: Δ= 0.22, t(181) = 2.34, *p* = .02; satisfying Δ= 0.32, t(180) = 2.83, *p* = .005) under which their performance was identical.

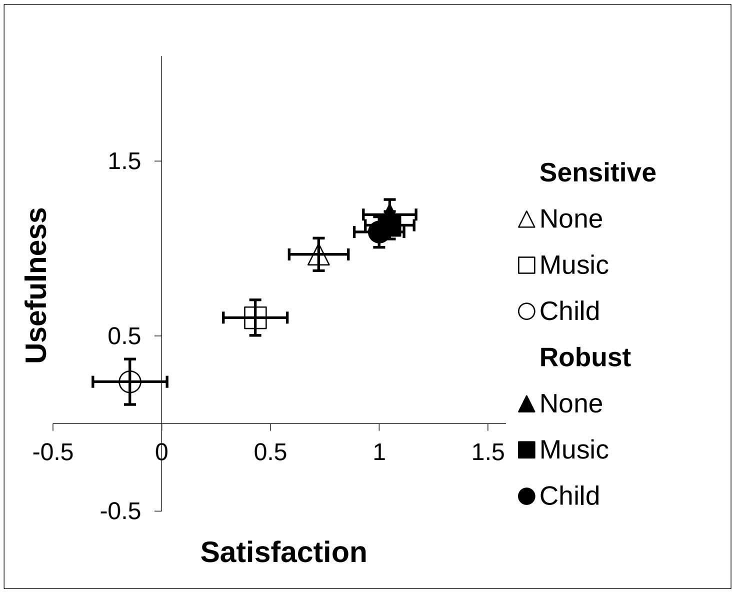


Figure 6: User acceptance of noise robust and noise sensitive systems (the error bars represent Standard Error)

In order to determine if the order in which the system was presented had an effect on user ratings, Welch two sample tests were performed on both scores. No significance was found for either usefulness (*t*(205.9) = 0.86, *p* = .40) or satisfaction (*t*(207.8) = 0.34, *p* = .74) scores between those participants who experienced the sensitive system before the robust system (n=20), and those who received robust system first (n=16). Therefore, the order the participants experienced the systems did not significantly influence their ratings.

Table 4: F-Statistics from mixed linear models on user acceptance

	System		**Noise**		**System *Noise**	
Response Variable	F-value	*p*	F-value	*p*	F-value	*p*
Usefulness	F(2,180) = 99.37	**<.0001**	F(2,180) = 19.99	**<.0001**	F(4,181) = 11.54	**<.0001**
Satisfaction	F(2,180) = 118.33	**<.0001**	F(2,180) = 19.17	**<.0001**	F(4,180) = 14.81	**<.0001**

4.2 Driving

The interaction between system and noise condition was not significant in the initial models of driving data, and was therefore removed in the final models. Thus, this section discusses the main effects of system and noise conditions.

For average speed (Figure 7), system was a significant main effect, but noise was not (Table 5). Post-hoc contrasts showed that while average speed was higher for the baseline compared to the use of either the robust or sensitive voice-controlled system (sensitive: Δ= 1.53, t(288) = 6.10, *p* < .0001; robust: Δ= 1.05, t(280) = 4.21, *p* < .0001), there was no significant difference between the two systems (Δ= 0.47, t(280) = 1.88, *p* = .06).

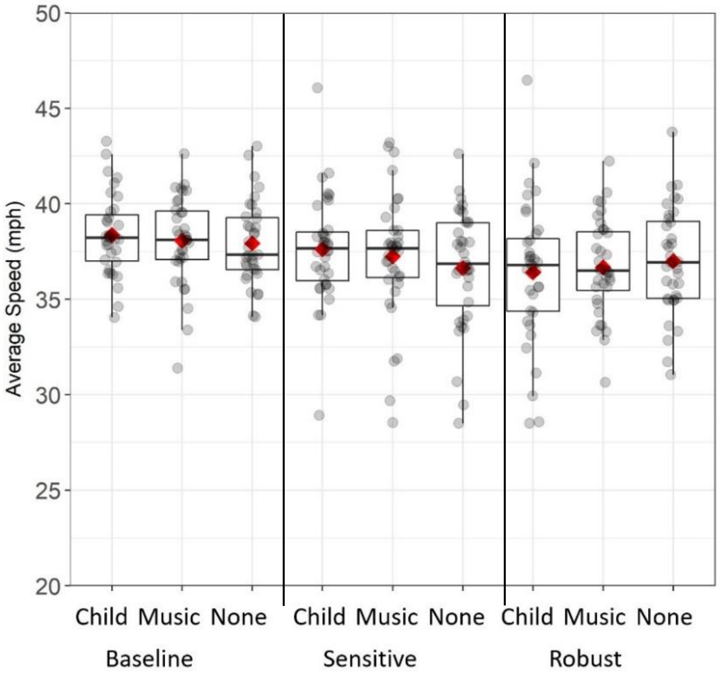


Figure 7: Box plots of average speed for different system and noise conditions (in this figure and in the following ones, the boxplots provide quartile information, the gray circles represent the data, and the red diamond represents the mean)

For standard deviation (SD) of speed (Figure 8), system was a significant main effect, but noise was not (Table 5). As with average speed, contrasts showed that while there was a difference between the baseline and the use of either the robust or the sensitive voice-controlled system (robust: Δ= -1.27, t(288) = -7.46, *p* < .0001; sensitive: Δ= -1.36, t(288) = -7.98, *p* < .0001;), there was no significant difference between the two systems (Δ= -0.09, t(288) = -0.52, *p* = .60). For standard deviation of lane position (Figure 9), there were no significant effects (Table 5).

Table 5: F-Statistics from mixed linear models on speed and lane position

	System		**Noise**	
Response Variable	F-value	*p*	F-value	*p*
Average Speed	F(2,288) = 19.50	**<.0001**	F(2,288) = 0.71	.49
SD Speed	F(2,288) = 39.84	**<.0001**	F(2,288) = 0.36	.70
SD Lane Position	F(2,288) = 1.74	.18	F(2,288) = 0.82	.44

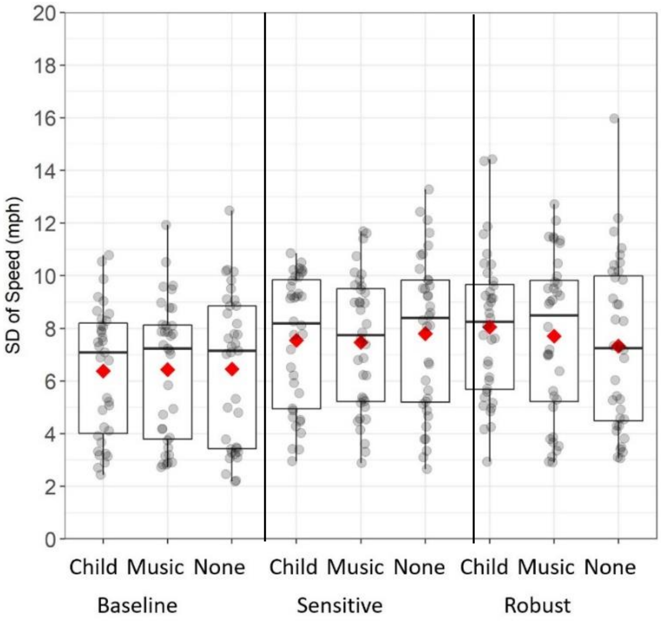


Figure 8: Box plots of standard deviation of speed for different system and noise conditions

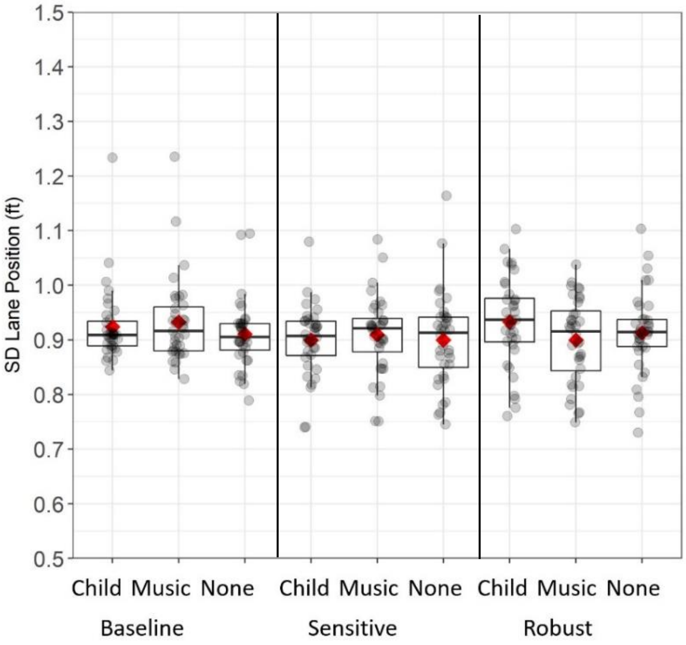


Figure 9: Box plots of standard deviation of lane position for different system and noise conditions

For accelerator release time (Figure 10), system was significant but noise was not (Table 6). Examining contrasts, it was found that there was also a significant difference between the baseline and the robust system, but the difference between the baseline and sensitive system was only approaching significance (robust: Δ= -0.11, t(227) = -2.84, *p* = .005; sensitive: Δ= -0.07, t(245) = -1.94, *p* = .053). Therefore, the robust system appeared to result in a longer accelerator release time than the baseline condition.

Table 6: F-Statistics from mixed linear models on accelerator release time, transition time, brake response time, maximum deceleration, and minimum TTC

	System		**Noise**		**Avg. Inverse Headway Distance**	
Response Variable	F-value	*p*	F-value	*p*	F-value	*p*
Avg. Accel. Release Time	F(2,227) = 4.26	**.02**	F(2,228) = 0.43	.65	F(1,247) = 15.30	< **.0001**
Avg. Transition Time	F(2,252) = 0.74	.48	F(2,252) = 1.06	.35	F(1,261) = 3.44	.06
Avg. Brake Response Time	F(2,252) = 4.13	**.002**	F(2,252) = 0.44	.46	F(1,263) = 28.43	< **.0001**
Avg. Max. Deceleration	F(2,252) = 1.88	.15	F(2,252) = 1.87	.16	F(1,261) =8.80	**.003**
Avg. Min. TTC	F(2,159) = 1.45	.24	F(2,156) = 0.36	.70	F(1,164) = 0.01	.93

With regards to brake transition time (Figure 11) there were no significant main effects. For brake response time (Figure 12), system was approaching significance, but there was no significant main effect for noise (Table 6). Examining contrasts, it was found that there was a significant difference between the baseline and the robust system, but the difference between the baseline and the sensitive system was only approaching significance (robust: Δ= -0.09, t(252) = -2.80, *p* = .006; sensitive: Δ= -0.06, t(246) = -1.90, *p* = .058). With regards to maximum deceleration (Figure 13), system was approaching significance but there was no significant main effect for noise. Examining contrasts, it was found that there was a significant difference between the baseline and the robust system (Δ= -0.12, t(285) = -2.38, *p* = .02). For minimum TTC (Figure 14), there were no significant main effects for either system or noise.

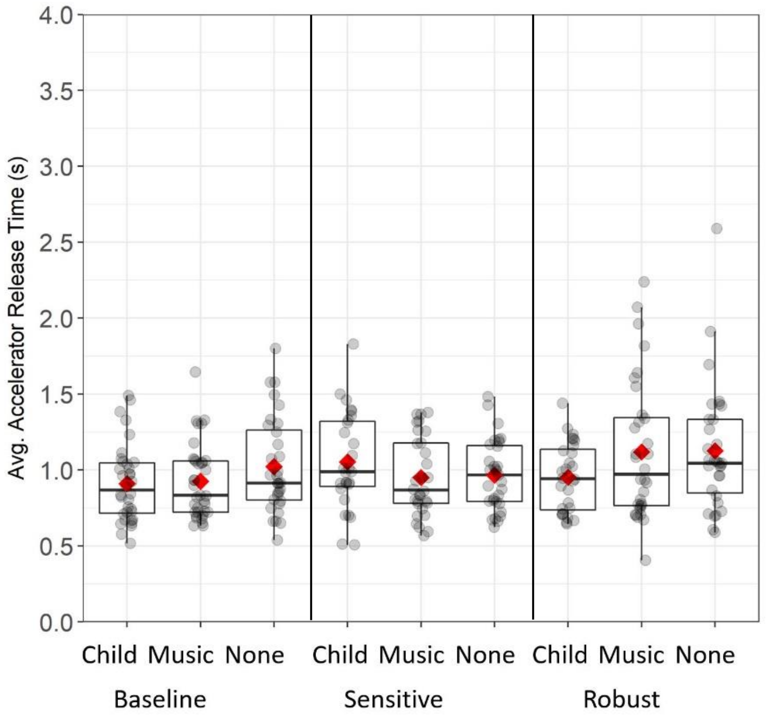


Figure 10: Box plots of accelerator release time for different system and noise conditions

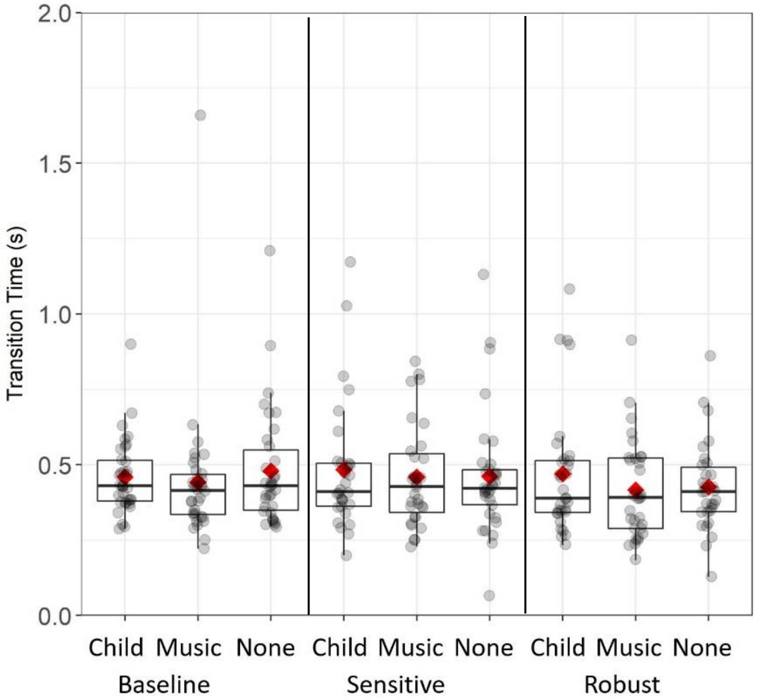


Figure 11: Box plots of transition time for different system and noise conditions

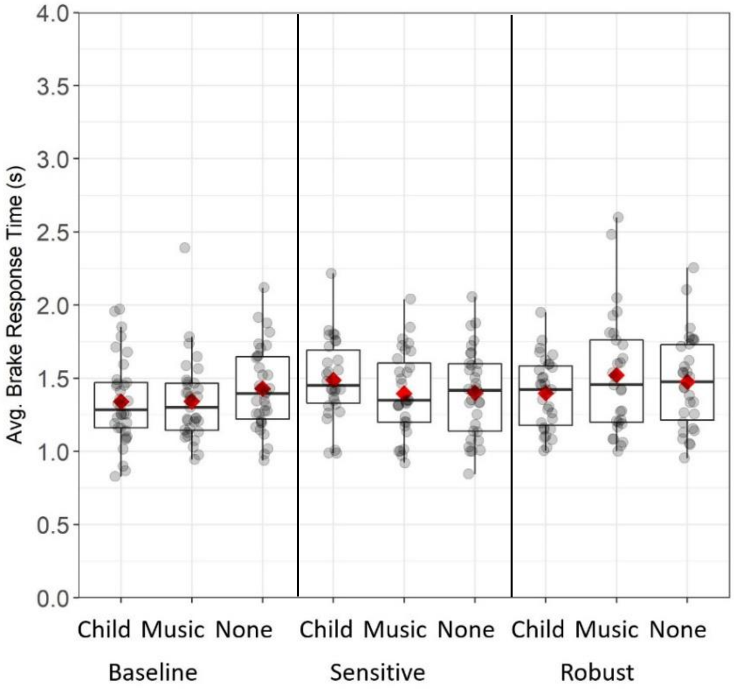


Figure 12: Box plots of brake response time for different system and noise conditions

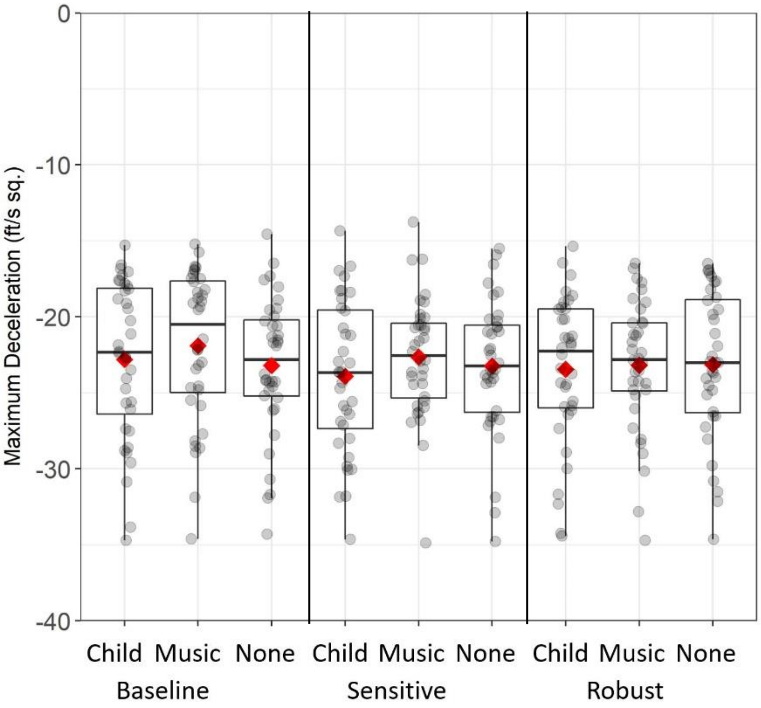


Figure 13: Box plots of maximum deceleration for different system and noise conditions

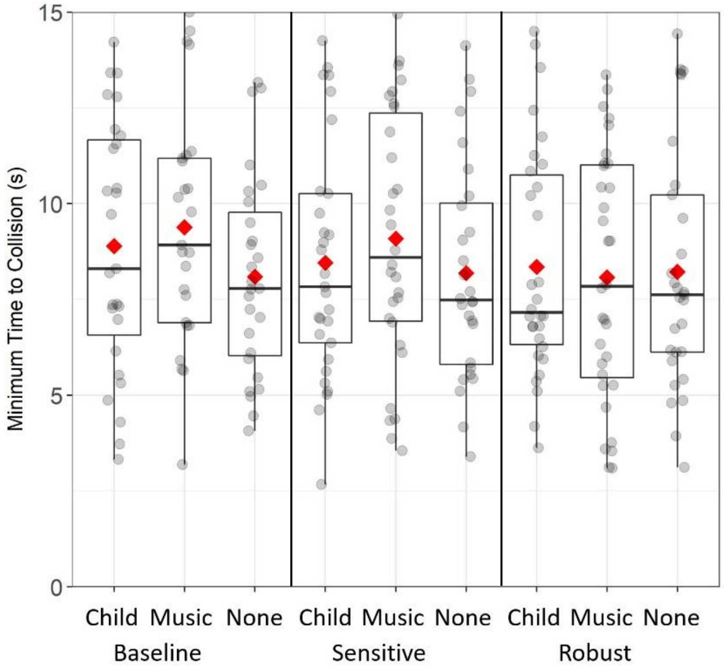


Figure 14: Box plots of minimum TTC for different system and noise conditions

4.3 Workload

For subjective workload (Figure 15), there were significant main effects of system and noise, as well as a significant interaction effect of system and noise (Table 7). The robust system resulted in lower subjective workload than the sensitive system under the music condition (Δ= -23.3, *t*(147) = -4.90, *p* < .0001). The difference was even more pronounced under the child noise condition (Δ = -38.8, *t*(145) = -8.07, *p* < .0001). Unlike with the acceptance scores, no significant difference was found between the workload scores of the two systems under the no noise condition (Δ = -8.57, *t*(145) = -1.75, *p* = .08).

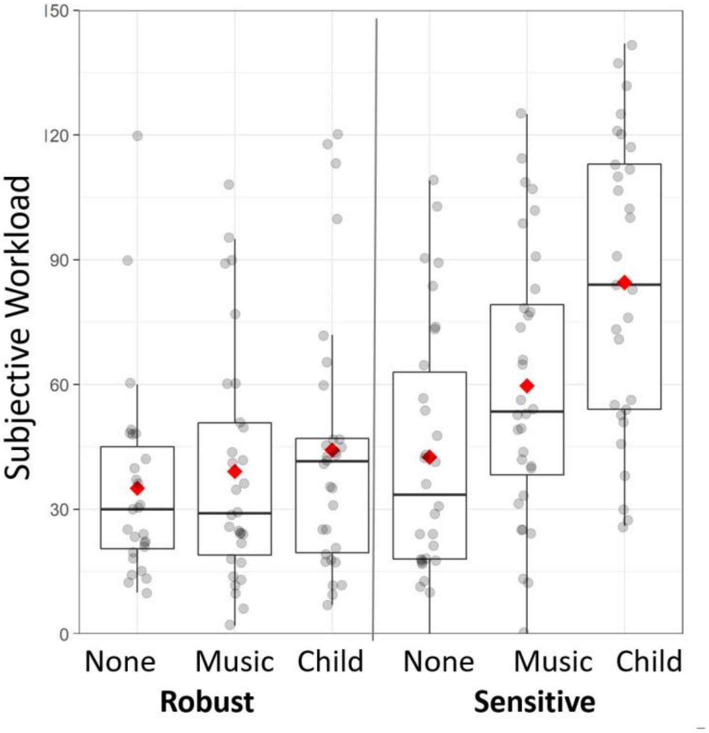


Figure 15: Box plots of RSME scores for different system and noise conditions

Table 7: F-Statistics from mixed linear models on subjective and physiological workload measures

	System		Noise		System *Noise	
Response Variable	F-value	*p*	F-value	*p*	F-value	*p*
Subjective (RSME)	F(1,146) = 71.06	**<.0001**	F(2,146) = 29.97	**<.0001**	F(2,145) = 9.76	**.0001**
GSR	F(2,259) = 0.63	.53	F(2,259) = 0.61	.55	-	
Heart Rate	F(2,259) = 5.93	**.003**	F(2,259) = 0.62	.91	-	

After removing the non-significant interaction term, there were also no significant effects of system or noise in GSR (Table 7, Figure 16).

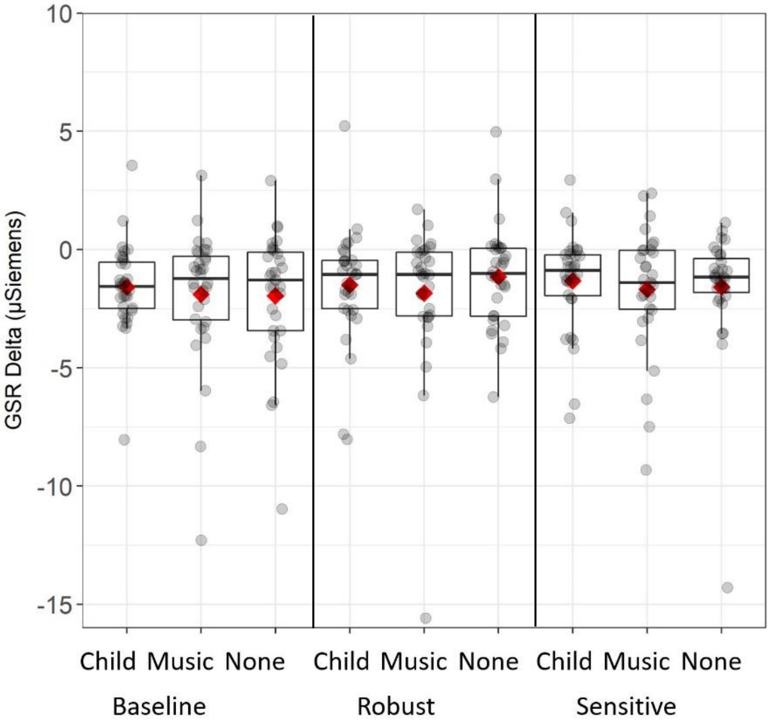


Figure 16: Box plots of GSR delta (GSR at start of drive – average GSR over task period) for different system and noise conditions

With regards to heart rate (Figure 17), one participant was dropped from analysis, as their heart rate data was an outlier, and experimental notes indicated that their body hair had caused issues with sensor contact and the ECG signal. A significant main effect was found for system, but not for noise (Table 7). Follow-up contrasts revealed that the significant difference was between the baseline and the robust system (Δ = -2.83, $t(259) = -2.88$, $p = .004$), as well as between the robust system and the sensitive system (Δ = 3.04, $t(259) = 3.07$, $p = .002$). That is to say, the robust system resulted in a higher heart rate than both the baseline and the sensitive system.

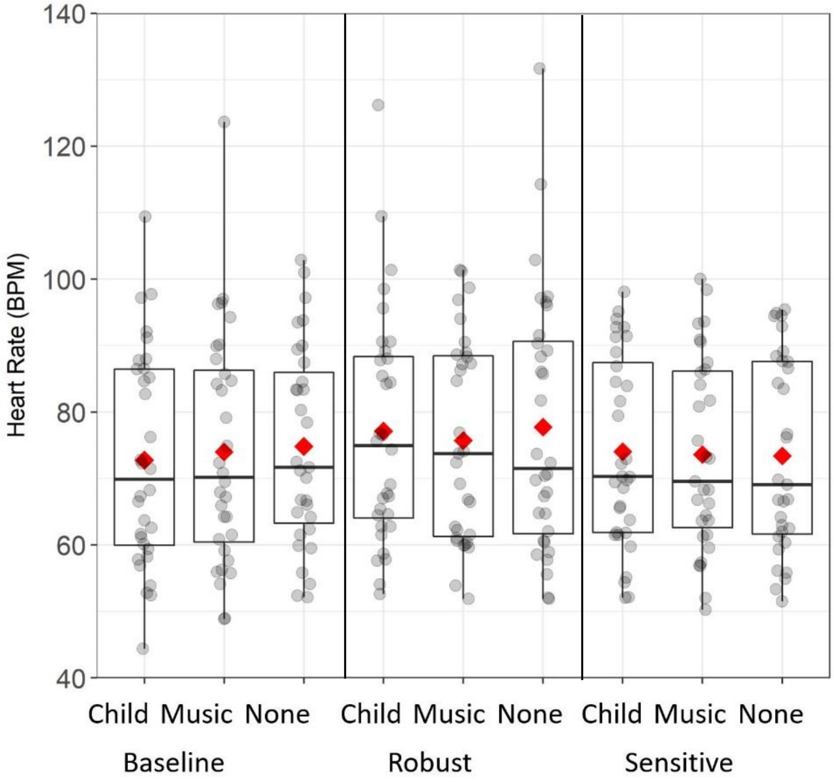


Figure 17: Box plots of heart rate for different system and noise conditions

<u>4.4 Glances to the display</u>

For average glance duration on the in-vehicle display (Figure 18), it was found that noise had a highly significant effect on average glance duration. Further, there was a significant interaction between system and noise (Table 8). System itself was not significant. Examining contrasts, it was found that average glance duration under the child noise condition was longer for the sensitive system than the robust system (Δ= 0.23, *t*(162) = 3.06, *p* = .003). With regards to the effect of noise, it was found that noise did not have an effect for the robust system, but for the sensitive system, average glance duration was longer during the child noise condition than under music (Δ= 0.30, *t*(162) = 3.99, *p* < .0001) and no noise (Δ= 0.30, *t*(162) 4.07, *p* < .0001), but not different between music and no noise (Δ= 0.006, *t*(162) = 0.08, *p* = .93).

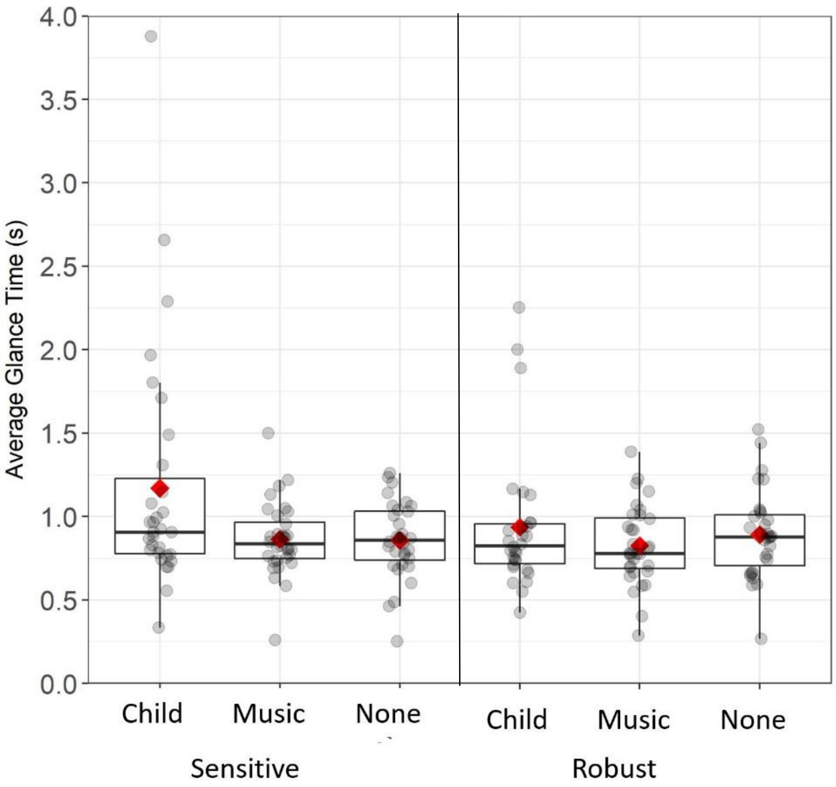


Figure 18: Box plot of average glance duration for system condition and noise

Table 8: F-Statistics from mixed linear model on average glance duration

	System		**Noise**		**System *Noise**	
Response Variable	F-value	*p*	F-value	*p*	F-value	*p*
Avg. Glance Duration	F(1,162) = 3.51	.06	F(2,162) = 8.62	**.0004**	F(2,162) = 3.20	**.042**

For all glance counts, the interaction between system and noise was found to be not significant and was dropped from the models; final model statics are reported in Table 9. For glances greater than 1.6 seconds (Figure 19), system was significant, but noise was not significant. The follow-up contrast revealed that the sensitive system had a higher rate of glances greater than 1.6 seconds (per minute of drive time) than the robust system (Δ= 174%, z = 3.71, p = .0002).

With regards to number of glances longer than 2 seconds and glances shorter than 1.6 seconds, there were no significant effects.

Table 9: Wald statistics from Poisson regression on glance frequency measures

	System		**Noise**	
Response Variable	Chi-Square	*p*	Chi-Square	*p*
Num. Glances ≥ 2s	χ^2 (1, N = 35) = 3.18	.07	χ^2 (2, N = 35) = 1.21	.54
Num. Glances ≥ 1.6s	χ^2 (1, N = 35) = 13.75	**.0002**	χ^2 (2, N = 35) = 0.81	.66
Num. Glances < 1.6s	χ^2 (1, N = 35) = 1.30	.25	χ^2 (2, N = 35) = 0.45	.80

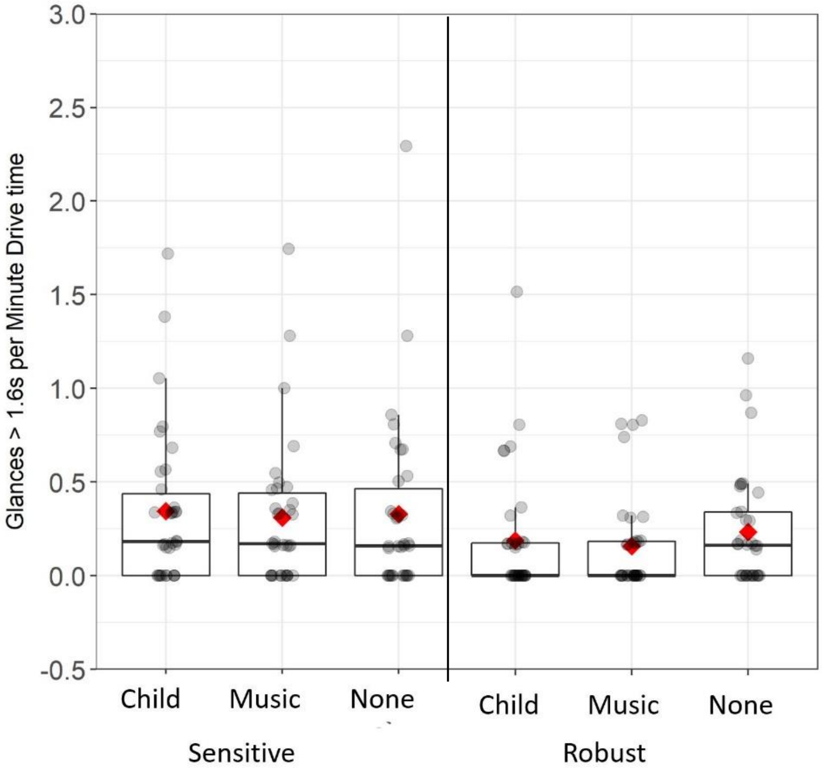


Figure 19: Box plots of glances ≥ 1.6s per min for different system and noise conditions

Chapter 5

5.0 Discussion

5.1 Acceptance and subjective workload

Subjective measures showed clear advantages to the robust system over the sensitive system. While results show that participants largely perceived both systems to be useful and satisfying, participants scored the noise-robust system higher in terms of both usefulness and satisfaction than the noise-sensitive system in all cases. Therefore we can confirm hypothesis H1.

There is a particularly interesting case with regards to the ‘None’ noise condition. When no background noise was present, both systems performed at an identical 90% accuracy rate. Despite this identical level of performance, participants still scored the sensitive system lower. This difference suggests that participants perceive their acceptance of the system based more on its overall performance than any individual use of the system. As mentioned in the introduction to this thesis, it has been suggested that participants’ may cultivate distrust in a system if it fails, even if that failure is ostensibly due to environmental factors outside its control (Lee & See, 2004). Other than trust, effectiveness and efficiency may also play a role here. Effectiveness is a goal of usability, which is described by Preece, Rogers, and Sharp (2015) as “how good a product is at doing what it is supposed to do”. A voice-controlled system that does not correctly respond to voice-commands is failing at this usability goal. Preece, Rogers, and Sharp (2015) describe efficiency as “the way a product supports users in carrying out their tasks”. A voice-controlled system that requires multiple interactions before completing a task is not efficiently supporting the users. By failing in these goals of usability, a noise-sensitive system would provide an objectively worse user experience.

Subjective workload showed a different trend in participants' perception about the systems. Based on the results of the RSME survey, it appears that participants strongly link the workload of using a voice-controlled system to the accuracy of that system. Participants did not view the noise-sensitive system to result in higher workload under the no noise condition. However under noisy conditions, the robust system was clearly viewed as imposing less workload. These results partially confirm hypothesis H2.

These subjective results have broad implications for the future development of in-car voice controlled systems. Manufacturers may invest an increasing amount of money into development and marketing of infotainment systems, but if users do not harbor enough of a positive view of the system to actually use it, that investment would be wasted. As the benefits seem clear with regards to the link between accuracy consistency under noisy conditions and acceptance, manufacturers of voice controlled infotainment systems would be wise to concentrate efforts on developing and adopting technologies to improve noise robustness.

5.2 Physiological workload, driving and glance behaviour

Compared to the subjective measures, objective measures did not completely support the advantages of the robust system. Glance behaviour was the one metric that did. Between average glance duration and number of glances longer than 1.6s, the robust system appears to offer a clear advantage over the sensitive system. Background noise (i.e., system accuracy) had no effect on the length of glances for the robust system; whereas, when the sensitive system was in use, glance duration was longer under conditions of lower accuracy. Compared to the robust system, the sensitive system also resulted in a larger frequency of glances longer than 1.6 seconds. Thus, both accuracy and consistency of accuracy have a positive effect on glance behaviour, confirming hypothesis H3. The reduction in visual distraction with noise-robust systems is a benefit that automotive manufacturers should take into account when creating future systems.

The glance behaviour observed also reiterates that - as prior studies by Maciej & Vollrath (2009) and Reimer et. al. (2013) have shown - voice-control does not equate to an absence of unsafe off-road glances. While some off-road glances may be inevitable in certain situations, such as confirming a correct navigation entry, a successful system should have minimal visual feedback. If users are glancing more frequently due to a lack of trust in the system, then even if the manufacturers reduce the level of feedback provided, the number of unsafe glances may not be reduced as much as expected.

Results demonstrated that average speed was reduced and speed variability was increased by the use of either voice-controlled system, while lane keeping was not affected. These findings are in line with previous work by Kun et al. (2007). These effects were minor, not significantly more pronounced for the sensitive system, and in the case of average speed, the result was a reduction in speed. These findings potentially suggest that participants are adopting compensatory behaviours in response to the additional cognitive load associated with the voice-controlled task, similar to those seen in Reimer et al. (2013). However, accelerator release and brake response times indicate that the robust system resulted in significantly worse braking event response compared to the baseline, whereas no differences were observed for the sensitive system. This result is unexpected given the glance results, which show that participants glance away from the road for a shorter duration and less frequently when using the robust system compared to the sensitive system. Regardless, because of these results we cannot fully confirm hypothesis H4.

The lack of significance in GSR and the heart rate being significantly higher for the robust system were unexpected, given that the robust system was viewed by participants as resulting in lower workload than the sensitive system. However, given the sensitivity of the ECG measures (the difference was at most 3 beats per minute), there were a number of issues that could account for this unexpected result. First, issues with the sensor system itself (see section 6.2) may have resulted in too much noise with regards to the ECG data. ECG sensors were placed in positions recommended by the system vendor. However, upon investigation after issues were discovered in the data, no medical literature could be found recommending placement of a three lead ECG system in the manner suggested. These issues were discussed with Dr. Bruce Mehler (personal communication, June 26, 2017). Dr. Mehler has authored or co-authored a significant number of studies applying these measures in the driving domain (Mehler et al., 2009; Reimer et al., 2014, 2013). Based on his recommendations, future work should make use of an alternative sensor placement strategy, as well as modifications to our procedures to include the step of cleaning the site of the sensors with an alcohol wipe to remove dead skin cells. According to Dr. Mehler, both vastly improve R wave prominence, which make the peak finding and heart rate measurements more reliable. While knowing these shortcomings will be useful for future research, we unfortunately cannot trust the reliability of the physiological data collected for this thesis to draw firm conclusions, and hence cannot confirm or reject hypothesis H2.

It is also possible that our physiological setup did not impose enough stress or mental workload increase to be detectable. Studies such as Reimer et al. (2013) took place in on-road highway driving conditions, while ours took place in a simulator. Our participants would not feel the same level of stress when having to repeatedly interact with an unreliable in-car voice-controlled system in a safe, simulated environment as they would driving a real car on a real road. The reduced stress in the simulator compared to real driving may have affected participant responses to braking events, as unlike real driving, there is no danger in responding too slowly to such events in the simulator. Also, given that the simulator did not have motion, participant behaviour may have been affected, in particular, the ability to perceive and maintain speed. In summary, the results from this experiment do not conclusively support the hypothesis that noise-robustness of a voice-controlled system would enhance driving safety.

Chapter 6

6.0 Limitations and Future Work

6.1 Experimental design and setup

There were some factors in the design of the experiment that must be recognized as limitations. The first involves how the tasks were presented and how associated times were recorded by the custom display software. While there was a defined start and end of the task period, there was no such end for an individual task. Once the task was ‘successfully’ performed, the screen displaying the successful response remained displayed until the next interaction. The lack of a clear end point meant that accurate analysis of the measures on a task-by-task basis (instead of an overall task period basis) was not possible. In the future for similar experiments, asking the participant to confirm the correct result and then having the experimenter mark down that frame number with the software as a task end point would potentially eliminate this limitation.

6.2 Physiological sensors and recording system

During the course of this study, it became clear that the physiological data collected was affected by a number of factors related to the sensor system. The first potential limitation was the fact that for a number of the initial participants, no additional conductive gel was used. The application of conductive gel was a change made after noticing a significant loss of signal when attempting to run a male participant with a large amount of body hair. While the gel did improve the signal, in a few cases, sensor adhesion was affected if too much gel was accidentally used.

Another problem encountered was the limits of the sensitivity of the GSR sensors. The sensor system used could only record values between 0 and 50 μSiemens. Some participants did reach the maximum value, following which the measurement plateaued and any variation in skin conductance could not be captured. When this plateau occurred during a drive, an attempt was made to take a small break to try and have the value decrease back into the variable range and reapply the sensors if necessary. However, the data during the prior drive was irrevocably affected.

6.3 Noise conditions

It also must be recognized that there were some limitations resulting from noise conditions and how they were presented. The first is the nature of the noise. Music can be pleasant, relaxing, and entertaining. A song from the one of the best-selling albums of all time, such as the song Billie Jean by Michael Jackson, is likely to elicit such a reaction. Conversely the sound of children arguing can be stressful and irritating, with whether or not a participant has children or experience driving with small children a factor in their reaction. A solution to this issue would be instead to choose a single noise condition (i.e., just music) and only vary the volume, or keep the noise at a consistent volume and vary the nature of the noise.

6.4 Future work

There are several potential ways to expand on the results and fill in the shortcomings of the study. A future study can study how voice-controlled system accuracy affects users in the presence of background noise (i.e., accuracy degrading with higher noise levels) and when no background noise is present (i.e., accuracy changing in vacuum). This future work is required to bridge this thesis to prior work by Kun et al. (2007) and Reimer et al. (2013).

Additional measures such as a measure of participants' perceived risk while using the system should be collected, as prior work by Keuntje and Poormohammadroohafza (2014) showed that perceived risk was an important negative influence on user acceptance that could be offset by increased perceived ease of use. Similarly, collecting explicit subjective measures on system trust could help confirm the assertions made in this thesis about potential distrust in systems. In Chapter 1, it was discussed how users have been observed in earlier research to make orienting responses during voice-interactions. We were unable to analyze these responses due to our eye tracking system being configured for gaze tracking only, and not head tracking. Facelab does contain the ability to perform head tracking in addition to gaze tracking with additional configuration, and it would be useful to configure the system to add head tracking to allow for investigation of these orienting responses in future studies. Finally, this study asked participants if they owned or regularly operated a vehicle with a voice-controlled infotainment system, and the results showed that the majority did not. Collecting more detailed information on participants' experience with these systems could provide useful covariates in the statistical analysis of future studies.

Chapter 7

7.0 Conclusion

Voice-controlled in-car infotainment systems are only going to become more and more common in future years. The study presented in this thesis has demonstrated that a noise-robust system has several potential advantages over a noise–sensitive system in terms of user acceptance, subjective workload, and glance behaviour. However, with regards to driving performance, there does not appear to be any clear advantage over a noise sensitive system. Acceptance is a key factor for determining if manufacturers' improvements to in-car voice-controlled infotainment systems are being valued by users and leading to adoption of said systems. Therefore, the benefits in how users feel about noise-robust systems provide a promising insight for the future of in-car voice-controlled infotainment systems. However, manufacturers must be cautious not to claim that noise-robust systems can offer a safety improvement over noise-sensitive ones. The results of this thesis do not entirely support such an improvement, and in fact, may suggest a possible disadvantage in terms of safety.

Chapter 8

8.0 References

Arnett, J. J. (1996). Sensation seeking, aggressiveness, and adolescent reckless behavior. *Personality and Individual Differences*, *20*(6), 693–702.

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* (Vol. 1, pp. 212–217 vol.1).

Bagozzi, R. (2007). The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift. *Journal of the Association for Information Systems*, *8*(4), 244–254.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *The British Journal of Clinical Psychology / the British Psychological Society*, *21 (Pt 1)*, 1–16.

Casali, S. P., Williges, B. H., & Dryden, R. D. (1990). Effects of Recognition Accuracy and Vocabulary Size of a Speech Recognition System on Task Performance and User Acceptance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *32*(2), 183–196.

Consumer Reports. (2016, June 2). Brand-by-Brand Guide to Car Infotainment Systems. Retrieved August 17, 2017, from http://www.consumerreports.org/cars/infotainment-system-brand-by-brand-guide/

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–339.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636–2641.

Donmez, B., Boyle, L. N., & Lee, J. D. (2007). Safety implications of providing real-time feedback to distracted drivers. *Accident Analysis & Prevention*, *39*(3), 581–590.

European Naturalistic Driving Study. (n.d.). What is NDS. Retrieved September 6, 2017, from http://www.udrive.eu/index.php/about-udrive/what-is-naturalistic-driving

European Society of Cardiology. (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation*, *93*(5), 1043–1065.

Eysenck, S. B. G., Pearson, P. R., Easting, G., & Allsopp, J. F. (1985). Age norms for impulsiveness, venturesomeness and empathy in adults. *Personality and Individual Differences*, *6*(5), 613–619.

Feng, J., Marulanda, S., & Donmez, B. (2014). Susceptibility to Driver Distraction Questionnaire. *Transportation Research Record: Journal of the Transportation Research Board*, *2434*, 26–34.

Gellatly, A. W., & Dingus, T. A. (1998). Speech Recognition and Automotive Applications: Using Speech to Perform in-Vehicle Tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(17), 1247–1251.

Hankey, J., Dingus, T. A., Hanowski, R., Wierwille, W., & Andrews, C. (2000). *In-Vehicle Information Systems Behavioral Model and Design Support: Final Report* (No. 00-135). Blacksburg, VA: Virginia Tech Transportation Institute.

Hart, C. S. (2010). *Assessing the Impact of Low Workload in Supervisory Control of Networked Unmanned Vehicles*. Massachusetts Institute of Technology. Retrieved from http://www.dtic.mil/docs/citations/ADA531915

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

Hosking, S. G., Davey, C. E., & Kaiser, M. K. (2013). Visual cues for manual control of headway. *Frontiers in Behavioral Neuroscience*, *7*.

IHS Technology. (2013, March 19). Voice Recognition Installed in More than Half of New Cars by 2019. Retrieved October 17, 2016, from https://technology.ihs.com/427146/voice-recognition-installed-in-more-than-half-of-new-cars-by-2019

International Standards Organization. (2013). Road vehicles — Measurement of driver visual behaviour with respect to transport information and control systems — Part 1: Definitions and parameters. International Standards Organization.

Jones, A. (2014, October 9). Google Listening Image from: How To: Have Google Now remind you about paying bills. Retrieved September 29, 2017, from https://www.pocketmeta.com/google-now-remind-paying-bills-12996/

Kelley, J. F. (1984). An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.*, *2*(1), 26–41.

Keuntje, P., & Poormohammadroohafza, F. (2014, June). *Car Infotainment: An early analysis of driver perceptions towards apps in the car*. Lund University, Lund, Sweden.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358.

Kun, A., Paek, T., & Medenica, Z. (2007). The effect of speech interface accuracy on driving performance. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (pp. 1326–1329). Antwerp, Belgium.

Lajunen, T., Parker, D., & Stradling, S. G. (1998). Dimensions of driver anger, aggressive and highway code violations and their mediation by safety orientation in UK drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *1*(2), 107–121.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management*, *40*(3), 191–204.

Lowry, P. B., Gaskin, J., Twyman, N., Hammer, B., & Roberts, T. (2012). *Taking "Fun and Games" Seriously: Proposing the Hedonic-Motivation System Adoption Model (HMSAM)* (SSRN Scholarly Paper No. ID 2177442). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2177442

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, *41*(5), 924–930.

McCallum, M. C., Campbell, J. L., & Richman, J. B. (2004). Speech Recognition and In-Vehicle Telematics Devices: Potential Reductions in Driver Distraction. *International Journal of Speech Technology*, *7*(1), 25–33.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult

Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*, 6–12.

Meshkati, N. (1988). Heart Rate Variability and Mental Workload Assessment. In P. A. H. and N. Meshkati (Ed.), *Advances in Psychology* (Vol. 52, pp. 101–115). North-Holland. Retrieved from http://www.sciencedirect.com/science/article/pii/S0166411508623845

National Highway Transport Safety Administration. (2012, February 15). Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (pp. 420–423). New York, NY, USA: ACM.

Owsley, C., McGwin, G., & McNeal, S. F. (2003). Impact of impulsiveness, venturesomeness, and empathy on driving by older adults. *Journal of Safety Research*, *34*(4), 353–359.

Park, E., & Kim, K. J. (2014). Driver acceptance of car navigation systems: integration of locational accuracy, processing speed, and service and display quality with technology acceptance model. *Personal and Ubiquitous Computing; London*, *18*(3), 503–513.

Parker, D., Reason, J. T., Manstead, A. S. R., & Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. *Ergonomics*, *38*(5), 1036–1048.

Patten, C. J. D., Kircher, A., Östlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, *36*(3), 341–350.

Preece, J., Sharp, H., & Rogers, Y. (2015). In *Interaction Design: Beyond Human-Computer Interaction, 4th Edition* (pp. 19–20). West Sussex, United Kingdom: Wiley. Retrieved from http://www.wiley.com/WileyCDA/WileyTitle/productCd-EHEP003334.html

Ranney, T. A., Harbluk, J. L., & Noy, Y. I. (2005). Effects of Voice Technology on Test Track Driving Performance: Implications for Driver Distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(2), 439–454.

Reason, J. T., Manstead, A. S. R., Stradling, S. G., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction? *Ergonomics*, *33*(10-11), 1315–1332.

Rebman, C. M., Aiken, M. W., & Cegielski, C. G. (2003). Speech recognition in the human–computer interface. *Information & Management*, *40*(6), 509–519.

Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2013). *The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance* (MIT AgeLab Technical Report No. 2013-17A). Cambridge, MA.: Massachusetts Institute of Technology,.

Reimer, B., Mehler, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Rumpold, A. (2014). Effects of an "Expert Mode" Voice Command System on Task Performance, Glance Behavior & Driver Physiology. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 23:1–23:9). New York, NY, USA: ACM.

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, *35*(11), 991–1009.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying Fixations and Saccades in Eye-tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (pp. 71–78). New York, NY, USA: ACM.

Scherer, M. (1986). *Values in the creation, prescription, and use of technological aids and assistive devices for people with physical disabilities*. University of Rochester, Rochester, NY.

Shutko, J., Mayer, K., Laansoo, E., & Tijerina, L. (2009). *Driver Workload Effects of Cell Phone, Music Player, and Text Messaging Tasks with the Ford SYNC Voice Interface versus Handheld Visual-Manual Interfaces* (SAE Technical Paper No. 2009-01-0786). Warrendale, PA: SAE Technical Paper. Retrieved from http://papers.sae.org/2009-01-0786/

Smiley, A. (2005). What is distraction? Presented at the International Conference on Distracted Driving, Toronto, Ontario, Canada.

Strayer, D. L., & Johnston, W. A. (2001). Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychological Science*, *12*(6), 462–466.

Van Der Laan, J. D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*(1), 1–10.

van Driel, C. J. G., Hoedemaeker, M., & van Arem, B. (2007). Impacts of a Congestion Assistant on driving behaviour and acceptance using a driving simulator. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(2), 139–152.

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 186–204.

Waard, D. D., & Brookhuis, K. A. (1997). Behavioural adaptation of drivers to warning and tutoring messages: results from an on–the–road and simulator test. *International Journal of Heavy Vehicle Systems*, *4*(2-4), 222–234.

Waugh, J. D., Glumm, M. M., Kilduff, P. W., Tauson, R. A., Smyth, C. C., & Pillalamarri, R. S. (2000). Cognitive Workload while Driving and Talking on a Cellular Phone or to a Passenger. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(33), 6–276.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (pp. 63–102). New York: Academic Press.

Winsum, W. V., & Heino, A. (1996). Choice of time-headway in car-following and the role of time-to-collision information in braking. *Ergonomics*, *39*(4), 579–592.

Wu, J., Chang, C.-C., Boyle, L. N., & Jenness, J. (2015). Impact of In-vehicle Voice Control Systems on Driver Distraction Insights From Contextual Interviews. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1583–1587.

Zijlstra, F. R. . (1993). *Efficiency in Work Behavior. A Design Approach for Modern Tools*. TU Delft, Delft, Netherlands.

Appendix 1: Screening Questionnaire

University of Toronto, Human Factors and Applied Statistics Lab

You are invited to participate in a research study conducted by the Human Factors and Applied Statistics Lab (Director: Prof. Birsen Donmez) at the Department of Mechanical and Industrial Engineering, University of Toronto. The experiment will take place in the Rosebrugh Building, located on the University of Toronto St. George campus. Participants will be compensated at a rate of $15/hour, for approximately two hours ($30 total), and has a chance to earn a performance bonus of up to $5.

The goal of this study is to understand driver behaviour and make our roads safer. If you choose to participate, you will be presented with questions about yourself and your driving behaviour. You will also be asked to perform simple tasks interacting with a voice control system while driving in a driving simulator.

Please note that all information collected will be held in the strictest confidentiality. Personal data will be stored securely in the Human Factors and Applied Statistics Lab at the University of Toronto. Under no circumstances will personal data be revealed to any third party, for any purpose. Research findings that we disseminate via scientific publications and reports will be at an aggregated level, such that no individual may be identified by any means.

At this moment, we invite drivers with a full, valid driver's license (G driver license or equivalent), normal to corrected vision, and normal hearing to complete the following questionnaire. This questionnaire will help us determine your eligibility for participating in our research. If you have any questions or concerns, please email us at nsokol@mie.utoronto.ca or 416.978.0881.

Screening Questionnaire

1. Your first and last name: ________________________________
2. Please provide your preferred method of contact:

e-mail address: ______________________________

and/or

phone number: ______________________________

3. What is your age? ______________________________
4. What is your sex? Male Fema
5. What valid driver's licenses do you currently hold?
   a. Full driver's license (e.g. G license in Ontario)
   b. Learner's license (e.g. G1 and G2 licenses in Ontario)
   c. Other licenses (please specify): ________________________
   d. I do not currently have a valid government issued driver's license

6. How often do you drive a car or other motor vehicle?

   a. Almost every day
   b. A few days a week
   c. A few days a month
   d. A few times a year
   e. Never

7. Over the last year, how many kilometers did you drive?

   a. Under 1600
   b. Between 1,601 and 8,000
   c. Between 8,001 and 16,000
   d. Between 16,001 and 24,000
   e. Between 24,001 and 32,000
   f. Over 32,001
   g. None
   h. I don't know

8. Please provide the city and province where you drive most often:

City: ______________________

Province: ___________________

9. How would you describe your **physical** well-being (over the past month including today)?

a. Excellent
b. Good
c. Average
d. Fair
e. Poor

10. Compared with others your age, how would you rate your overall **vision**? (If you wear glasses or contacts, rate your corrected vision when you are wearing them.)

a. Excellent
b. Good
c. Average
d. Fair
e. Poor

11. Compared with others your age, how would you rate your overall **hearing**?

a. Excellent
b. Good
c. Average
d. Fair
e. Poor

Some people tend to experience a type of motion sickness, called simulator sickness, when driving the simulator. The next questions are asked to help us identify if you might be prone to simulator sickness.

12. Do you frequently experience migraine headaches?

    a. Yes
    b. No

13. Do you experience motion sickness?

    a. Yes
    b. No

14. Do you experience claustrophobia?

    a. Yes
    b. No

15. Are you pregnant?

    a. Yes
    b. No

In this study, we will be collecting physiological data including heart rate.

16. Are you comfortable with temporary sensors being attached to your skin (e.g., Electrocardiogram sensors) ?

    a. Yes
    b. No

Appendix 2: Example Condition Orders

Total number of unique conditions was 36 (3! X 3!).

Three example orders of blocked condition orders:

Participant	**Order (Drive 1 to 9)**		
1	Noise Sensitive None Music Child	Baseline Music None Child	Noise Robust None Child Music
2	Noise Sensitive None Music Child	Noise Robust None Child Music	Baseline Child Music None
3	Baseline None Music Child	Noise Robust None Child Music	Noise Sensitive Child Music None

Appendix 3: Voice-Controlled System Tasks

The text below outlines the different tasks the participant will be asked to perform using each "Voice-Controlled System". Researchers will use the scripted lines to prompt users to engage in the task.

"Please pay careful attention to the following instructions. I will now ask you to use the [First/Second] voice-controlled system located to the right of the steering wheel. Using only verbal instructions, please attempt to carry out the tasks I indicate. To initiate a command, use the phrase "Okay VC", and then wait for a chime from the system before proceeding. While performing these tasks, keep in mind your primary goal is to drive as safely as possible".

Task Code	**Prompt**
T1 (Navigation task)	"Please use the voice-controlled system to locate pizza restaurants around the University of Toronto"
T2 (Contact task)	"Please use the voice-controlled system to find the phone number for Bob Johnson"
T3 (Contact task)	"Please use the voice-controlled system to find the address for Sally"
T4 (Contact task)	"Please use the voice-controlled system to find the address for Homer Simpson
T5 (Music task)	"Please use the voice-controlled system to play the song "Rocketman"
T6 (Music task)	"Please use the voice-controlled system to play music by the band "The Rolling Stones"
T7 (Music task)	"Please use the voice-controlled system to play the song "Let it Be"
T8 (Navigation task)	"Please use the voice-controlled system to locate hotels near Union Station
T9 (Contact task)	"Please use the voice-controlled system to find the phone number for 'Andy'
T10 (Navigation task)	"Please use the voice-controlled system to locate bars around the University of Toronto
T11(Contact task)	"Please use the voice-controlled system to find the address for "Aubrey"
T12 (Navigation task)	"Please use the voice-controlled system to find a Starbucks nearby"

T13 (Music task)	“Please use the voice-controlled system to find music by the band ‘Pink Floyd”
T14 (Navigation task)	“Please use the voice-controlled system to find directions from here to the CN Tower”
T15 (Contact task)	“Please use the voice-controlled system to find the phone number for Tracey”
T16 (Music task)	“Please use the voice-controlled system to Find songs by artist 'Drake'”
T17 (Navigation task)	“Please use the voice-controlled system to Show directions to 'Union Station'”
T18 (Navigation task)	“Please use the voice-controlled system to Show Spas near the University of Toronto”
T19 (Navigation task)	“Please use the voice-controlled system to Show Tennis Courts in Toronto”
T20 (Contact task)	“Please use the voice-controlled system to Show contacts for Sam Patrick”
T21 (Music task)	“Please use the voice-controlled system to Show songs by the band “Rush"”
T22 (Navigation task)	“Please use the voice-controlled system to Show directions to the Airport”
T23 (Navigation task)	“Please use the voice-controlled system to Show Gas stations near the University of Toronto”
T24 (Contact task)	Please use the voice-controlled system to find a phone number for Mikey”
T25 (Music task)	“Please use the voice-controlled system to Find the song "The Queen" by "Lady Gaga"”
T26 (Navigation task)	“Please use the voice-controlled system to Find Shopping malls in Toronto”
T27 (Contact task)	“Please use the voice-controlled system to Find a phone number for Jenny”

Appendix 4: Task Order Example

Task selection and order for each condition is flexible and will be randomly determined and recorded prior to each participant, however it must meet the following conditions:

- In each condition, at least one of each task type (Map, Contact, Music) must be present.
- No tasks are repeated within a single block of drives

A label of 'S' indicated the interaction is a success for that task. A label of 'F' indicated the interaction should be a failure

Condition	Task Order
VoiceControl1_NoNoise	T1_S T2_S T10_S T3_S T9_S T4_F T4_S T5_S T7_S T6_S
VoiceControl1_Music	T2_S T3_F T3_S T1_S T4_S T10_F T10_S T9_S T6_F T6_S
VoiceControl1_ChildNoise	T4_F T4_F T4_S T5_F T5_F T5_S T1_F T1_F T1_F T1_S
VoiceControl2_NoNoise	T4_S T5_S T3_S T6_S T2_S T7_S T1 T8_F T8_S T9_S
VoiceControl2_Music	T5 _S T6_S T4_S T7_F T7_S T8_S T2_S T9_S T1_S T10_S
VoiceControl2_ChildNoise	T6_S T7_S T5_S T8_S T4_S T9_S T3_S T10_S T2_F T2_S

Appendix 5: Pre-Experiment Questionnaires

Demographics

The following are standard questions that allow researchers to determine how representative the group of participants in a study is of the general population. Remember, filling out this questionnaire is voluntary. Skipping any question that makes you feel uncomfortable will not exclude you from the study.

1. Please describe the highest level of formal education you have completed:

   a. Some high school or less
   b. High school graduate
   c. Some college
   d. College graduate
   e. Some graduate education
   f. Completed graduate or professional degree (e.g. Masters, LCSW, JD, Ph.D., MD, etc.)

2. Are you: (Please circle all that apply.)

   a. A full time student
   b. A part time student
   c. Unemployed
   d. Retired
   e. Employed full time
   f. Employed part time
   g. A full time caregiver (e.g. children or elder)
   h. A part time caregiver (e.g. children or elder)
   i. None of the above

3. Are you:

   a. Married
   b. Divorced
   c. Widowed
   d. Single living with partner
   e. Single never married
   f. Prefer not to answer

4. What best describes your total household income?

a. Less than $25,000
b. $25,000 – $49,999
c. $50,000 – $74,999
d. $75,000 – $99,999
e. $100,000 – $124,999
f. $125,000 – $149,999
g. $150,000 or more
h. I don’t know

Driving History

5. When did you obtain your first driver’s license (after your knowledge test)? (YYYY)

____

6. When did you obtain your full driver’s license? (YYYY)

____

7. On a scale of 1 to 10, with 1 being very unsafe and 10 being very safe, how safe a driver do you think you are?

1	2	3	4	5	6	7	8	9	10
Very									Very
Unsafe									Safe

8. In the past five years, how many times have you been stopped by a police officer and received a **warning** (but no citation or ticket) for a moving violation (i.e. speeding, running a red light, running a stop sign, failing to yield, reckless driving, etc.)?

   Enter a number (enter 0 for none.): ____________

9. In the past five years, how many times have you been stopped by a police officer and received a **citation or ticket** for a moving violation?

   Enter a number (enter 0 for none.): ___________

10. In the past five years, how many times have you been in a **vehicle crash** where you were the driver of one of the vehicles involved?

    Enter a number (enter 0 for none.): ___________

Technology Use

1. On a scale of 1 to 10, with 1 being very inexperienced and 10 being very experienced, how would you rate your level of experience with technology (e.g. cell phones, automatic teller machines, digital cameras, computers, etc.)?

   1 2 3 4 5 6 7 8 9 10

   Very Inexperienced — Very Experienced

2. Some people prefer to avoid new technologies as long as possible while others like to try them out as soon as they become available. In general, how would you rate yourself as being an avoider or an early adopter of new technology?

   1 2 3 4 5 6 7 8 9 10

   Avoid as long as possible — Try as soon as possible

3. On average, how often do you use an electronic navigation system in a car or truck (using a built-in navigation system, portable navigation unit or a smart phone)?

   a. More than once a day
   b. About once a day
   c. A few times a week

d. A few times a month
e. A few times a year
f. Never

4. Do you own or regularly operate a vehicle with a voice command interface system?

a. Yes

b. No

5. How often do you use vehicle based voice command interface systems?

a. More than once a day
b. About once a day
c. A few times a week
d. A few times a month
e. A few times a year
f. Never

Appendix 6: System Acceptance Questionnaire

System Acceptance

Participant ID ______________________________

System Code ______________________________

Please describe the voice control system you just used

I find such a system / the (...) system (please tick a box on every line)

		Strongly	Somewhat	Neutral	Somewhat	Strongly	
1.	Useful	○	○	○	○	○	Useless
2.	Pleasant	○	○	○	○	○	Unpleasant
3.	Bad	○	○	○	○	○	Good
4.	Nice	○	○	○	○	○	Annoying
5.	Effective	○	○	○	○	○	Superfluous
6.	Irritating	○	○	○	○	○	Likeable

7.	Assisting	○	○	○	○	○	Worthless
8.	Undesirable	○	○	○	○	○	Desirable
9.	Raising Alertness	○	○	○	○	○	Sleep-Inducing

Appendix 7A: Susceptibility to Driver Distraction Questionnaire (SDDQ)

	1. When driving, I…	Never	Rarely	Sometimes	Often	Very Often
a.	Have phone conversations.	○	○	○	○	○
b.	Manually interact with a phone (e.g., sending text messages).	○	○	○	○	○
c.	Adjust the settings of in-vehicle technology (e.g., radio channel or song selection).	○	○	○	○	○
d.	Read roadside advertisements.	○	○	○	○	○
e.	Continually check roadside accident scenes if there are any.	○	○	○	○	○
f.	Chat with passengers if you have them.	○	○	○	○	○
g.	Daydream.	○	○	○	○	○

2.	**I think, it is alright for me to drive and…**	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a.	Have phone conversations.	○	○	○	○	○
b.	Manually interact with a phone (e.g., sending text messages).	○	○	○	○	○
c.	Adjust the settings of in-vehicle technology (e.g., radio channel or song selection).	○	○	○	○	○
d.	Read roadside advertisements.	○	○	○	○	○

e.	Continually check roadside accident scenes.	○	○	○	○	○
f.	Chat with passengers.	○	○	○	○	○

3.	**I believe I can drive well even when I…**	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a.	Have phone conversations.	○	○	○	○	○
b .	Manually interact with a phone (e.g., sending text messages).	○	○	○	○	○
c.	Adjust the settings of in-vehicle technology (e.g., radio channel or song selection).	○	○	○	○	○
d .	Read roadside advertisements.	○	○	○	○	○
e.	Continually check roadside accident scenes.	○	○	○	○	○
f.	Chat with passengers.	○	○	○	○	○

4.	**Most drivers around me drive and…**	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a.	Have phone conversations.	○	○	○	○	○
b .	Manually interact with phones.	○	○	○	○	○
c.	Adjust the settings of in-vehicle technology (e.g., radio channel or song selection).	○	○	○	○	○
d .	Read roadside advertisements.	○	○	○	○	○
e.	Continually check roadside accident scenes if there are any.	○	○	○	○	○
f.	Chat with passengers if there are any.	○	○	○	○	○

5.	**Most people who are important to me think, it is alright for me to drive and…**	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a.	Have phone conversations.	○	○	○	○	○

b.	Manually interact with phones.	○	○	○	○	○
c.	Adjust the settings of in-vehicle technology (e.g., radio channel or song selection).	○	○	○	○	○
d.	Read roadside advertisements.	○	○	○	○	○
e.	Continually check roadside accident scenes.	○	○	○	○	○
f.	Chat with passengers.	○	○	○	○	○

	6. While driving, I find it distracting when…	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Never happens
a.	My phone is ringing.	○	○	○	○	○	○
b	I receive an alert from my phone (e.g., incoming text message).	○	○	○	○	○	○
c	I am listening to music.	○	○	○	○	○	○
d	I am listening to talk radio.	○	○	○	○	○	○
e	There are roadside advertisements.	○	○	○	○	○	○
f	There are roadside accident scenes.	○	○	○	○	○	○
g	A passenger speaks to me.	○	○	○	○	○	○
h	Daydreaming.	○	○	○	○	○	○

Appendix 7B: Distraction: Everyday Experience and Beliefs (Cognitive Failures)

	1. **Please answer the following questions**	Never	Rarely	Sometimes	Often	Very Often
a.	Do you read something and find you haven’t been thinking about it and must read it again?	○	○	○	○	○
b.	Do you find you forget why you went from one part of the house to the other?	○	○	○	○	○
c.	Do you fail to notice signposts on the road?	○	○	○	○	○
d.	Do you find you confuse right and left when giving directions?	○	○	○	○	○
e.	Do you have trouble making up your mind?	○	○	○	○	○
f.	Do you daydream when you ought to be listening to something?	○	○	○	○	○
g.	Do you start doing one thing at home and get distracted into doing something else (unintentionally)?	○	○	○	○	○
h.	Do you find you can’t quite remember something although it’s ‘on the tip of your tongue’.	○	○	○	○	○

Appendix 7C: Self Description – Section I (Eysneck Impulsiveness)

Please answer each question by putting a circle around the “Yes” or the “No” following the questions. There are no right or wrong answers, and no trick questions. Work quickly and do not think too long about the exact meaning of the question.

Please answer “Yes” or “No”

Would you enjoy water skiing?	Yes	No
Usually do you prefer to stick to brands you know are reliable, to trying new ones on the chance of finding something better?	Yes	No
Do you quite enjoy taking risks?	Yes	No
Would you enjoy parachute jumping?	Yes	No
Do you often buy things on impulse?	Yes	No

Do you generally do and say things without stopping to think?	Yes	No
Do you often get into a jam because you do things without thinking?	Yes	No
Do you think hitch-hiking is too dangerous a way to travel?	Yes	No
Do you like driving off the highboard?	Yes	No
Are you an impulsive person?	Yes	No

Do you welcome new and exciting experiences and sensations, even if they are a little frightening and unconventional?	Yes	No
Do you usually think carefully before doing anything?	Yes	No
Would you like to learn to fly an aeroplane?	Yes	No
Do you often do things on the spur of the moment?	Yes	No
Do you mostly speak without thinking things out?	Yes	No

Do you often get involved in things you later wish you could get out of?	Yes	No

Do you get so ‘carried away’ by new and exciting ideas, that you never think of possible snags?	Yes	No
Do you find it hard to understand people who risk their necks climbing mountains?	Yes	No
Do you sometimes like doing things that are a bit frightening?	Yes	No
Do you need to use a lot of self-control to keep out of trouble?	Yes	No

Would you agree that almost everything enjoyable is illegal or immoral?	Yes	No
Generally do you prefer to enter cold sea water gradually, to diving or jumping straight in?	Yes	No
Are you often surprised at people’s reactions to what you do or say?	Yes	No
Would you enjoy the sensation of skiing very fast down a high mountain slope?	Yes	No
Do you think an evening out is more successful if it is unplanned or arranged at the last moment?	Yes	No

Would you like to go scuba diving?	Yes	No
Would you enjoy fast driving?	Yes	No
Do you usually work quickly, without bothering to check?	Yes	No
Do you often change your interests?	Yes	No
Before making up your mind, do you consider all the advantages and disadvantages?	Yes	No

Would you like to go pot-holing?	Yes	No
Would you be put off a job involving quite a bit of danger?	Yes	No
Do you prefer to ‘sleep on it’ before making decisions?	Yes	No
When people shout at you, do you shout back?	Yes	No
Do you usually make up your mind quickly?	Yes	No

Appendix 7C: Self Description – Section II (Arnett Inventory of Sensation Seeking)

For each item, indicate how well it describes you.

1. **For each item, indicate how well it describes you.**	Very well	Somewhat	Not very well	Not at all
I can see how it would be interesting to marry someone from a foreign country.	○	○	○	○
When the water is very cold, I prefer not to swim even if it is a hot day.	○	○	○	○
If I have to wait in a long line, I am usually patient about it.	○	○	○	○
When I listen to music, I like to be loud.	○	○	○	○
When taking a trip, I think it is best to make as few plans as possible and just take it as it comes.	○	○	○	○

2. **For each item, indicate how well it describes you.**	Very well	Somewhat	Not very well	Not at all
I stayed away from movies that are said to be frightening or highly suspenseful.	○	○	○	○
I think it's fun and exciting to perform or speak before a group.	○	○	○	○
If I were to go to an amusement park, I would prefer to ride the rollercoaster or other fast rides.	○	○	○	○
I would like to travel to places that are strange and far away.	○	○	○	○
I would never like to gamble with money, even if I could afford it.	○	○	○	○

3. **For each item, indicate how well it describes you.**	Very well	Somewhat	Not very well	Not at all
I would have enjoyed being one of the first explorers of an unknown land.	○	○	○	○
I like a movie where there are a lot of explosions and car chases.	○	○	○	○
I don't like extremely hot and spicy foods.	○	○	○	○
In general, I work better when I'm under pressure.	○	○	○	○
I often like to have the radio or TV on while I'm doing something else, such as reading or cleaning up.	○	○	○	○

4. **For each item, indicate how well it describes you.**	Very well	Somewhat	Not very well	Not at all
It would be interesting to see a car accident happen.	○	○	○	○
I think it's best to order something familiar when eating in a restaurant.	○	○	○	○
I like the feeling or standing next to the edge on a high place and looking down.	○	○	○	○
If it were possible to visit another planet or the moon for free, I would be among the first in line to sign up.	○	○	○	○
I can see how it must be exciting to be in a battle during a war.	○	○	○	○

Appendix 7D: Manchester Driver Behaviour Questionnaire

Nobody is perfect. Even the best drivers make mistakes, do foolish things, or bend the rules at some time or another. For each item below you are asked to indicate HOW OFTEN, if at all, this kind of thing has happened to you. Base your judgments on what you remember of your driving. Please indicate your judgments by circling ONE of the numbers next to each item. Remember we do not expect exact answers, merely your best guess; so please do not spend too much time on any one item.

	1. How often do you do each of the following?	Never	Hardly ever	Occasionally	Quite often	Frequently	Nearly all the time
a.	Try to pass another car that is signaling a left turn.	○	○	○	○	○	○
b.	Select the wrong turn lane when approaching an intersection.	○	○	○	○	○	○
c.	Fail to ‘Stop’ or ‘Yield’ at a sign, almost hitting a car that has the right of way.	○	○	○	○	○	○
d.	Misread signs and miss your exit.	○	○	○	○	○	○
e.	Fail to notice pedestrians crossing when turning onto a side street.	○	○	○	○	○	○
f.	Drive very close to a car in front of you as a signal that they should go faster or get out of the way.	○	○	○	○	○	○
g.	Forget where you parked your car in a parking lot.	○	○	○	○	○	○

h.	When preparing to turn from a side road onto a main road, you pay too much attention to the traffic on the main road so that you nearly hit the car in front of you.	○	○	○	○	○	○
i.	When you back up, you hit something that you did not observe before but was there.	○	○	○	○	○	○
j.	Pass through an intersection even though you know that the traffic light has turned yellow and may go red.	○	○	○	○	○	○
k.	When making a turn, you almost hit a cyclist or pedestrian who has come up on your right side.	○	○	○	○	○	○
l.	Ignore speed limits late at night or very early in the morning.	○	○	○	○	○	○
m.	Forget that your lights are on high beam until another driver flashes his headlights at you.	○	○	○	○	○	○
n.	Fail to check your rear-view mirror before pulling out and changing lanes.	○	○	○	○	○	○
o.	Have a strong dislike of a particular type of driver, and indicate your dislike by any means that you can.	○	○	○	○	○	○
p.	Become impatient with a slow driver in the left lane and pass on the right.	○	○	○	○	○	○
q.	Underestimate the speed of an oncoming vehicle when passing.	○	○	○	○	○	○

r.	Switch on one thing, for example, the headlights, when you meant to switch on something else, for example, the windshield wipers.	○	○	○	○	○	○
s.	Brake too quickly on a slippery road, or turn your steering wheel in the wrong direction while skidding.	○	○	○	○	○	○
t.	You intend to drive to destination A, but you ‘wake up’ to find yourself on the road to destination B, perhaps because B is your more usual destination.	○	○	○	○	○	○
u.	Drive even though you realize that your blood alcohol may be over the legal limit.	○	○	○	○	○	○
v.	Get involved in spontaneous, or spur-of-the moment, races with other drivers.	○	○	○	○	○	○
w.	Realize that you cannot clearly remember the road you were just driving on.	○	○	○	○	○	○
x.	You get angry at the behavior of another driver and you chase that driver so that you can give him/her a piece of your mind.	○	○	○	○	○	○

Appendix 8: Informed Consent Form

Participant Consent Form

Title: Designing feedback to help induce safer driving behaviours

Investigators: Neil Sokol (416.978.0881); nsokol@mie.utoronto.ca)

Winnie Chen (416 978 0881); winnie.hy.chen@mie.uotoronto.ca)

Dr. Birsen Donmez (416.978.7399; donmez@mie.utoronto.ca)

You are being asked to take part in a research study. Before agreeing to participate in this study, it is important that you read and understand the following explanation of the proposed study procedures. The following information describes the purpose, procedures, benefits, discomforts, risks and precautions associated with this study. In order to decide whether you wish to participate in or withdraw from this research study, you should understand enough about its risks and benefits to be able to make an informed decision. This is known as the informed consent process. Please ask the investigator to explain any words you don't understand before signing this consent form. Make sure all your questions have been answered to your satisfaction before signing this document.

Purpose

This study aims to understand driver behaviour with and user acceptance of voice-control systems.

Procedure

First, you will be required to complete and sign a consent form prior to the start of the experiment. Once your consent is obtained, the experimenter will ask to see your driver's license to confirm you meet the participation requirements.

After the consent is obtained, there will be three parts to this study.

1. Introduction and Setup

You will fill out a questionnaire to provide your demographic information, as well as some information on your driving habits and familiarity with technology. You will be provided an introductory overview of the voice-control systems and the tasks you will be performing with them while driving. We will then help you attach physiological sensors to the correct location on your body and will configure the eye tracking system. The physiological sensors consist of electrocardiogram sensors on the chest, and galvanic skin response sensors on the foot. All sensors will be applied with the assistance of a researcher of the same gender as yourself.

2. Simulated Driving

a. You will first complete an introductory drive of 5 minutes in order to become accustomed to the simulator and to monitor for signs of simulator sickness.
b. After the introductory drive, you will undertake a 10 minute practice drive in order to familiarize yourself with the driving environment, the voice control system, and various events which can occur in the simulated environment. During the introductory and practice drives, researchers will answer any questions or concerns you may have about the experimental setup or tasks.
c. You will then complete 3 sets of 3 experimental drives; in 2 of these sets you will also be interacting with voice-control systems. Each set will take approximately 18 minutes (approx. six minutes per drive), and there will be 5 minute breaks in between sets. During the breaks you will be asked to fill out a short questionnaire describing your experience with the voice control systems.

We ask that you treat the simulation just like you were driving your own car, thinking of all elements of the simulation as if they were encountered in the real world. Multiple cameras will record your drive from various angles during this phase.

3. Post-drive Questionnaires

Once the drives are completed, you will be asked to fill out a final set of questionnaires concerning your driving habits and personality.

Risks

There are no major risks involved with this experiment, the tasks are not physiologically demanding or psychologically stressing. We want to make you aware of the possibility of simulator sickness (a form of motion sickness specific to simulators), however. Especially upon first using a driving simulator, there is a small chance of feeling dizzy, nauseous, or fatigued. If you feel any of these symptoms appear, please immediately stop the experiment and inform the investigator. The investigator will also monitor for any signs of simulator sickness.

Benefits

There are several benefits to conducting this study. The most important benefit is your contribution to research on in-car technologies, which will guide the development of new methods of automobile system interaction. You will also gain experience with academic research and be able to use and test out a state of the art driving simulator.

Compensation

You will receive $15/hr for your participation plus a possible task performance bonus at the end of this study of up to $5.

Confidentiality

All information obtained during the study will be held in strict confidence. You will be identified with a study number only, and this study number will only be identifiable by the primary investigator. No names or identifying information will be used in any publication or presentation. No information identifying you will be transferred outside the investigators in this study.

Please be advised that we video-record the experimental trials with four small web-cameras. One camera will be pointed at you, one will capture the steering wheel, one the pedals, and the final camera the overall scene. We will use four other cameras on and near the dashboard to track and record where you are looking during the experiment. The videos will only be seen by the investigators, the primary investigator's research assistant, and research collaborators. Faces will be blurred in any video used in public presentations. Audio recordings of your interaction with the voice control systems will also be made. In any public presentation, we will obscure your voice to maintain confidentiality.

You will be asked to fill out several questionnaires regarding your driving behavior, including possibly illegal activities such as speeding. Your responses to these questions will be held in strict confidentiality and no information from these questionnaires will be shared with any government or police authority.

The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team.

Participation

Your participation in this study is voluntary. You can choose to not participate or withdraw at any time and still be compensated at a pro-rated basis of $15/hr for your participation to that point. Furthermore you can also choose to skip survey questions with no penalty.

Questions

If you have any general questions about this study, please call 416.978.0881 or email nsokol@mie.utoronto.ca.

If you have questions concerning how you have been treated as a research participant, please contact the Office of Research Ethics at ethics.review@utoronto.ca or 416 946 3273.

Consent

I have had the opportunity to discuss this study and my questions have been answered to my satisfaction. I consent to take part in the study with the understanding I may withdraw at any time. I have received a signed copy of this consent form. I voluntarily consent to participate in this study

______________________________   ____________________   ____________________

Participant’s Name (please print)   Signature   Date

I confirm that I have explained the nature and purpose of the study to the participant named above. I have answered all questions.

______________________________ ____________________ ____________________

Investigator’s Name Signature Date