Effects of Modality, Urgency and Situation on Responses to Multimodal Warnings for Drivers

Ioannis Politis

Submitted in Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

College of Science and Engineering
School of Computing Science
University of Glasgow

December 2016

Abstract

Signifying road-related events with warnings can be highly beneficial, especially when imminent attention is needed. This thesis describes how modality, urgency and situation can influence driver responses to multimodal displays used as warnings. These displays utilise all combinations of audio, visual and tactile modalities, reflecting different urgency levels. In this way, a new rich set of cues is designed, conveying information multimodally, to enhance reactions during driving, which is a highly visual task. The importance of the signified events to driving is reflected in the warnings, and safety-critical or non-critical situations are communicated through the cues. Novel warning designs are considered, using both abstract displays, with no semantic association to the signified event, and language-based ones, using speech. These two cue designs are compared, to discover their strengths and weaknesses as car alerts. The situations in which the new cues are delivered are varied, by simulating both critical and non-critical events and both manual and autonomous car scenarios. A novel set of guidelines for using multimodal driver displays is finally provided, considering the modalities utilised, the urgency signified, and the situation simulated.

Acknowledgements

I would like to thank my supervisor, Professor Stephen Brewster, for his continuous support and inspiration during this work. I would also like to thank my second supervisor, Professor Frank Pollick, for his involvement and sincere interest in the research.

I would like to thank Derek Beattie and Andrew Birnie from Freescale Semiconductor (now NXP) for the constructive discussions and useful suggestions during the PhD. My visits to East Kilbride site offered me insights in how research can be applied in the industry.

I thank Dr Gary Burnett and Dr Helen Purchase for their insightful comments on the work during my viva, leading to a set of thesis corrections.

I thank my colleagues in Multimodal Interaction Group, and also past members of the group, for their help with ideas, comments and resources throughout the PhD, and for the pleasant conversations on various topics, whether related to HCI or not: Dr Graham Wilson, Dr Mark McGill, Dr Dong-Bach Vo, Dr Alexander Ng, Dr Euan Freeman, Dr Matthew Jamieson, Dr Ross McLachlan, Dr Marilyn McGee-Lennon, Dr Shimin Feng, Dr David Warnock, Christopher McAdam, Professor David McGookin, Dr Julie Williamson, Gözel Shakeri, Rui Huan, Noor Izdhora, Guillaume Aquilina.

I thank the people closest to me, my family and friends. I am grateful to my beloved wife, Flora, for her love, encouragement and understanding at all times. I am grateful to my beloved parents, Maria and Takis, for having taken good care of me, and for being loving and supportive. I am thankful to my old and new friends for being unconditional in their friendship and sincere in their advice: Fotis, Pol, Paula, Rafal, Miro, Beant, Dali, Christos, Michalis, Sofia, Leonidas, Lida, Meili, Jose, Manos M., Zach, Antonis, Manos T., Emily, Neil, Stuart, Heather, Iraklis, Petros, Panos, Jane.

I am finally thankful to the longing for truth.

This thesis is dedicated to my daughter, expected to arrive in April 2017.

This research was partly funded by Freescale Semiconductor Inc. (now NXP), Automotive Microcontroller Product Group.

Declaration & Contributing Publications

The research presented in this thesis is entirely the author's own work. This thesis exploits only the parts of these papers that are directly attributable to the author:

Experiments 1 and 2 from Chapter 3 have been published in AutomotiveUI 2013:

Politis, I., Brewster, S., & Pollick, F. (2013). Evaluating Multimodal Driver Displays of Varying Urgency. In *AutomotiveUI 2013* (pp. 92–99). ACM Press.

Experiment 3 from Chapter 4 has been published in CHI 2014:

Politis, I., Brewster, S., & Pollick, F. (2014). Evaluating Multimodal Driver Displays under Varying Situational Urgency. In *CHI 2014* (pp. 4067–4076). ACM Press.

Experiments 4 and 5 from Chapter 5 have been published in AutomotiveUI 2014:

Politis, I., Brewster, S., & Pollick, F. (2014). Speech Tactons Improve Speech Warnings for Drivers. In *AutomotiveUI 2014* (pp. 1–8). ACM Press. Best Paper Award.

Experiments 6 and 7 have been published in CHI 2015:

Politis, I., Brewster, S., & Pollick, F. (2015). To Beep or Not to Beep? Comparing Abstract versus Language-Based Multimodal Driver Displays. In *CHI 2015* (pp. 3971–3980). ACM Press.

Experiments 8 and 9 have been published in AutomotiveUI 2015:

Politis, I., Brewster, S., & Pollick, F. (2015). Language-Based Multimodal Displays for the Handover of Control in Autonomous Cars. In *Automotive UI 2015* (pp. 3–10). ACM Press.

Experiment 10 is in press, to be published in the International Journal of Mobile Human Computer Interaction:

Politis, I., Brewster, S., & Pollick, F. (2016). Using Multimodal Displays to Signify Critical Handovers of Control to Distracted Autonomous Car Drivers. In *International Journal of Mobile Human Computer Interaction* (In Press).

Summaries of this research have been published in Doctoral Consortia in CHI 2014, and AutomotiveUI 2014:

Politis, I. (2014). A Beep, a Flash, a Rumble? Evaluating Multimodal Displays for Drivers. In *CHI 2014 Extended Abstracts* (pp. 303–306). ACM Press.

Politis, I. (2014). The Effects of Modality, Urgency and Message Content on Responses to Multimodal Driver Displays. In *AutomotiveUI 2014 Adjunct Proceedings* (pp. 1–5). ACM Press.

Finally, ideas developed in this thesis have contributed to a series of workshops on User Experience for Autonomous cars, in AutomotiveUI 2014, CHI 2015, AutomotiveUI 2015 and CHI 2016, co-organised by the author:

Meschtscherjakov, A., Tscheligi, M., Szostak, D., Ratan, R., McCall, R., Politis, I., & Krome, S. (2014). 2nd Workshop on User Experience of Autonomous Driving. In *AutomotiveUI 2014 Adjunct Proceedings* (pp. 1–3). ACM Press.

Meschtscherjakov, A., Tscheligi, M., Szostak, D., Ratan, R., McCall, R., Politis, I., & Krome, S. (2015). Experiencing Autonomous Vehicles: Crossing the Boundaries between a Drive and a Ride. In *CHI 2015 Extended Abstracts* (pp. 2413–2416). ACM Press.

Meschtscherjakov, A., Tscheligi, M., Szostak, D., Ratan, R., McCall, R., Politis, I., Krome, S., Riener, A., Jeon, M. & Terken, J. (2015). 3rd Workshop on User Experience of Autonomous Driving. In *Automotive UI 2015 Adjunct Proceedings* (pp. 1–3). ACM Press.

Meschtscherjakov, A., Tscheligi, M., Szostak, D., Krome, S., Ratan, R., Pfleging, B., Politis, I., Baltodano, S., Miller, D., Ju,W. (2016). HCI and Autonomous Vehicles: Contextual Experience Informs Design. In *CHI 2016 Extended Abstracts* (pp. 3542–3549). ACM Press.

Education Use Consent

I hereby give my permission for this thesis to be shown to other University of Glasgow students and to be distributed in an electronic format.

Ioannis Politis

Table of contents

List of Tables

List of Figures

Definitions

A set of definitions of specialist concepts used in this thesis is provided below, taken form ("Cambridge Dictionary," 2016; Freeman, Wilson, Vo, Politis, & Brewster, 2017).

Auditory Icons	Caricatures of natural sounds occurring in the real world, used to represent information from a computer interface.
Earcons	Structured abstract audio messages, made from rhythmic sequences called motives. Motives are parameterized by audio properties like rhythm, pitch and timbre.
Haptic	A sensation coming from the skin (cutaneous), also referred to as tactile.
Modality	A particular way of doing or experiencing something.
Multimodal	Utilising multiple modalities.
Tactile	Rendering a percept of the cutaneous sense, for example, using vibration (vibrotactile), temperature, texture, or other material properties to encode information.
Tactons	Structured abstract tactile messages that use properties of vibration to encode information, also referred to as tactile icons.
Urgency	The state of being very important and needing to take action immediately.

Glossary of Acronyms

A	Audio
AT	Audio + Tactile
ATV	Audio + Tactile + Visual
AV	Audio + Visual
CD	Car to Driver
DC	Driver to Car
LatDev	Lateral Deviation
LDaH	Lateral Deviation after Handover
LDU	Level of Designed Urgency
L_H	Level High (referring to LDU)
L_L	Level Low (referring to LDU)
L_M	Level Medium (referring to LDU)
NoM	Number of Modalities
PA	Perceived Annoyance
PAE	Perceived Alerting Effectiveness
PU	Perceived Urgency
RecA	Recognition Accuracy

RecT	Recognition Time
ResA	Response Accuracy
ResT	Response Time
RMSE	Root Mean Square Error
SteAng	Steering Angle
T	Tactile
TV	Tactile + Visual
V	Visual

1. Introduction

1.1. Motivation

Driving is a highly dynamic task, demanding in terms of attentional resources (Endsley, 1995; C. D. Wickens, 1980, 1992; Christopher D. Wickens, 2002). When operating a vehicle, a driver needs to make decisions and act based on the continuously changing state of the environment (Endsley, 1995). Performance in this task is influenced by the attentional resources at the driver's disposal, which involve the processing of surrounding information, the modalities used to achieve this processing as well as the resulting responses of the driver (C. D. Wickens, 1980, 1992; Christopher D. Wickens, 2002). When there is a bottleneck in the above process, driving workload increases (Horrey & Wickens, 2004). Such an effect can be catastrophic, leading to increased fatalities on the road (Redelmeier & Tibshirani, 1997; F. A. Wilson & Stimpson, 2010). Warnings can be an effective means to attract attention back to the road when needed (John D. Lee, McGehee, Brown, & Reyes, 2002; Liebermann, Ben-David, Schweitzer, Apter, & Parush, 1995; Schweitzer, Apter, Ben-David, Liebermann, & Parush, 1995).

Multimodal displays have been used as warnings in a variety of driving studies to address this problem (Christy Ho & Spence, 2008). With the term "multimodal", one refers to warnings displayed across several feedback channels of sensory communication. In more detail, "unimodal" refers to a single channel (*e.g.* a sound), "bimodal" to two channels (*e.g.* a sound and a visual signal), "trimodal" to three channels (*e.g.* a sound, a visual signal and a vibrational signal), *etc*. Modalities utilised in driving studies include sound (Graham, 1999; Robert Gray, 2011; Cristy Ho & Spence, 2005; Sullivan & Buonarosa, 2009), visuals (Ablaßmeier, Poitschke, Wallhoff, Bengler, & Rigoll, 2007; Inuzuka, Osumi, & Shinkai, 1991; Liebermann et al., 1995; Medenica, Kun, Paek, & Palinko, 2011), vibration (Enriquez, Afonin, Yager, & Maclean, 2001; Cristy Ho, Tan, & Spence, 2005; Hogema, De Vries, Van Erp, & Kiefer, 2009; A. Riener, Zia, Ferscha, Ruiz Beltran, & Minguez Rubio, 2012) and combinations of these (Erp & Veen, 2001; Cristy Ho, Spence, & Tan, 2005; Lindgren, Angelelli, Mendoza, & Chen, 2009; Mollenhauer, Lee, Cho, Hulse, & Dingus, 1994). However, there has been less consideration of how to use the warnings multimodally to signify events of varying urgency, although there are guidelines on how to design warnings of varying urgency outside the driving context (C. L. Baldwin et al., 2012; Chapanis, 1994;

Judy Edworthy, Loxley, & Dennis, 1991b; E. Hellier, Weedon, Edworthy, & Walters, 2000; Elizabeth Hellier, Edworthy, Weedon, Walters, & Adams, 2002; B. A. Lewis & Baldwin, 2012; Bridget A Lewis, Eisert, & Baldwin, 2014). This is important, because events on the road are not always equally critical, for example an impending collision versus an incoming message. A significant research step is therefore to combine the above notions and design truly multimodal displays, using all combinations of audio, visual and tactile modalities, varying in urgency and evaluate them as warnings. In this way, the effects of combining modalities in vehicle alerts will be identified, the applicability of urgency design guidelines will be evaluated in a multimodal setting, and better alerts will be designed. This motivates the research question of how multimodal displays varying in urgency affect performance.

In terms of warning design, previous studies have evaluated speech (Lai, Cheng, Green, & Tsimhoni, 2001; Serrano, Di Stasi, Megías, & Catena, 2011), abstract messages (Erp & Veen, 2001; Cristy Ho, Tan, et al., 2005) or other message designs semantically associated with the signified event (Cristy Ho & Spence, 2005; Denis McKeown & Isherwood, 2007). However, there has been no direct comparison between cue designs in all multimodal combinations of warnings. This is partly because warnings have not been effectively synchronised in all modalities, especially speech warnings. Using speech on the tactile modality by retaining some aspects of speech rhythm and intensity could reveal new advantages of the resulting cues. Therefore, two opportunities arise from this fact. Firstly, one can design truly multimodal warnings that vary in urgency and message content, by transferring abstract pulses and speech to all combinations of audio, visual and tactile cues. The transferring of speech to vibration is a particularly novel and less explored aspect (Salminen et al., 2012; Spens, Huss, Dahlgvist, & Agelfors, 1997), which one needs to solve first to achieve the above transfer. Secondly, one can compare the resulting warnings multimodally and derive guidelines on the use of multimodal warnings varying in message content. This topic is also less explored, with limited available studies that attempt comparisons unimodally (Carryl L Baldwin & May, 2014; J. Edworthy, Walters, Hellier, & Weedon, 2000). Creating truly multimodal alerts of varying urgency and message content and comparing their effectiveness will further inform warning design, by evaluating a much wider and more flexible set of messages that can possibly be interchangeable. This motivates the research question of how abstract and language-based multimodal displays varying in urgency compare to each other in terms of performance.

Warnings can be used in a variety of situations, which can be more or less critical. Other than designing warnings that vary in urgency, one can directly observe the added benefit of warnings when delivering them in a set of contexts varying in criticality, or contexts of varying situational urgency. This notion has been partly explored in previous studies, by investigating the added benefit of warnings when delivered in the presence or absence of a critical event (false alarms) (Lees & Lee, 2007; Maltz & Shinar, 2004, 2007) or by evaluating responses to a critical event with warnings or without warnings (Chun et al., 2012; Cristy Ho, Reed, & Spence, 2007; Scott & Gray, 2008). However, these evaluations did not use all multimodal combinations of warnings, while the above cases were not combined in the same study, *i.e.* false alarms, critical events with absence of warnings and critical events with warnings. Such an investigation would clarify the added benefit of warnings in all possible variations of situational urgency and confirm that they are actually useful in improving reactions. This motivates the research question of how does situational urgency affect responses to multimodal warnings varying in urgency.

Autonomous cars are gaining in popularity (Kyriakidis, Happee, & Winter, 2014), which motivates research on how driving an autonomous car is different to driving a manual car (Brandenburg & Skottke, 2014; de Winter, Happee, Martens, & Stanton, 2014). Since driving involvement when operating an autonomous car is fundamentally different to a manual car, there are new implications on how to design safe interfaces for such vehicles. One particularly relevant aspect of the interaction are the points where control is transferred between the car and driver: the handovers of control. This is an important scenario, since autonomy in cars is not yet complete and will not be complete without a transition to partial autonomy first (National Highway Traffic Safety Administration, 2013; SAE J3016 & J3016, 2014). Warnings for handovers of control have been designed (Eriksson, Marcos, Kircher, Västfjäll, & Stanton, 2015; Koo et al., 2014; Naujoks, Mai, & Neukum, 2014), however the scenarios that would demand such warnings are underexplored. This is essential to investigate, since such scenarios can vary in criticality, which should also be reflected by the warnings designed for these scenarios. The influence of modality and urgency content is also rarely explored in this context (Naujoks et al., 2014), while the influence of message content has not been explored. Using available urgency design guidelines and warning designs to signify handovers of control will extend knowledge in this new use case (Kyriakidis et al., 2014; Schoettle & Sivak, 2014). Therefore, the research question of how multimodal driver displays varying in urgency and message content affect performance during autonomous car handovers is motivated by this lack of research.

This thesis answers the above questions by investigating a set of multimodal driver displays varying in urgency in a series of experiments. Initially, it investigates the use of abstract multimodal warnings in terms of how they are perceived subjectively, as well as how participants react to them. These warnings are then investigated in varying situational urgency contexts, in order to confirm the utility of the cues and their added benefits in driving. Cue design is the next topic in question, and a set of language-based warnings varying in urgency across all modalities is designed and evaluated in terms of subjective responses and objective reactions. As a next step, the two types of cues are compared to each other in order to better understand their relative performance as driving alerts. The cues are finally redesigned to fit the context of autonomous cars and evaluated in that context, addressing scenarios of control transfer (handovers) between car and driver. This scenario is simulated in different urgency levels, since handovers are expected to vary in terms of time criticality. This further extends the investigation of this thesis in terms of cue modality and design, to conclude with a set of guidelines that relate to both manual and autonomous car driving scenarios.

A simulated driving task is used in the studies of this thesis. This is a widely used practice, found in the majority of studies investigating critical road events (see Chapter 2). This approach was chosen due to practical limitations of studying critical situations in real driving (see also Section 3.1 and Section 8.6.1). This is an acceptable approach in the research community, and was selected as a safe way to study safety related events in a controlled environment.

1.2. Thesis Statement

Driver displays are essential in capturing attention. This thesis investigates the use of multimodal displays as warnings, utilizing all combinations of audio, visual and tactile modalities varying in urgency and message content. The warnings are tested in situations of varying urgency, as well as in the context of autonomous cars, where control is transferred between the car and the driver. Novel guidelines are thus provided in a range of contexts on the use of multimodal displays as warnings.

1.3. Research Questions

This thesis aims to answer the following research questions:

- *RQ-1: How do multimodal driver displays varying in urgency affect performance?* (Experiments 1 and 2)
- *RQ-2: How does situational urgency influence responses to multimodal driver displays varying in urgency?* (Experiment 3)
- *RQ-3: How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?* (Experiments 4, 5, 6 and 7)
- *RQ-4: How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?* (Experiments 8, 9 and 10)

1.4. Thesis Outline

Chapter 2 presents a literature review on the topic of multimodal driver displays. The use of alerts as a strategy to attract attention is discussed, based on theories on Situation Awareness, Multiple Resources and Signal Detection. A set of studies investigating performance of audio, visual, tactile and multimodal displays is then presented, highlighting the potential of using such displays to improve reaction in driving events. Available guidelines on how to vary perceived urgency of warnings are reviewed, concluding that they are not fully implemented in multimodal driver displays. A set of studies comparing warnings varying in semantic content are also presented, highlighting the fact than comparisons between multimodal displays varying in message content are lacking. A presentation of available work on the influence of situational urgency follows, identifying a research potential of varying the urgency of the events as well as the urgency of the warnings to evaluate multimodal displays. Finally, studies on the use of handovers of control in autonomous cars are presented, concluding that neither the situations requiring a handover nor the appropriate multimodal warnings to signify it have been fully investigated.

Chapter 3 presents the general experimental methodology of this thesis, and discusses the use of driving simulators in research, the driving metrics used in the studies of this thesis, and the use of participants in the experiments conducted.

Chapter 4 presents Experiments 1 and 2, answering RQ-1. A set of abstract multimodal displays along all unimodal, bimodal and trimodal combinations of audio, visual and tactile modalities were designed using available guidelines, and their perceived urgency and perceived annoyance were assessed (Experiment 1). The warnings were then tested in a driving simulator in terms of recognition time and accuracy, to investigate performance of participants when identifying the urgency of the cues (Experiment 2). In this way, both subjective and objective measures were used to evaluate the designed warnings and provide guidelines on the use of abstract multimodal driver displays varying in urgency. See Table 1-1 for an overview of the thesis research questions, the experimental research questions (constructed to answer the thesis research questions) the dependent variables, and the experimental conditions of the experiments presented in this thesis. See Figure 1-1 for a framework of the experiments described in this thesis, in terms of the displays, driving conditions, subjective and objective measures examined.

Chapter 5 presents Experiment 3, answering RQ-2. The warnings of Experiments 1 and 2 were evaluated along different contexts of situational urgency: when presented along with a critical event in the driving simulator, when presented without the critical event and when the event was presented without warnings. Reaction time and accuracy to the above presentations were measured. In this way, objective measures regarding the usage of the warnings in situations varying in urgency were taken, confirming the utility of the cues and producing further guidelines on situational urgency in multimodal driver displays.

Chapter 6 presents Experiments 4 and 5, and Chapter 7 presents Experiments 6 and 7, all answering RQ-3. Chapter 6 describes the design of a set of language-based warnings varying in urgency. A set of tactile cues (speech Tactons) were designed, based on speech and retaining rhythmic and intensity features of speech. They were also varied in urgency according to available guidelines on urgency of speech messages and tactile messages. They were then evaluated in terms of perceived urgency, annoyance and alerting effectiveness (Experiment 4), as well as in terms of recognition time and accuracy (Experiment 5). In this way, an effective cue design for language-based warnings was identified, which was then used in comparison with abstract cues, in Chapter 7. In this chapter, multimodal cues varying in urgency and message content were compared in a driving simulator, in contexts varying in criticality. Their recognition time and accuracy was first compared in a driving simulator with a non-critical driving task, *i.e.* with no critical event present (Experiment 6). Their reaction time and accuracy was then compared when the cues were signifying a critical event (Experiment 7). In this way, the cues designed were evaluated in a broad set of contexts both with subjective and objective measures, and new guidelines on the use of truly multimodal cues varying in urgency and message content were created.

Chapter 8 presents Experiments 8, 9 and 10, answering RQ-4. A set of language-based cues, able to convey rich information on a situation, were designed for the context handovers of control in autonomous vehicles. Using the process developed in earlier experiments, and envisioning new situations of varying urgency where a handover could occur, a set of language-based warnings also varying in urgency was designed. These were first evaluated in terms of perceived urgency, annoyance and alerting effectiveness (Experiment 8). They were then used in a driving simulator to signify handovers of control varying in urgency and evaluate responses to the cues during these handovers (Experiment 9). In the particularly critical case of automation failures, both abstract cues designed in previous experiments and language-based cues designed in Experiment 10 were compared, in order to identify the best warnings to signify an automation failure. In this way, a set of warnings varying in urgency and message content were designed and evaluated with subjective and objective measures, to derive guidelines on warning design for this unexplored scenario.

Finally, Chapter 9 presents a general discussion of the findings of this thesis, a set of guidelines deriving from the experimental results, and a set of contributions provided by this work, as well as general conclusions of this thesis.

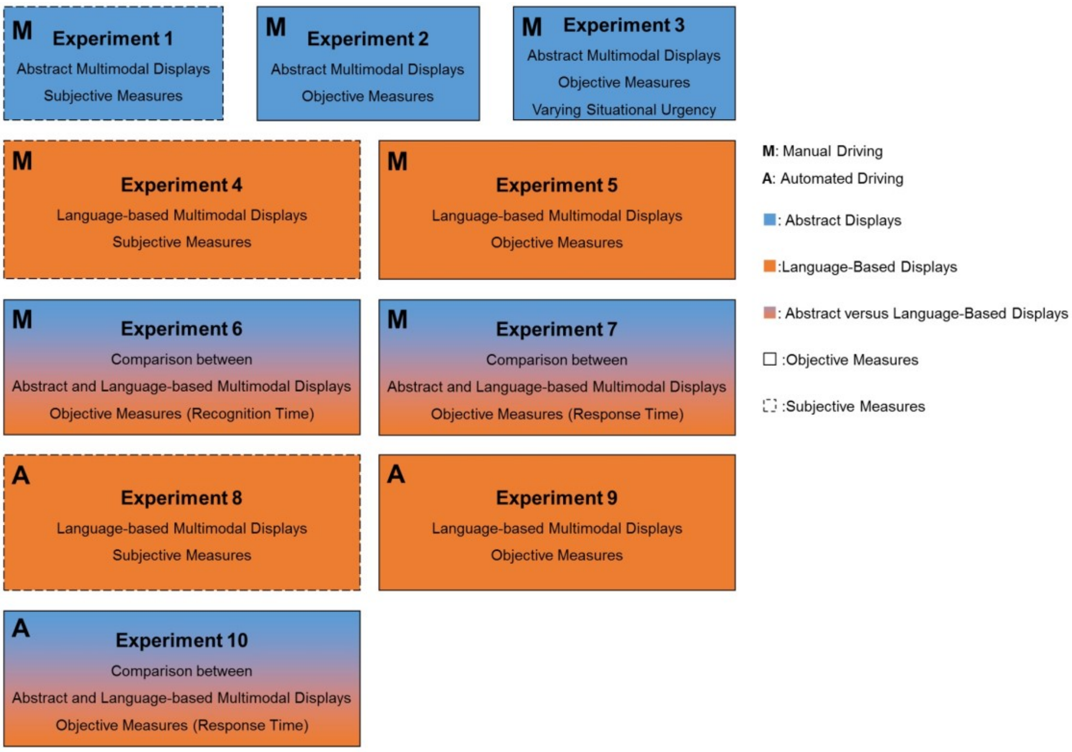


Figure 1-1: A framework of the experiments described in this thesis.

Experiment - Chapter	**Thesis Research Question**	**Experimental Research Question, Answering Thesis Research Question**	**Dependent Variables (Experimental Conditions)**
1 - 4	RQ-1: How do multimodal driver displays varying in urgency affect performance? (Experiments 1 and 2)	What is the perceived urgency and annoyance of multimodal driver displays varying in designed urgency?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)*
2 - 4		How quickly and accurately can multimodal driver displays varying in designed urgency be identified?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)*
3 - 5	RQ-2: How does situational urgency influence responses to multimodal driver displays varying in urgency? (Experiment 3)	How quick and accurate are reactions to multimodal driver displays varying in designed urgency, when delivered in situations of varying urgency?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)* Situational Urgency *(Three situations, including a critical event with warnings, the warnings without the critical event, and the critical event without the warnings)*
4 - 6	RQ-3: How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance? (Experiments 4, 5, 6 and 7)	What is the perceived urgency, annoyance and alerting effectiveness of language-based multimodal driver displays varying in designed urgency and tactile design?	Modality *(Audio, Tactile and Audio-Tactile)* Designed Urgency *(Six messages of varying designed urgency)* Tactile Design *(Four designs, where intensity, roughness, both intensity and roughness, and neither intensity nor roughness were introduced to the tactile cues)*
5 - 6		How accurately are language-based tactile driver displays varying in designed urgency and tactile design perceived?	Designed Urgency *(Six messages of varying designed urgency)* Tactile Design *(Four designs, where intensity, roughness, both intensity and roughness, and neither intensity nor roughness were introduced to the tactile cues)*

Experiment - Chapter	Thesis Research Question	Experimental Research Question, Answering Thesis Research Question	Dependent Variables (Experimental Conditions)
6 - 7		How do abstract versus language based multimodal driver displays varying in designed urgency compare in terms of recognition time and accuracy in a non-critical driving context?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)* Message Content *(Abstract cues and language-based cues)*
7 - 7		How do abstract versus language based multimodal driver displays varying in designed urgency compare in terms of reaction time and accuracy in a critical driving context?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)* Message Content *(Abstract cues and language-based cues)*
8 - 8	RQ-4: How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car? (Experiments 8, 9 and 10)	What is the perceived urgency, annoyance and alerting effectiveness of language-based multimodal driver displays varying in designed urgency, designed for critical and non-critical handovers of control in autonomous cars?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)* Situation *(Three levels of handover criticality in an autonomous car)*
9 - 8		How quick and accurate are reactions to language based multimodal driver displays varying in designed urgency, in critical and non-critical handovers in autonomous cars?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Designed Urgency *(Three levels of increasing designed urgency)*
10 - 8		How do abstract versus language based multimodal driver displays varying in urgency and location compare in terms of reaction time and accuracy during critical handovers in autonomous cars?	Modality *(Audio, Visual, Tactile and all their bimodal and trimodal combinations)* Message Content *(Abstract cues and language-based cues)* Location *(From a driving simulator or a tablet)*

Table 1-1: Overview of the of the thesis research questions, the experimental research questions (constructed to answer the thesis research questions) the dependent variables, and the experimental conditions of the experiments presented in this thesis.

2. Literature Review

Cars are becoming increasingly advanced in the technologies they provide. In parallel, technologies outside the car are becoming increasingly mobile, enabling drivers to carry smart devices in their pockets and in their cars. This creates a potential for multiple sources of distraction, which can lead to critical situations. The use of audio, visual tactile and multimodal displays has been shown to provide benefits when driver attention needs to be attracted. There has been consideration of warning design, to convey the appropriate degree of urgency, as well as the warning's semantic content, to better communicate desired messages. Further, implications of presenting the warnings in various road situations and with various levels of car automation have been examined.

This chapter reviews available literature on the above topics, and explains the research opportunities identified by this review. In the following paragraphs a description of the topics reviewed is provided, and how they relate to the research questions of this thesis. Note that the review follows a different order compared to the research questions. It first presents literature related to warning design and comparisons of different warning types, and then introduces other topics such as situational urgency (which was addressed in an experiment before different warning designs were compared), and warnings for autonomous cars. This presentation was followed in order to discuss more related topics together.

The initial topic of the literature review is the attentional demands of the driving task and how alarms can be beneficial in driving (Section 2.1). This section does not directly relate to a research question of this work, but answers a more primary question, which was the motivation of the work:

- *RQ-0: Why are driver warnings useful?*

To attract attention and assist the completion of the desired task, studies using multimodal warnings for drivers are reviewed (Section 2.2). Since urgency is a decisive factor on the relevance of the warnings, current literature on how to design urgency in multimodal displays is then presented (Section 2.3). These sections highlight the research space for investigating the use of all multimodal combinations of audio, visual and tactile abstract warnings varying in urgency to alert drivers. This investigation would improve available warning mechanisms and effectiveness of attracting driver attention when signifying scenarios of varying importance. Sections 2.2 and 2.3 motivate the research question:

- *RQ-1: How do multimodal driver displays varying in urgency affect performance?*

The semantic content of the warnings has been shown to fundamentally influence reactions. Therefore, a review of abstract and language-based warnings is then provided, being two prevalent warning categories found in driver displays (Section 2.4). There is a distinct lack of work on using multimodal combinations of language-based displays to alert drivers, as well as comparing the effectiveness of abstract and language-based warnings to each other. Therefore, there is opportunity in investigating these topics, providing new guidelines on abstract and language-based warning utility and designing better alerts. Section 2.4 motivates the research question:

- *RQ-3: How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?*

The road is a highly dynamic environment that can alter how drivers react to car warnings. A set of studies considering the influence of criticality of road events (situational urgency) to drivers' reactions to warnings is therefore presented (Section 2.5). This review highlights the absence of work in evaluating the warnings in all variations of situational urgency. That is, when there is a critical event and warnings, when there are warnings with no critical event and when there is a critical event without warnings. In this way the influence of the environment to reactions and the actual value of warnings will be better understood. This section motivates the research question:

- *RQ-2: How does situational urgency influence responses to multimodal driver displays varying in urgency?*

Since the recent development of autonomous cars creates new challenges for keeping drivers attentive and warning them when their intervention is required, a review of relevant studies is provided (Section 2.6). This presentation concludes that studying warning mechanisms to alert drivers about control handovers between them and the vehicle is a promising research opportunity, which is not investigated. Investigating this topic will enhance understanding in this new domain and lead to safer autonomous cars. It motivated the research question:

- *RQ-4: How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?*

2.1. Understanding the Purpose of Driver Warnings

This section is not directly tied to a research question of this thesis, however it is used as a means of motivating the work and providing some background on the demands of driving as a dynamic task, as well as the attentional resources required for it, initially presenting the notion of Situation Awareness (Endsley, 1995) and then the Multiple Resources Theory, a widely accepted model of attentional resources (Christopher D. Wickens, 2002). To uncover the mechanisms taking place when an alarm is presented to a user, and the potential outcomes of this process, an overview of the Signal Detection Theory is presented (D. Green & Swets, 1966). Finally, with the purpose of managing the demands of the driving task and attracting attention, being the main motivation of this thesis, the utility of alarms as a Human Factors strategy is presented (Stanton, 1994). Contrary to the rest of the literature review, where research questions will be posed after presenting the relevant work, the question related to this section will be posed before, since it motivated the work:

- *RQ-0: Why are driver warnings useful?*

2.1.1. Situation awareness

As defined by Endsley (Endsley, 1995), "situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future". Situation Awareness (SA) is essential when interacting with dynamic systems. In her work, Endsley describes how SA helps form the basis for decision making when operating an aircraft, controlling air traffic, and also driving. Figure 2-1 depicts the model presented by Endsley, where the state of the environment is used as an input for SA, which will lead to decision and action. The results of the action will update the state of the environment and the same process can be repeated. This process is informed by factors relating to the task or system, but also individual factors. A brief description of Endsley's model will follow.

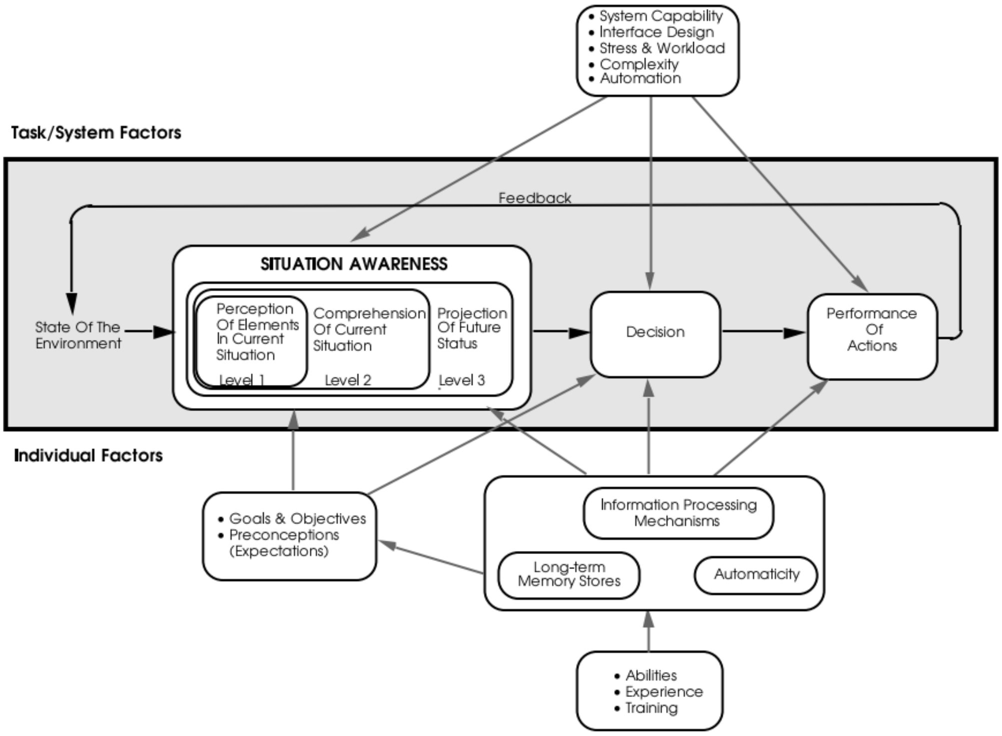


Figure 2-1: Model of situation awareness in dynamic decision making. Image taken from (Endsley, 1995).

The first stage of SA (Level 1 SA) is the perception of the environment one operates in. For driving, one should be aware of the location of other vehicles, obstacles, as well as the dynamics of the vehicle. The salience of cues related to the above will affect the resulting SA. Comprehending the situation (Level 2 SA) is the next essential stage towards decision making. Forming a holistic picture of the environment and understanding the significance of objects and events are part of this stage. For driving, one could detect that a vehicle in front is approaching rapidly and is on a collision course. Finally, projecting the future status of the elements of the environment (Level 3 SA) is the highest level of SA. For driving, one needs to detect that the critical situation described could lead to a collision, and act accordingly. Poor response to a situation may arise from poor SA.

As mentioned above, SA is influenced by individual factors when interacting with dynamic systems. It is required in the context of goals and expectations for a situation, for example to drive home safely. Further, factors related to the abilities and experience of the individual will influence the way information is processed and recalled, and how effortless is the response to a situation. Regarding task and system factors that influence SA, Endsley mentions the capabilities of a system, which are reflected in the interface design of that system, as well as the complexity, stress and level of automation related to a particular task.

The task and systems factors that affect SA are not without significance for the driving task. Interpreting Endsley's pointers, one can see a continuously increasing level of complexity of cars as systems, which has strong implications for their interface design. As mentioned in (Endsley, 1995), interface displays need to provide information rather than data, in a way that is relevant and easily understandable by operators. The risk of information overload by failing to manage what aspect of a complex system is visible to the user cannot be neglected. Since highly dynamic tasks, such as driving, can increase workload, maintaining SA through the interface cues is essential. To summarise, in complex and dynamic environments, the decisions made and responses executed utilising SA can be affected by high attentional demands. To explore this further, the next section will present a prevalent theory on attentional resources, Multiple Resources Theory.

2.1.2. Multiple Resources Theory

Multiple Resources Theory, proposed by Wickens (C. D. Wickens, 1980, 1992; Christopher D. Wickens, 2002) is particularly useful when attempting to explain human performance in dynamic tasks with a high workload, especially in multiple tasks executed at the same time (concurrent tasks). It presents four distinct dimensions that affect performance in such tasks, which are: processing stages, perceptual modalities, visual channels, and processing codes. Each of these dimensions is dichotomous, meaning that it can have two distinct levels in the model. Concurrent tasks utilising different levels in each dimension tend to interfere less with each other.

The processing stages related to a task response can be either perceptual / cognitive, mainly utilising working memory, or response-related, *i.e.* producing a response based on a piece of information that has been perceived. Concurrent tasks focusing on either of the above distinct stages tend to show low interference with each other. For example, a verbal acknowledgement of the state of an aircraft by an air traffic controller would not strongly interfere with their mental map of the airspace. In terms of perceptual modalities, cross-modal time sharing tends to be more effective than intra-modal time-sharing. For example, route guidance when driving, which is a largely visual task, is better presented only auditorily than only visually. Visual channels are nested in the visual modality and relate to whether a task will utilise mainly focal vision (required for fine detail and pattern recognition, *e.g.* reading text, recognising objects) or ambient vision (required for peripheral vision and perception of orientation and motion). These two types of vision tend to show low interference with each other. For example, a driver is capable of keeping the vehicle in the centre of the lane (ambient vision) while reading a road sign (focal vision). Finally, processing codes refer to whether the resources used for a task are mainly manual / spatial (utilising motoric skills) or vocal / verbal (utilising language). Response tasks utilising the above resources do not strongly compete with each other when one task is manual / spatial and the other vocal / verbal. For example, driving while dialling or texting, all being manual tasks, are more poorly time-shared compared to driving and voice dialling, one manual and one vocal task. See Figure 2-2 for a three-dimensional representation of the described theory, the Multiple Resources Model.

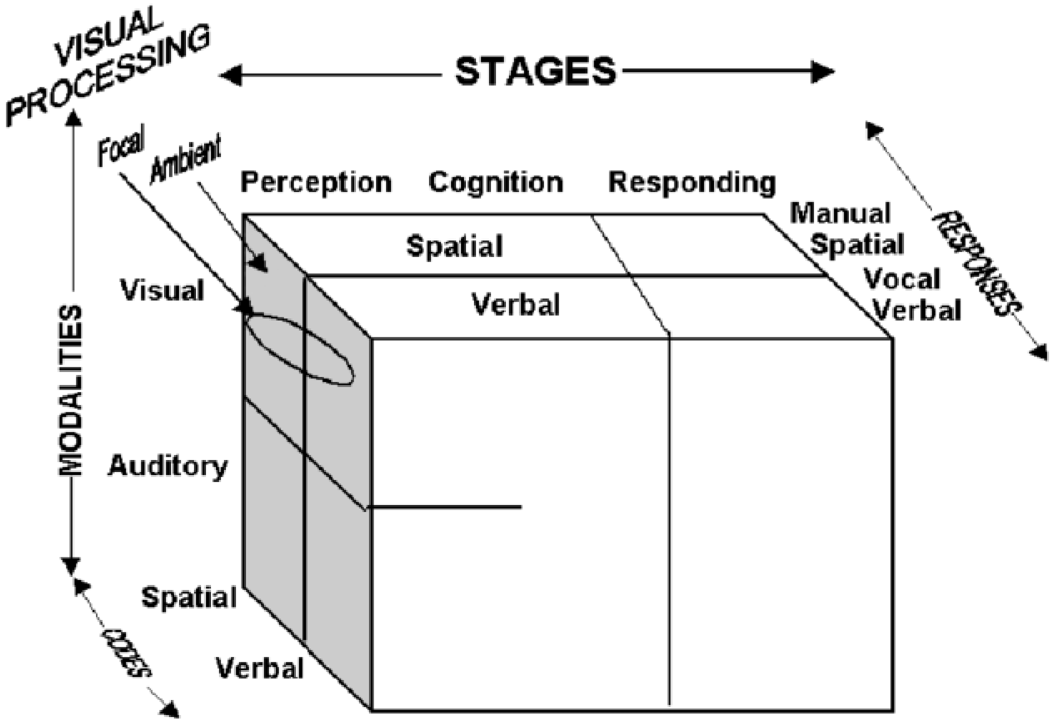


Figure 2-2: Three-dimensional representation of the structure of multiple resources. The fourth dimension (visual processing) is nested within visual resources. Image taken from (C. D. Wickens, 2008).

The Multiple Resources Theory sets a basis for explaining why a driver's performance may suffer when in high workload, and also for predicting concurrent task performance. When occupying competing resources, workload can quickly exceed safe levels and become overload. Aspects of this theory have been demonstrated in empirical studies, for example (Horrey & Wickens, 2004). In that study, Horey & Wickens used a driving simulator to study driving while voice dialling phone numbers. They showed that a focal task (voice dialling phone numbers presented visually) did not significantly interfere with an ambient task (lane keeping and speed control), at least in the absence of a critical event. Conversely, when a critical event occurred, which consisted of reacting to a danger presented visually, there was task interference. Horrey & Wickens argue that this was due to the new focal nature of the visual aspect of the driving task in presence of a danger, a discrete visual stimulus on the road, making the two tasks compete for resources. In terms of modalities used in the task, auditory presentation of the digits created quicker responses to visual presentation, as would be expected. However, it also disrupted driving behaviour in absence of hazard more than the visual presentation, which contrasts the expectations from Multiple Resources Theory. The authors accounted this fact to a possible elevated workload arising for the nature of the auditory task, which they referred to as pre-emption. Pre-emption related to some natural properties of audition, involving abrupt presentation and requiring mental rehearsal of the digits, as opposed to a visual inspection which was less disruptive and less salient.

To summarise, despite occasional findings that are less in line with this model, Wickens' theory has stood the test of time and is frequently used, especially when accounting for elevated mental workload situations due to dual-tasks. A more recent review by Wickens (C. D. Wickens, 2008) gives examples of such studies, but also highlights some limitations of the model, including the absence of the tactile modality in the perceptual modalities considered, as well as auditory pre-emption, the natural saliency of the auditory modality, which is difficult to ignore and therefore may be disruptive rather than helpful for an ongoing visual task (see also (Christopher D. Wickens, Dixon, & Seppelt, 2005)). This saliency is however what also makes audio suitable as a warning modality. The promising nature of audio as a warning mechanism, as well as the underexplored nature of vibration were both incentives to focus on multimodal displays, utilising a variety of modalities for this thesis.

2.1.3. Signal Detection Theory

Having presented the challenges of keeping an operator aware of their environment (Endsley, 1995), without over-utilizing their attentional resources (C. D. Wickens, 2008), the use of artificial stimuli as interface elements will now be discussed. As will be explained in the following section (Section 2.1.4), the challenge of creating effective alarm stimuli is a challenge that extends throughout the whole routine of interaction with a system. However, before discussing the topic of alarms as an effective behaviour changing strategy in Human Factors, one needs to discuss what detecting any signal directed to a user involves. This is because, in order for one to be able to react to a stimulus, one first needs to decide whether a stimulus is indeed an alarm directed at them or whether is something irrelevant to the task at hand, *i.e.* noise. This topic has been modelled by the Signal Detection Theory (SDT), initially presented by (D. Green & Swets, 1966), an overview of which will be provided below, as presented in the original work, and in (Abdi, 2007).

SDT is a way to model an observer's interpretation of a stimulus, regarding whether it is relevant to the task at hand (*i.e.* signal) or irrelevant to it (*i.e.* noise). A classic example, which also motivated the theory, is interpreting radar signals by operators, where the decision needs to be made on whether they are planes (signal) or something else (noise). SDT can be used in other contexts as well, for example when deciding whether a stimulus has been presented or not in a cue recognition experiment, or whether a finding in a medical image is suspicious or not. In all cases, the probability distributions of signal and noise differ to each other by a factor typically referred to as *d'* (see Figure 2-3). The task of the observer is to decide whether a stimulus belonged to the signal or the noise distribution. In other words, whether there was a signal or no signal (*i.e.* noise). As can be observed in Figure 2-3, the closer the signal and the noise distributions are to each other (the smaller *d'* is), the harder is the problem to tell them apart. Performance to this task can have four distinct outcomes:

- The observer might decide that there was a signal, when in reality indeed there was a signal. This is referred to as a **Hit**.
- The observer might decide that there was no signal, when in reality there was a signal. This is referred to as a **Miss**.
- The observer might decide that there was a signal, when in reality there was no signal. This is referred to as a **False Alarm**.
- The observer might decide that there was no signal, when in reality indeed there was no signal. This is referred to as a **Correct Rejection**.

Another element of SDT is the strategy of the observer, regarding whether they will call a stimulus signal or noise. This is reflected in a cut-off point, referred to as the criterion *c*. Every stimulus beyond this point (*i.e.* to the right of this point on the horizontal axis) will be called a signal by the observer, and every stimulus to the left will be called noise. If the criterion is set too high on the horizontal axis, the observer will tend to decide that stimuli are noise (conservative strategy). This strategy is likely to increase correct rejections (decrease false alarms), but it will also decrease hits (increase misses). In the opposite case, if the criterion is set too low on the horizontal axis, the observer will tend to decide that stimuli are signals (liberal strategy). In this case, hits are expected to increase (misses will decrease), but false alarms will also increase (correct rejections will decrease). A trade-off is therefore uncovered, on where a good criterion for the observer is to be set, which depends on the particular nature of the task at hand, and can be different across domains.

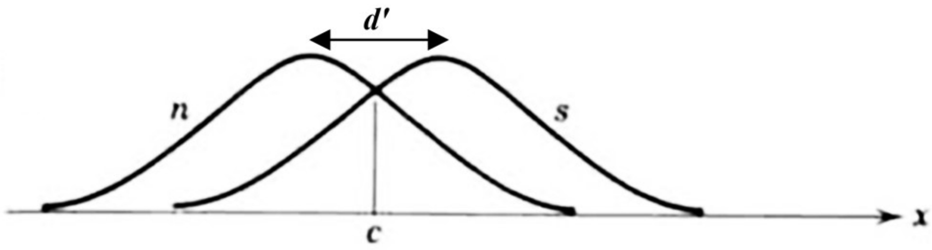


Figure 2-3: The distribution of noise (*n*) and of signal (*s*), the distance between the two (*d'*), and the criterion whereupon the observer will decide a stimulus is signal or noise (*c*). Image modified from (D. Green & Swets, 1966).

The significance of SDT is that, with enough observations, the strategy of the observer can be understood, based on their tendency to decide liberally or conservatively. This is a relevant exercise when uncovering psychological mechanisms that function when responding to stimuli. Further, SDT encodes the space of responses to alarms (hits, misses, false positives, correct rejections) enabling the study of each of these outcomes. Warning design aims to facilitate true alarms (hits) and correct rejections, and to hinder false alarms and misses. This topic will also be discussed in Section 2.5, where the urgency of the situation will be presented as another factor that can influence whether an alarm is true or false. In any case, creating alarms that are sufficiently different to noise is essential. In the car, this can mean creating alarms that differ to the background auditory, visual or tactile environmental noise, and are salient enough to facilitate appropriate responses (see Section 2.2 for an analysis of warnings as alarms in all of these modalities). In this way, responses will be present when needed (hits), and absence of alarms will be interpreted as such (correct rejections). Conversely, with a poor alarm design, alarms could be ignored (misses), or with an overutilization of alarms, alarms could be perceived when absent (false positives). To uncover the functions that take place when exposed to alarms, their use as an effective Human Factors strategy will be discussed below.

2.1.4. Alarms as a Human Factors Strategy

Having described the properties of a signal directed to an operator as opposed to noise, one can now discuss the use of signals alerting of a situation that requires attention, *i.e.* alarms.

As defined by (Stanton, 1994) in his book on the topic, an alarm has the role to give warning of impending danger, in varying degrees of severity. It can indicate for example an unexpected change in system state, a means of signalling state changes, a means of attracting attention, a means of arousing someone or a change in the operator's mental state. A system model of alarms was presented by Stanton in the same work (see Figure 2-4). In this particular model, the system was referred to as 'plant', possibly revealing a connotation to nuclear plants, where this model was extensively used. The arrows in the model indicate transition of information. When there is a change in the system (Prerequisite events), the operator needs to be aware of it. Such a change affects the threshold value of the system (Threshold set point), making it too high or too low. The human operator needs to receive communication of this change in some modality (Alarm panel). This information reception (Attraction) will necessitate the operator's response or non-response by making them aware of the situation (Aroused state). The operator then needs to decide on the appropriate course of action (Decision), and take the that action (Behaviour). If this process is successful, the appropriate action will be taken by the operator. If not, then an error will occur.

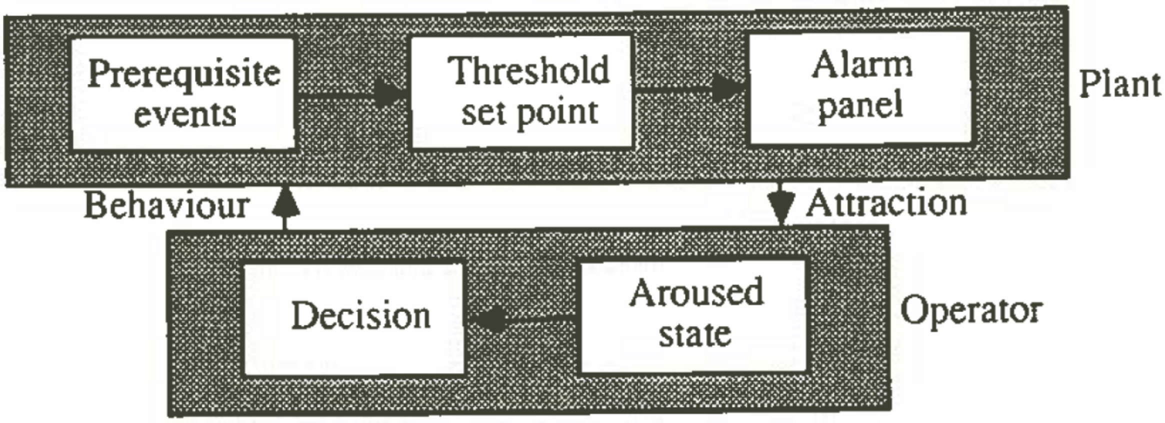


Figure 2-4: A systems model of alarms. Taken form (Stanton, 1994).

As mentioned by Stanton later in the same work, creating a model of alarm handling is necessary to guide research, *"so that we may ask appropriate questions and utilize empirical techniques to yield answers"*. Figure 2-5 presents Stanton's model of alarm initiated activities, where the interaction between system and operator during an alarm is modelled. As already identified by Stanton, and elaborated in urgency research by (E. J. Hellier, Edworthy, & Dennis, 1993) and others (see also Section 2.3), there needs to be a reflection of the criticality of the task to be performed when interacting with alarms. In Stanton's model the dotted lines represent critical incidents, while the plain lines represent routine incidents. The numbered arrows represent the possible routes of actions to be taken in the presence of the alarm. These are to ignore the alarm (1), to monitor it (2), to routinely deal with the cause (3), to further investigate and correct the cause (4), or, if the alarm cannot be cleared, to go in an investigative mode to seek the cause, correct and monitor further (5).

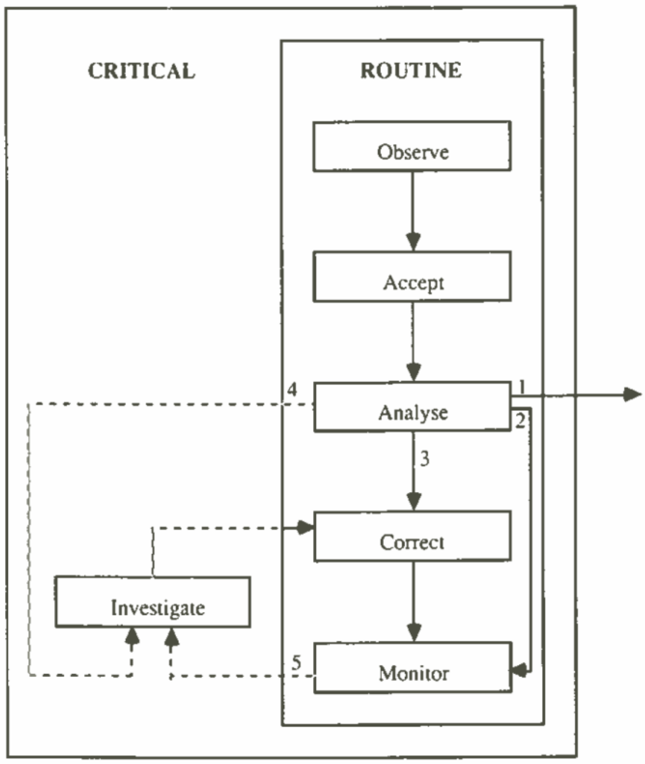


Figure 2-5: Model of alarm initiated activities. Taken from (Stanton, 1994).

To further analyse the modes described in Figure 2-5, the mode "Observe" involves an initial detection of abnormal system behaviour. This assumes an effective display of the system in regards to its own abnormal behaviour, that is able to attract the operator's attention. There can be limitations in this process, such as system failure, signal and noise being too similar to each other (see also previous section), and the alarm being ignored, misinterpreted, or not observed due to too much information available (the way information can compete for resources of the operator has also been discussed in Section 2.1.2). All these considerations need to be taken into account, in order to attract the user's attention. As Stanton mentions, attracting attention necessarily means distracting the user from the other aspects of the task. Therefore, alarms should highlight a problem rather than force an action, unless immediate reaction is needed due to a critical situation.

The mode "Accept" involves the acknowledgement of the alarm by the operator. This may mean performing a physical action in order to deactivate the alarm. Combinations of alarm modalities in this case will increase saliency and aid the acknowledgement of the alarm.

Possible problems in this process is in the case of multiple alarms, where a group acknowledgment of all of them may lead to missed signals and loss of information. This group acknowledgement may come from the operator's intention to mute the alarms and avoid distraction. This creates a limitation, namely the operator avoiding the reception of information by the system, which information is designed to be useful and elicit certain action. This is why overutilization of alarms is not ideal, as will be discussed in experimental chapters of this thesis. This also motivated the use of subjective measures of annoyance in this thesis, and in prior work (see also Section 3.2), in order to be able to assess how intrusive the warnings designed might be to users.

The mode "Analyse" involves the assessment of the situation signified by the alarm, and the decision on which will be the actions to address it. The alternative routes to be followed by the operator can be the ones described earlier in this section, namely to ignore it, to monitor it, to routinely correct the cause, to further investigate and correct the cause, or to continue investigation, since the cause could not be determined. As Stanton mentions, conveying the right amount of information is essential in this step, so as to avoid unnecessary effort. As also mentioned, the right degree of urgency needs to be embedded into the alarm, making it clear in terms of the criticality of the situation it signifies. Further, the use of abstract cues or speech cues, text or pictures, as well as different modalities are possible design elements to be considered. Stanton also quotes Wickens' theory, suggesting that using cues in modalities that do not compete in resources with the modalities utilised by the main task may be helpful in creating effective alarms. Finally, it is suggested that using a combination of codes in the alarms may further assist the analysis task.

The mode "Investigate" refers to discovering the cause of the alarm, if it is not routinely available, with the goal to address the underlying problem the alarm signifies. This exercise is a problem solving one, where breaking the task into subtasks and sequentially address those seems to be a mechanism typically followed by operators. Since operators are individuals with differences in their mental representation of the system, Stanton claims it is difficult to prescribe one mode of alarms that will assist everyone. However, facilitating the problem solving sequence, by highlighting the actions to be taken in order to solve the problem seems to be a viable strategy in alarm design. Stanton's pointers were followed in the warning design of this thesis, where clarity of meaning in the alarms and the actions to be taken were emphasized.

The mode “Correct” involves the corrective actions taken by the operator to address the alarm. As suggested, corrective actions taken are not always correct, while limited actions taken does not necessarily indicate low mental workload. Therefore, effective alarms should aid the process of correcting the error. One potential way to achieve this, is to present the alarms in modalities of the expected response. For example, speech cues could be used when speech response is required, while spatially distributed abstract cues could be used for manual responses. However, there are situations where cues are not easily separated in terms of expected response, especially in tasks where frequent manual input is intrinsically involved. Stanton mentions plant control rooms as an example of this, but it is claimed here that driving is another example of a highly manual task, where there needs to be expectancy for manual responses to alarms. This thesis treated alarm designs equally in terms of expected responses, requiring manual responses to the designed warnings. This was necessary in order to be able to make comparisons of performance in the presence of fixed response tasks. However, deciding on different response modalities for different alarm designs would also be a possible future investigation.

Finally, the mode “Monitor” includes the assessment of the outcome of one’s actions to address the alarm. Contrary to the mode “Analyse”, where the strategy to be followed is decided, in this mode the effect of the actions already performed by the operator is evaluated. In this stage, system feedback on the actions of the operator is important, since it makes the results of user input prominent. If the user’s understanding of the system is not complete, then the outcomes of their actions may be unclear, which will delay the problem solving process. It is suggested that systems should guide the operator’s activities in the context of alarms, facilitate minimal need to seek information by the operators, and focus on display design that corresponds to the limited attentional responses of the operator. Again, Stanton’s pointers were followed in the design of alarms in this thesis, by using available guidelines to design concise and clear warnings, that effectively signify the event in question.

To conclude, the purpose of driver warnings is to effectively alert the driver and alter their behaviour when this is necessary. This is the final desired outcome of the warnings (or alarms), however there are more elements to be considered when presenting these warnings throughout the complex interaction taking place during driving. This section presented an overview of prominent literature justifying why an operator needs to be aware of their environment when interacting with complex systems, by discussing Situation Awareness theory (Section 2.1.1). The attentional resources utilised when interacting with such a complex interface and how they can compete to each other was then discussed, by presenting the Multiple Resources Theory (Section 2.1.2). In order to attract attention, establish Situation Awareness, and overcome the struggles presented through high utilisation of resources, alarms were suggested as one viable solution. Since alarms need to be effectively interpreted by system operators, the property of a signal being an alarm as opposed to a non-alarm was discussed through Signal Detection Theory (Section 2.1.3). Finally, why alarms are a viable means of interface design and how they operate in potentially altering behaviour was discussed in Section 2.1.4. It is noted, that the focus of this thesis was in the early stages of alarm initiated activates (Figure 2-5), and drivers of the experiments presented in later chapters would be asked to apply specific corrective actions to address the alarms designed, with little ambiguity on how to react. However, the pointers of Stanton's model and his considerations on how alarm design needs to be investigated in terms of warning modalities used, designed urgency, as well as message content, were followed throughout this thesis experimental work. The following section will discuss examples of utilizing warnings of various modalities and effectively alerting drivers of events on the road, presenting applications of the use of alarms in HCI, and further demonstrating why this technique is viable as a means of attracting attention.

2.2. Multimodal Driver Displays

Multimodal driver displays have been used from the early days of Human Factors research and they have achieved good results in attracting driver attention. This section will review available literature on multimodal displays, utilizing audio, visual and tactile modalities. In order to provide a better overview of available work, literature regarding each modality will be addressed in a separate subsection, and work on multimodal combinations of displays will be presented in a separate subsection as well. This section addresses the "*multimodal*" aspect of RQ-1: "*How do **multimodal** driver displays varying in urgency affect performance?*", and the next one (Section 2.3) addresses the "*urgency*" aspect of RQ-1.

Even from earlier research in driving behaviour, the importance of expectancy to reactions during a critical event on the road was emphasized. Schweitzer *et al.* & Liebermann *et al.* (Liebermann et al., 1995; Schweitzer et al., 1995) performed a road study where young athletes were asked to brake as a response of a lead vehicle braking in real traffic conditions. Such a study would be rare to find in more recent literature due to high safety concerns of performing repeated emergency stops in real traffic. It was found that when a critical event was expected by informing the participants before the experiment, reactions were quicker as opposed to when participants were naïve to the event. The authors therefore emphasize that the use of pre-cues can aid braking responses in emergency stops. The importance of warnings to aid reactions whether drivers are distracted or not was also confirmed by Lee *et al.* (John D. Lee et al., 2002), who advocated the use of this mechanism to alert car drivers. More recently, Moheby, Gray & Tan (Mohebbi, Gray, & Tan, 2009) specifically addressed the case where a driver was distracted by a phone conversation and found that warnings and especially vibrotactile cues, helped attract attention and react quicker to imminent events in simulated driving.

Ho & Spence provide an extensive review of how multimodal warnings can assist responses to driving–related events (Christy Ho & Spence, 2008). After discussing literature on driver distraction, as well as psychological studies on auditory, tactile and multimodal cuing of driver attention, they conclude that non-visual multimodal displays are particularly promising in capturing attention and facilitating quick reactions to unexpected events on the road. In the following sections, the presentation approach of (Christy Ho & Spence, 2008) will be followed, by presenting research related to the modalities of interest (audio, visual, tactile and multimodal), and identifying the research potential of studying all multimodal combinations of these modalities as vehicle alerts.

2.2.1. Audio Displays

The use of sound has shown to be an effective means communication in human-computer interfaces. Brewster, Wright & Edwards (S. A. Brewster, Wright, & Edwards, 1993) found that the use of Earcons (introduced in (Blattner, Sumikawa, & Robert, 1989)), abstract synthetic sounds, was effective in communicating messages related to computer usage. Graham (Graham, 1999) used Auditory Icons (introduced in (Gaver, 1986)) as warnings (the sound of skidding tyres and of a car horn) and compared them to abstract sounds and a simple speech message (a voice saying “ahead”). It was found that Auditory Icons produced quicker reaction times but more inappropriate responses to assessing whether road video footage was depicting a critical driving event or a stationary vehicle. This task had low ecological validity, but showed that audio can effectively be used as an alert in a driving scenario.

Lai *et al.* (Lai et al., 2001) compared synthetic and recorded human speech for delivering navigational messages, email messages as well as news stories while driving a simulated vehicle. Listening to the messages did not disturb the driving metrics (standard deviation of lateral lane position and of steering angle), but when asked questions regarding the content of the messages, it was found that comprehension of human speech was higher. Ho & Spence (Cristy Ho & Spence, 2005) investigated the use of car horn sounds and speech cues as warning signals. Participants responded quicker to a critical event when the cues were coming from the direction of the event (front or back) and when their attention was directed to the correct direction through a speech cue (*"front"* or *"back"*). The task used by Ho & Spence was a reaction task, where a rural road was depicted with either a car in front braking towards the participant's car (*"front"* warning) or the car on the back performing this action (*"back"* warning, see Figure 2-6). These studies showed the utility of speech as a warning.

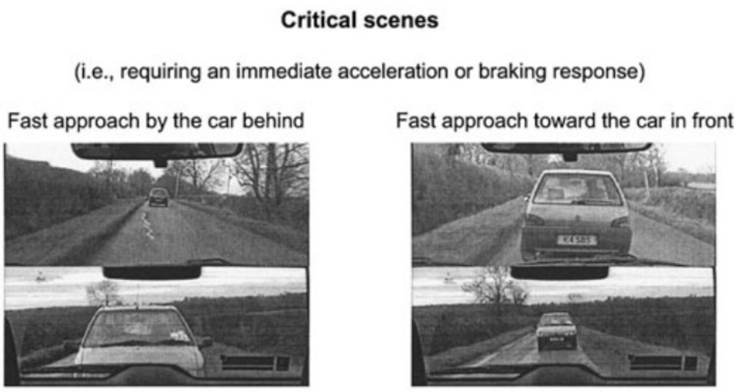


Figure 2-6: Sample video stills taken from the video clips used in (Cristy Ho & Spence, 2005). The upper half of each still shows the view of the windscreen seen directly in front of participants, whereas the lower half shows the rear view seen by the participants in the rear view mirror. Image taken from (Cristy Ho & Spence, 2005).

McKeown & Isherwood (Denis McKeown & Isherwood, 2007) evaluated a set of auditory warnings with different semantic associations to driving events. Four suites of warning sounds were designed, with nine sounds in each suite. Abstract sounds (*e.g.* tones or a siren), environmental sounds with no relation to driving (*e.g.* a baby sneezing or footsteps), environmental sounds related to driving events (*e.g.* car speeding past) and speech messages (*e.g. "Exceeding speed limit"* or *"Petrol is low"*) were used. It was found that abstract sounds had the highest response times and the lowest identification accuracy. Speech and auditory icons related to driving had the lowest response times and the highest accuracy. Speech was perceived as more pleasant and less urgent compared to abstract sounds. This study further confirmed the effectiveness of using speech warnings as alerts.

Cummings *et al.* (Cummings, Kilgore, Wang, Tijerina, & Kochhar, 2007) tested five different auditory warnings, one master alarm (addressing a critical case independent of what it was) and four individual alarms (addressing four specific critical cases examined by the authors). These cases were a forward collision, a lane departure either to the left or to the right and a fast approach from a vehicle behind, which were also reproduced in a driving simulator. All alarms used were abstract repeated pulses of varying durations and interpulse intervals, which were used in a real car. No difference was found in reaction time and accuracy in the above simulated events when using a general master alarm as opposed to directional specific alarms. In addition, reactions to forward collision warnings were the quickest, while low alarm reliability (false alarms) negatively influenced vehicle speed control. This study suggested that using a warning with little semantic association to the signified event as a general alert can also be effective.

Sullivan & Buonarosa (Sullivan & Buonarosa, 2009) tested three suites of sound warnings, namely semantic (natural sounds, semantically associated with the events they signified), less-urgent-semantic (same as semantic, but modified for attributes such as pitch, pulse rate and duration to convey lower urgency) and abstract. The semantic sounds were an appliance beep, a car horn, the sound screeching tires and a rumble strip sound, while the abstract sounds were repeated pulses. All sounds were used along with pictures of urgent road scenarios (forward collision, high speed in a curve, lane change when a vehicle approaches from the back and lateral drift towards a vehicle on the right). Participants were asked to identify the situation depicted and it was found that semantic warnings induced the fastest reaction time and highest recognition rate for this task. Thus the distinction of various critical scenarios was shown to be better facilitated by sounds semantically relevant to the scenarios. However, the task used was not highly naturalistic.

Serrano *et al.* (Serrano et al., 2011) presented a set of speech messages to drivers (*"Look out on the left / ... on the right /... on the road!"*). They were followed by pictures of either hazardous or non-hazardous road scenes, asking drivers to identify whether there was a hazard. Reaction times to this task were shorter and responses more accurate when the speech messages were presented from the direction of the hazard as opposed to a random direction. Messages presented from the correct direction created even shorter reaction times, when that direction was uttered in the message as opposed to not specified. This demonstrated the effectiveness of short speech warnings when delivered from the appropriate direction relative to the threat. However, the subjective responses of the warnings as well as the recognition accuracy were not assessed. This is essential when drivers need to interpret the meaning of messages and act appropriately.

Gray (Robert Gray, 2011) introduced looming auditory warnings, *i.e.* critical warnings of intensity varying as a function of the leading vehicle proximity. Motivated by the high error rate when responding to symbolic cues (*e.g.* a car horn), Gray designed a set of auditory warnings consisting of a simple tone increasing in intensity as a lead simulated vehicle braking suddenly would approach the participants' vehicle. These were compared to a car horn sound, a constant intensity tone, a repeated pulse and a tone increasing in intensity independently of the lead vehicle distance. Looming warnings outperformed all other cues in terms of reaction times to the critical event as well as reaction accuracy (false activations of the brake). This study showed the effectiveness of conveying distance information through audio intensity in a critical scenario.

Finally, Fagerlönn, Lindberg, & Sirkka (Fagerlönn, Lindberg, & Sirkka, 2012) used an alternative strategy to alert drivers, by panning the sound of the radio towards the driver's opposite side or reducing the sound level of the radio. Participants responded positively to panning the radio and identified this as quickly as a mild abstract warning signal consisting of repeated tones. These alerts were used as early warning strategies, meaning they were not designed to signify critical events. As such, the intervention of panning the radio away from the driver provided positive results and was found to be interchangeable in terms of effectiveness with an abstract sound.

To summarize, there have been numerous studies examining audio as a warning mechanism. Speech, tones and sounds with some semantic association with the events signified have been used, with results showing benefit on using either of these cues, depending on the situation. Further, aligning the direction of the cues to the direction of an approaching threat has been found to improve reactions. However, using more modalities to alert drivers has shown even higher benefits, as will be further discussed in Section 2.2.4. Even when displays are multimodal, an exhaustive combination of the modalities used has rarely been studied, as will also be discussed in that section. Additionally, the presented studies did not largely vary the urgency of the warnings used and the situations signified. They either studied critical events or non-critical ones with warnings of fixed urgency. Varying the urgency of the warnings would help generalise results on the warning utility in differently urgent contexts. This is also something lacking in most studies in this section, which will lead to the discussion on how to explicitly vary the urgency designed in the warnings in order to create alerts for more versatile scenarios (Section 2.3).

2.2.2. Visual Displays

Although prevalent in vehicles, unimodal visual displays as in-vehicle alerts are often discouraged, since they can intervene with the primarily visual divining task, see for example (Hirst & Graham, 1997; Scott & Gray, 2008; Van Erp & Van Veen, 2004). However, they have been used in many previous studies, as well as in more recent years, with the development of Augmented Reality (AR) in the car. This subsection reviews available studies on unimodal visual driver displays.

The most traditional instruments of the vehicle consist of mostly visual displays (see Figure 2-7 for an example of a car dashboard). Aside to the car dashboard, another set of visual displays, targeted to users outside the car are the brake lights, turn signals, reversing lamps, as well as emergency vehicle lights. The appearance of these devices is largely standardised and regulated (*e.g.* (United Nations, 1959, 1968)). Examining the dashboard indicators and outside light displays of the car is beyond the primary scope of this thesis. However, the studies that will be discussed in this section include some that have looked into the utility of a subset of these displays.

Figure 2-7: Dashboard design of a Lancia Orca, a 1982 concept car. The uncommon design includes a large number of visual displays and input devices, which can be hard to interact with. Image taken from: blog.petervidani.com.

Sivak, Post & Olson (Sivak, Post, & Olson, 1981) compared reaction times between traditional brake lights and the addition of high mounted brake lights fitted in the top of the car's boot. This was a real traffic study and it was found that fewer brake responses were missed with the additional lamps. Liebermann *et al.* (Liebermann et al., 1995), found that responses were quicker and there were less missed responses when the onset of brake lights was due to an actual deceleration of the lead vehicle as opposed to a false alarm (in which case they were referred to as "dummy" brakes). These studies combined point to a desired saliency but also validity of the brake lights to improve their effectiveness as alerts for following cars.

Dingus *et al.* (Dingus et al., 1997) created a set of visual displays to help drivers assess the distance of the lead vehicle. They compared three displays mounted on the dashboard, one showing a car icon increasing in size as the lead vehicle approached, one with coloured bars that would increase in number and one that would only depict two blocks, one orange and one red (see Figure 2-8). The car icon and bar displays would flash red once the distance to the lead vehicle was critical, while the block display would flash orange once the lead vehicle was close and red once a collision was imminent. Participants drove in real traffic, were naïve of the purpose of the experiment and were distracted by in-vehicle tasks. The car icon and bar displays helped participants maintain safer headways, both when simply following a lead car and when the car would brake slightly. No critical events were tested for safety reasons; however, the results showed how constant visual feedback can help maintain safe distance to a lead car.

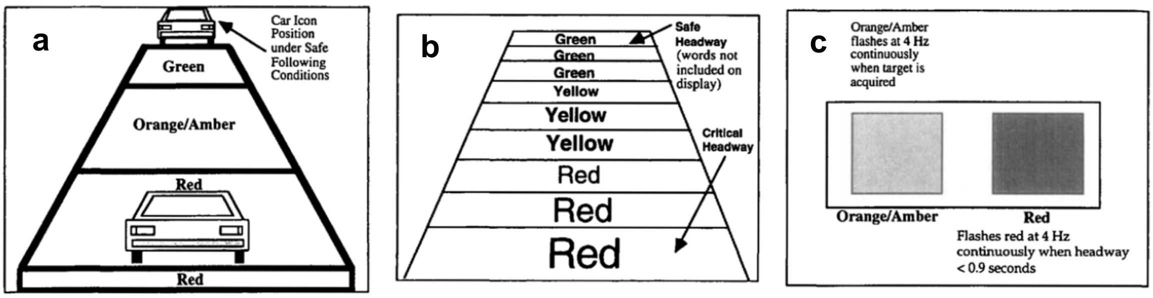


Figure 2-8: The visual displays created by Dingus *et al.* (Dingus et al., 1997), using a car icon (a), bars (b) or blinking blocks (c). Image taken from (Dingus et al., 1997).

Head-up displays (HUDs) are another in-car technology that has achieved popularity over the years. Introduced commercially almost three decades ago (see (Weihrauch, Meloeny, & Goesch, 1989)), they were inspired from aviation displays, where they provided some benefits when operating an aircraft (see for example (Martin-Emerson & Wickens, 1997)).

They are able to display driving related information closer to the driver's visual field compared to the dashboard (see Figure 2-9). HUDs have been studied since the early days of their adoption in vehicles, *e.g.* by Inuzuka, Yoshimasha & Shinkai (Inuzuka et al., 1991), who found that they improved recognition time of a speed indication and disrupted the eye gaze less compared to a traditional instrument panel.

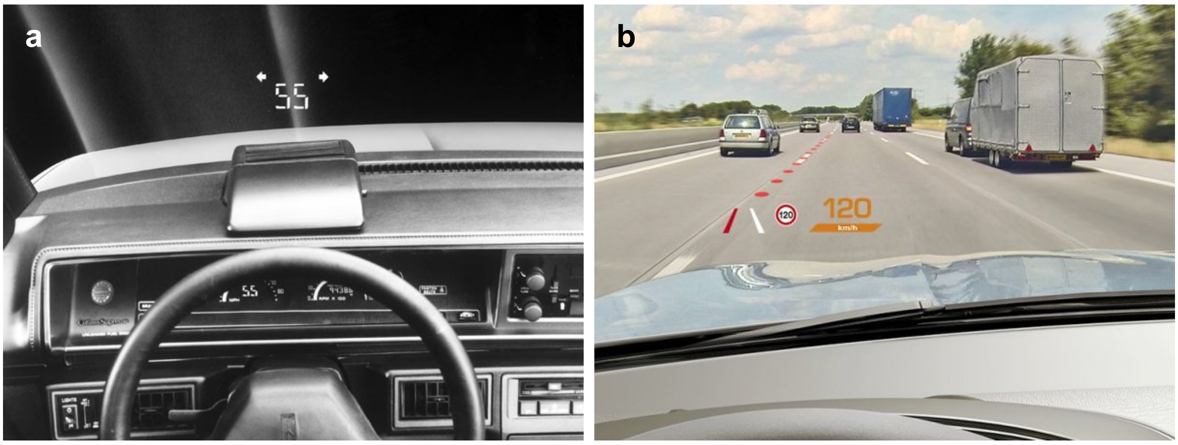


Figure 2-9: Two HUDs from different eras. (a) The first commercial automobile HUD in a General Motors 1988 Oldsmobile Cutlass Supreme, able to display turn indications and speed (image taken from: dailytech.com). (b) An Augmented Reality HUD from Continental, currently used in vehicles. It is able to display rich content and align to the driver's visual field. In this picture a lane departure warning, current speed and speed limit are displayed (image taken from: continental-head-up-display.com).

Liu & Wen (Y.-C. Liu & Wen, 2004) conducted an on-road study and compared a HUD with a visual display placed on the level of the instrument panel and towards the centre of the vehicle, a Head-down display (HDD). They used the displays in commercial vehicles to provide information during goods delivery, navigation and speed-related information, as well as provide warnings related to road conditions or vehicle conditions that required prompt attention. They found no difference between HUD and HDD in terms of driving behaviour when receiving navigational instructions and goods delivery instructions, concluding that a prolonged exposure to visual cues of low criticality by these displays creates similar results. However, they observed quicker reactions to visual warnings and better adherence to speed-related information when using the HUD, leading them to suggest it for high load road conditions. Similar results were presented by Ablaßmeier *et al.* (Ablaßmeier et al., 2007) and Doshi, Cheng & Trivedi (Doshi, Cheng, & Trivedi, 2009). They conducted on-road studies presenting driving-related information on HUDs, such as speed, and observed less gazes off the road using a HUD and higher acceptance of this device compared to displays positioned in the centre of the vehicle or the dashboard. These results show the potential of HUDs in attracting driver attention, but do not examine critical situations, on which an imminent reaction is required.

A more recent approach in visual displays for cars is the use of Augmented Reality (AR). This technology is able to display context-related information which is aligned to the driver's field of view (see Figure 2-9.b) and has been investigated in a number of studies. Medenica *et al.* (Medenica et al., 2011) presented a prototype of an AR-HUD that displayed navigational information by projecting linear trails in front of the driver. They compared this display with a traditional map-based navigational device and a street-view navigational device, both placed near the car centre when used. The AR-HUD required fewer glances off the road, induced less subjective workload and was preferred to the other devices. Rush *et al.* (Rusch et al., 2013) used an AR-HUD to highlight roadside hazards and found some benefits in detection rates of pedestrians and road signs, but not vehicles. Wernecke & Vollrath (Werneke & Vollrath, 2013) used such a device for collision warnings at intersections and found benefits in driving behaviour when presenting a visual warning well in advance of a critical event in an intersection, at the area of the visual field where the event was expected to occur. This functionality would require higher connectivity of the vehicle and higher awareness of its surroundings, but proved more effective in terms of driving behaviour compared to late visual warnings when a lead vehicle would already be visible.

The above studies showed the utility of visual displays, mostly for non-critical situations where driver attention is required. The use of HUD and AR displays have shown some added benefits, since they are placed closer to the driver's field of view and can be less disruptive. Informed by these findings the visual parts of the displays used in this thesis were also designed to be close to the drivers' field of view, as will be described in the experimental chapters. As will be discussed in Section 2.2.4, the benefits of visual displays to alert drivers can be more pronounced when combined with alerts in the audio and tactile modalities, since, when used alone, they are not as effective in signifying critical events.

2.2.3. Tactile Displays

Tactile displays have been investigated as a means of transferring information in the car for a number of decades. Early studies by Fenton (Fenton, 1966) and Fenton & Montano (Fenton & Montano, 1968) investigated the possibility of using haptic means to convey information through a control stick, a device resembling a joystick. This device would be able to control the vehicle's direction by pulling it back, pushing it forward and turning it left or right (see Figure 2-10). It also had a "finger", a protuberance that would protrude or recess in a manner proportional to the lead car headway. It substituted the pedals and steering wheel in a driving simulator (Fenton, 1966) and assisted participants in maintaining the desired distance to the lead vehicle. The observed advantage persisted in an on-road study (Fenton & Montano, 1968). However, in the following decades, there appeared to be less focus in developing in-car tactile displays, with literature in such displays being mainly focused in aviation (see for example (Gilliland & Schlegel, 1994; Zlotnik, 1988)). This fact was acknowledged by Burnett & Porter (Burnett & Porter, 2001), who stressed the potential benefits of utilising the tactile modality in the car. They argued that the tactile modality can provide information without using the visual system and can help the older population, with possible decreases in visual and auditory capabilities. Studies by Van Erp & Van Veen ((Erp & Veen, 2001; Van Erp & Van Veen, 2004)) also appeared around that time, which presented the merits of using tactile and visual cues together for in-car navigation (see Section 2.2.4).

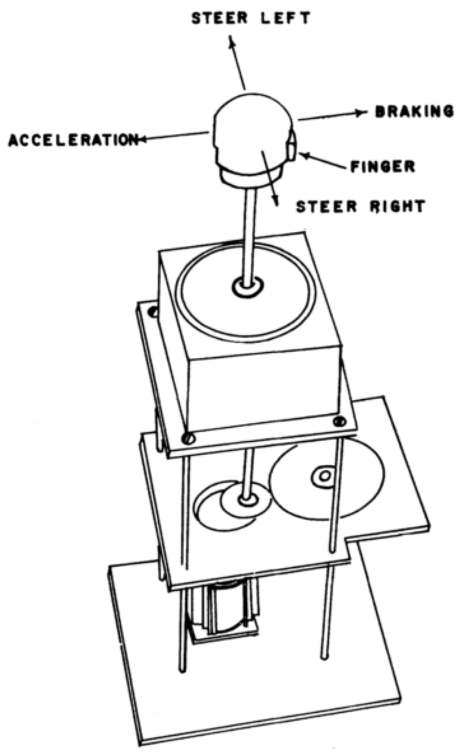


Figure 2-10: The control stick designed by Fenton. Image taken from (Fenton, 1966).

One location used to deliver tactile sensation is the steering wheel. Enriquez *et al.* (Enriquez et al., 2001) investigated a pneumatic system to provide vibrations on the wheel in an abstract gauge reading task. Their intervention improved recognition time of errors when one out of a set of gauges presented visually went to an error state. Steele & Gillepsie (Steele & Gillespie, 2001) tested the use of steering assistance by applying torque to retain a simulated vehicle on its required path. They found that this improved performance and reduced visual workload in a path following and obstacle avoidance task. Suzuki & Jansson (Suzuki &

Jansson, 2003) delivered vibrations through the steering wheel to signify lane departures and found that this improved time and accuracy of recovery when there was no prior exposure to the warning. With prior exposure, abstract audio pulses outperformed vibration. Steering torque was the poorest of the strategies tested, often inducing the opposite result, since participants corrected the wheel movement thinking it was created in error or by a wind gust. This effect was ameliorated by Mulder, Abbink & Boer (Mark Mulder, Abbink, & Boer, 2008) by reducing the torque applied and creating a haptic guidance that achieved a slight improvement in steering performance during curves. Finally, Chun *et al.* (Chun et al., 2012) found that a vibration on the wheel induced quicker reaction times to a simulated collision event compared to a vibrating seat belt which also improved reactions compared to a no warning condition. This benefit persisted when detecting a vehicle in the blind spot (Chun et al., 2013). These studies show the utility of vibrations on the steering wheel. However, as acknowledged by authors of the studies, the wheel requires physical contact in order for a vibration to be effective, which is not constantly the case while driving.

Investigating tactile cues presented on the torso, Ho, Tan & Spence (Cristy Ho, Tan, et al., 2005) studied a set of spatially predictive vibrotactile cues. The cues used were simple vibrations delivered through a belt either on the abdominal area or in the back of the participants. They were used to warn of a rapidly approaching car from the back or from the front, which was presented on video. They were either spatially predictive (indicating the correct direction of approaching car in 80% of the cases) or non-predictive (indicating the correct direction in 50% of the cases). Participants were required to brake if the approach was from the front and accelerate if it was from the back. Both spatially predictive and non-predictive cues presented from the same direction as the approaching car (front or back) decreased drivers' reaction times compared to cues presented from the opposite direction. Ho, Reed & Spence (Cristy Ho, Reed, & Spence, 2006) studied a similar setup in a driving simulator and confirmed the above results, also finding that participants responded quicker and maintained a safer distance to the lead vehicle when receiving spatially predictive cues as opposed to when receiving no cues. These studies contributed to available knowledge firstly by presenting the efficacy of tactile cues on the torso as a warning mechanism and secondly by demonstrating that these cues are also more effective when presented from the direction of the approaching threat, as has already been discussed for the case of audio cues. Additionally, the location of the torso does not suffer from the risk of absence of contact like the steering wheel does, which makes it promising for delivering vibrations, for example in form of a seatbelt.

A further location for delivering tactile cues is the pedals. Several studies have experimented with this location. Mulder *et al.* (M. Mulder, Mulder, van Paassen, & Abbink, 2008; Mark Mulder, Abbink, Van Paassen, & Mulder, 2011) designed a haptic accelerator pedal with varying force required to depress it, depending on the distance of the lead vehicle (the closer the distance the more force required). They found that observed car following performance improved, since the headway of participants' simulated car increased and their workload decreased with this haptic support mechanism. De Rosario *et al.* (de Rosario et al., 2010) used a haptic accelerator pedal that vibrated when a collision was imminent and found it improved reaction times compared to a visual icon displayed on the screen of a simulator (see Figure 2-11). Finally, Birell, Young & Weldon (Birrell, Young, & Weldon, 2013) investigated a haptic accelerator pedal in the context of economical driving. This pedal vibrated when depressed more than a threshold that was considered acceptable for low fuel consumption. The observed mean acceleration values and excess throttle use decreased with this intervention, while the perceived workload when simply asked to drive economically was higher compared to when the haptic pedal was activated. These studies present some advantages of delivering vibration through the pedals, it should be noted, however, that such cues might be missed, even more so than when delivered on the steering wheel, due to possible periods when the foot is not in contact with the pedal.

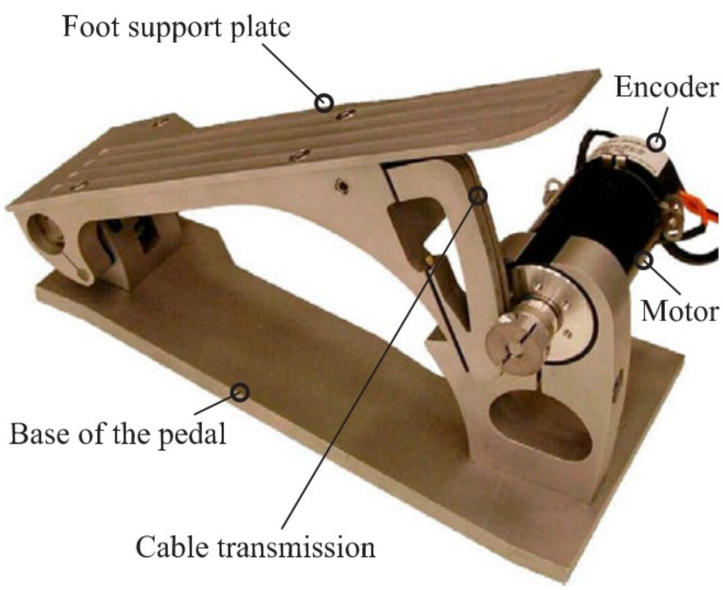


Figure 2-11: The haptic accelerator pedal designed by de Rosario. Image taken from (de Rosario et al., 2010).

The seat is a location with constant contact with the driver, which is an advantage compared to the steering wheel and pedals. Studies have used this location to deliver tactile cues, signifying both alerts as well as directional information. Hogema *et al.* (Hogema et al., 2009) mounted a set of tactile actuators in the seat pan. In an on-road study, they evaluated a set of eight tactile patterns, representing eight locations around the participant (front, back, left, right and their adjacent combinations). They observed very high performance of localization in different road conditions (brick road or highway). Fitch *et al.* (Fitch, Hankey, Kleiner, & Dingus, 2011) used both the pan and the back of the seat and created tactile patterns signifying locations around the participants. In an on-road study they found that alerts with unique tactile patterns, localised in either the seat pan or the seat back are easier to distinguish compared to varying patterns. Testing the seat pan as a location for delivering collision warnings, they also found that it improved reaction times when avoiding a barricade. Finally, Riener *et al.* (A. Riener et al., 2012) created a vibrating seat that would instruct drivers to perform various road manoeuvres such as change lanes, accelerate or decelerate. The seat was not empirically evaluated in that study, however Riener (Andreas Riener, 2012) evaluated a further use case of continuous feedback on fuel consumption to achieve fuel-efficient driving. He found in an on-road study that harmonic and low intensity vibrations for low consumption driving and disharmonic and higher intensity ones for high consumption indeed led to more fuel-efficient driving. Harmonic vibrations had similar frequency to the one used in medicine for releasing spasm of muscles and massaging, while disharmonic were experimentally determined with a superimposition of two frequencies and were more uncomfortable than harmonic ones. The vibrations in this case achieved better performance when delivered on the seat belt as opposed to the seat pan. However, as also mentioned by the author, the perceived annoyance of constant tactile feedback was not evaluated. This is an important next step, since perceived annoyance can be increased in tactile cues, as will be shown in the experimental chapters of this thesis.

More recently, there has been interest in evaluating tactile signals whose intensity and location change with time. An initial study by Ho, Spence & Gray (Cristy Ho, Spence, & Gray, 2013) confirmed the good results of looming intensity for audio (described in the previous subsection (Robert Gray, 2011)), while no additional benefit of looming intensity was found in the tactile modality. Gray, Ho & Spence (Rob Gray, Ho, & Spence, 2014) however found decreased response times in the tactile modality compared to constant pulses, when looming intensity was combined with apparent motion towards the drivers' head, created by a vertical array of three tactile actuators (Tactors) attached on the abdomen and activated in an upward manner. A variation of apparent motion was tested by Meng *et al.* (F. Meng, Gray, Ho, Ahtamad, & Spence, 2014; Fanxing Meng, Ho, Gray, & Spence, 2014) by activating vibrotactile cues first on participants' hands and then on their torso, creating a sense of cues moving towards the torso (see Figure 2-12). This intervention produced lower response times compared to static cues, while looming intensity of vibration showed again no additional benefits. The above studies present an interesting application of varying tactile intensity or location to alert drivers.

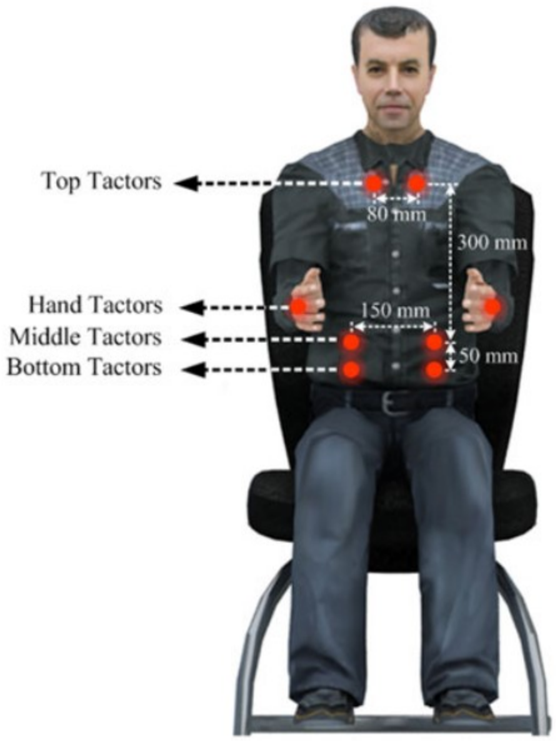


Figure 2-12: The configuration of tactors used by Meng *et al.* Image taken from (Fanxing Meng et al., 2014).

While not directly related to driving, the use of Tactons, tactile icons of varying frequencies, duration, amplitudes and rhythms were introduced by Brewster & Brown (S. Brewster & Brown, 2004), and has been used in a wide range of applications since. Brown, Brewster & Purchase (L. M. Brown, Brewster, & Purchase, 2006; L. M. M. Brown, Brewster, & Purchase, 2005) investigated the use of Tactons as a means to convey more complex information with the tactile modality. Varying parameters such as rhythm and roughness (amplitude modulation on the original waveform that provides the vibration) of such messages enabled the design of richer cues without cost in their recognition accuracy, as long as a reduced number of different levels for roughness was used. Hoggan and Brewster (E. E. Hoggan & Brewster, 2006) extended this work and evaluated parameters of Tactons when used in conjunction with audio messages. They suggested rhythm and roughness of vibration as a means to convey information effectively when used along with audio. Further, the use of intensity was regarded as requiring further investigation in the cues. These findings, although having high potential of increasing the informative content of cues have not been extensively applied in the driving context.

The potential for transferring some speech features into vibration has also been investigated outside the driving context and can provide another means to convey useful information. Applying these ideas to driving could be helpful, in order to design richer cues addressing a wider variety of car scenarios. Spens *et al.* (Spens et al., 1997) developed a handheld tactile aid to lip reading, which vibrated synchronously to speech and was designed for people with hearing impairments. Li (K. A. Li, 2009) investigated the use of synthesized Tactons, consisting of repeated pulses of varying duration and intensity. They were mapped through forced choice responses to simple speech messages that frequently occur when texting (*e.g.* *"hello?"*, *"goodbye."*, *"where are you?"*). The number of syllables, intonation and stress of the spoken messages were identified as important features to be mapped through vibration. The number of syllables was mapped to number of pulses, while the intonation and stress were mapped to vibration intensity. Salminen *et al.* (Salminen et al., 2012) investigated the use of audiotactile messages, where the vibration mimicked the amplitude changes of speech. The audiotactile messages were presented through a handheld device of form factor similar to (Spens et al., 1997). Participants rated the audiotactile messages as more arousing and dominant compared to the audio ones. Tuuri, Eerola & Pirhonen (Tuuri, Eerola, & Pirhonen, 2010) used intonation and rhythm of speech messages to create pure tones that were then delivered either through audio or through vibration. The messages *"Slow"*, *"Urge"* and *"Ok"* showed high recognition rates in both modalities. The authors conclude that the two modalities can be used interchangeably for interface design.

The above studies showed the benefits of using tactile cues to alert drivers about various events. Whether located on the steering wheel, the torso or a combination of these locations, they have shown to create elevated alertness when responding to varying road events. In this thesis' experimental work, locations that would not present the risk of absence of contact were chosen, *i.e.* the abdomen and the wrist, to achieve a higher saliency of the tactile displays designed. The positive effect of directionality in the presented cues was also utilised by delivering them in locations in accordance with the event signified. These choices will be discussed and elaborated in the experimental chapters. Further, an investigation of the use of Tactons in driving, and especially Tactons that can convey speech features was conducted, using ideas from studies presented in the previous paragraph. As discussed in these studies, the effects of these patterns can be even more pronounced when they utilise the audio modality as well. This finding is often observed in driver displays, therefore it will be discussed further in the next section, reviewing studies using multimodal driver displays.

2.2.4. Multimodal Combinations of Displays

As recently discussed by Hass & Van Erp (Ellen C Haas & Erp, 2014) the use of multimodal warnings can be particularly effective to convey risk. Several studies have used multiple modalities as warnings and consistently found improvements over unimodal warnings. This section will discuss a set of such studies and conclude with the first research question of this thesis, motivated by the limited available work in evaluating audio, visual and tactile displays in all their unimodal, bimodal and trimodal combinations. Due to the variety of possible combinations of displays in the candidate modalities, it was considered more straightforward to present the studies of interest in order of time. Some studies that do not use multimodal combinations of displays, but rather compare unimodal displays of different modalities will also be reviewed.

As mentioned in the previous section, since tactile displays for cars gained higher popularity as a research topic only in the last two decades, earlier multimodal studies focus mostly on audio and visual displays or their combinations. Mollenhauer *et al.* (Mollenhauer et al., 1994) used visual or auditory displays to signify road signs during a simulated driving task. It was found that auditory presentation resulted in better recall of the signs but lower performance and increased ratings of perceived distraction and annoyance. This disadvantage of audio was unexpected, however the lack of description of the cues used in this study as well as the fact that this finding is not discussed by the authors makes it difficult to speculate on possible reasons for this. Dingus *et al.* (Dingus et al., 1997) extended the investigation mentioned in Section 2.2.2 and added voice instructions to the visual bar display depicting the headway to the lead vehicle (Figure 2-8.b). They compared the unimodal visual display, unimodal audio instructions (a synthesised voice saying *"Look ahead"* when distance was moderately close and *"Brake!"* when it was very close), and their combination. The observed headways were safest in the multimodal combination.

A similar advantage of the audio and visual combination was found by Liu (Y. C. Liu, 2001). A Head-up Display was used in a simulator study to provide a variety of navigational information, as well as vehicle and terrain-related messages, while the equivalent information would be uttered by synthesized speech. It was found that when using audio and visuals combined, adherence to speed limits was higher, perceived workload was lower and lane keeping behaviour was better. Participants also preferred the multimodal display to the unimodal ones. Liu's study is one of the earliest driving studies that acknowledge the breadth of information of varying urgency that may be provided to the driver. Both warning messages (*e.g.* high engine temperature) and less urgent ones (*e.g.* instruction to turn right) were presented, however the appropriate warning design to convey more or less urgency was not discussed. In the time of Liu's study, Edworthy, Loxley & Dennis (Judy Edworthy et al., 1991b) had already provided guidelines as to how to design urgency in audio warnings (see also Section 2.3). However, such guidelines were adopted later by vehicle warning designers, while their applicability in all combinations of audio, visual and tactile displays was not explored, which motivated the first research question of this thesis.

Van Erp & van Veen (Erp & Veen, 2001) presented a study combining tactile and visual stimuli, as well as unimodal cues, to present route guidance in a simulated driving task. They used vibrations on the left side of the seat pan to indicate a left turn, on the right side for right turn and on both sides for going straight. Similar instructions were presented visually with text and arrows on a display next to the steering wheel. It was found that unimodal tactile created the lowest perceived workload and earliest compliance to the instructions. In this case, the bimodal combination did not perform better than the tactile only cues, which was accounted to possible time costs of checking both displays to perform the appropriate action. In a further simulator study however (Van Erp & Van Veen, 2004), the authors did observe quicker reactions to the multimodal combination of displays. These studies show some evidence of the added value in combining tactile cues with other modalities to alert drivers effectively, which became more prominent in literature thereafter.

Ho, Spence & Tan (Cristy Ho, Spence, et al., 2005) presented a study comparing all three modalities in terms of reaction time to a simulated collision, with a setup similar to (Cristy Ho, Tan, et al., 2005), where participants reacted to a set of videos depicting imminent collision events (see Figure 2-13). All cues were unimodal; the sound of a car horn, a vibrotactile cue and a LED light for the visual cue. It was found that the vibrotactile cue elicited the quickest reactions. As in (Cristy Ho, Tan, et al., 2005), the cues were localised depending on whether the threat was from the back or the front. In this study vibrotactile and audio cues originating from the back elicited quicker reactions compared to the front. This confirmed the utility of encoding spatial characteristics in cues described in other studies. A further study by Ho, Tan & Spence (Cristy Ho, Tan, & Spence, 2006) with the described setup showed however that audio cues (a car horn sound) presented from the direction of a threat also facilitated visual attention (quicker recognition of a change in number plate colour of the approaching car), while vibrotactile cues did not. This led to the suggestion that using vibrotactile cues can elevate alertness in both directions, but that they may not succeed in also shifting visual attention towards the target for more complex visual events. Using bimodal signals, Ho, Reed & Spence (Cristy Ho et al., 2007) showed the potential of audiotactile presentation in front-to-rear-end collision warnings, using vibration on the torso and a car horn sound. These bimodal warnings led to lower reaction times in a simulated driving task compared to the unimodal variants. Finally, using audio, tactile as well as visual modalities for alerting drivers, Scott & Gray (Scott & Gray, 2008) found that unimodal vibrotactile warnings on the abdomen, simulating seat belt warnings, can induce quicker reactions in a critical driving situation compared to an abstract tone or to a visual warning on the dashboard through a LED array. These studies shed further light in the utility of cue directionality, and in how vibrotactile cues can be particularly alerting compared to unimodal audio and visual ones, while increasing the number of modalities can improve reactions.

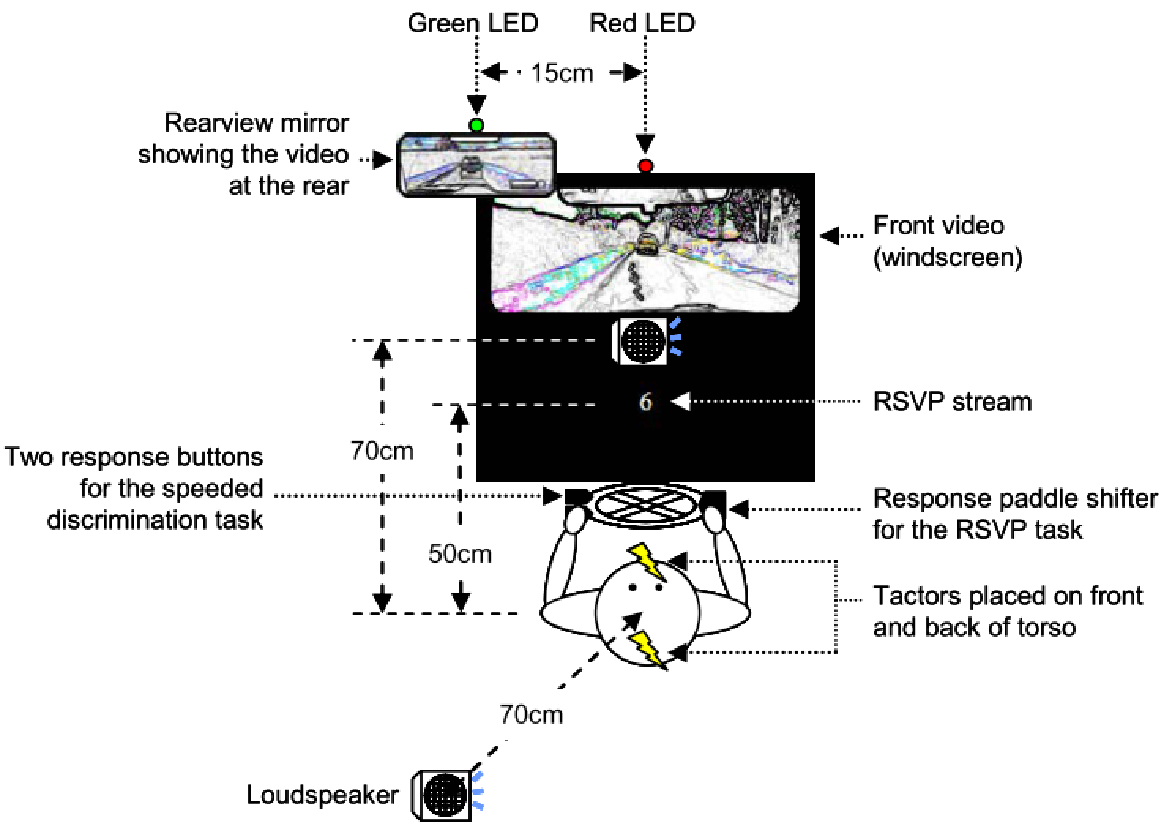


Figure 2-13: The setup of the experiment in Ho, Spence & Tan (Cristy Ho, Spence, et al., 2005). Since there was no driving simulator used in the study, participants were instead engaged in a Rapid Serial Visual Presentation (RSVP) task, where target digits would be identified in a stream of distractor letters. Image taken from (Cristy Ho, Spence, et al., 2005).

As mentioned earlier, an important factor when designing warnings is to consider the criticality of the signified event. In the studies described so far there is little consideration in this factor. Before presenting a set of available design guidelines for varying cue urgency, a set of studies using multimodal displays that addressed scenarios varying in criticality will be described. Kaufmann *et al.* (Kaufmann, Ohg, Risser, Geven, & Sefelin, 2008) presented a set of guidelines for the use of audio, tactile and visual warnings along three priority levels. The authors defined high priority warnings as requiring immediate action, while medium required no immediate reaction and low priority ones had no immediate relevance to the driving task. Audio and tactile modalities were suggested as suitable for high priority messages, visual and tactile for medium and audio and visual for low priority ones. The suggestions were based on studies measuring speed and steering performance of participants and from a workshop with experts. Lindgren *et al.* (Lindgren et al., 2009) investigated a set of integrated visual and auditory warnings for events in the driving task varying in criticality. Auditory warnings similar to commercially available ones for collision avoidance, lane departure and speeding on a curve were provided with or without advisory warnings in a driving simulator. The advisory warnings were visual indicators, graphically showing the distance of the car to a vehicle in front, behind or relative to a curve. It was found that the presence of warnings did not influence the driving speed or how often drivers moved their gaze off the road. Additionally, participants drove with higher average lateral deviation from the centre of the lane when no warnings were present, indicating a lower level of vigilance. These studies indicate that criticality in cues is recognised as a factor to consider. However, warnings are only using a subset of possible multimodal cues and urgency is not explicitly considered in the cue design.

Presenting some more considerations on the importance of the event signified, Cao *et al.* (Cao, van der Sluis, Theune, op den Akker, & Nijholt, 2010) investigated the use of audio and tactile cues conveying four different levels of urgency. Number of pulses and inter-pulse-interval were manipulated for all cues to signify urgency (see Figure 2-14). Additionally, pitch was manipulated for the audio cues and intensity for the tactile ones. The main task in this study was visual tracking with different levels of auditory distractions (namely radio, conversation and noise) but no driving task was simulated. A general trend of higher urgency = faster response was found, indicating that the designed urgency of the cues was successfully perceived. Vibration cues were also identified more accurately but sound cues more quickly. Finally, sound cues were reported as easier to distinguish by the participants. Cao *et al.* (Cao, Mahr, et al., 2010; Cao, Theune, & Müller, 2010) investigated the use of speech, abstract audio cues and visuals for presenting road obstacle warnings in a simulated driving task. Speech messages along with pictures led to good recall of the signified events but slow reaction times. Thus, the authors suggested the use of speech along with pictures of signified events for tasks not requiring quick responses, such as navigation. Speech along with pictures was also perceived as most useful in various driving contexts, *i.e.* low visibility, under fatigue and high demand. The speech cues used in these studies were relatively long, *e.g. "Broken vehicle in 180 meters on the right roadside"*, resulting to longer utterances. This may be a limitation when quick reactions are needed.

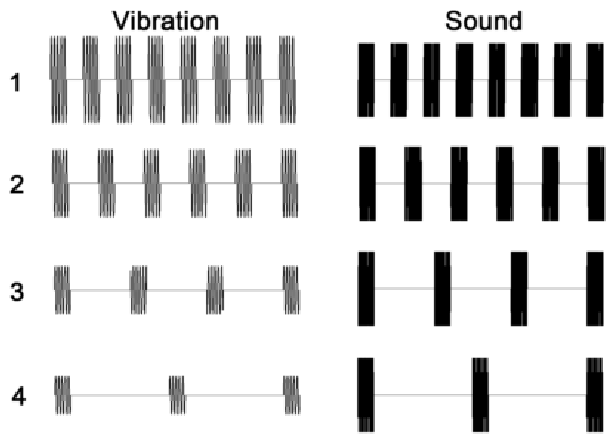


Figure 2-14: Illustration of the design of sound cues and vibration cues in Cao *et al.* (Cao, van der Sluis, et al., 2010). Numbers indicate priority levels. The x and y axes of each oscillogram are time and amplitude. This study applied guidelines by Edworthy, Loxley & Dennis (Judy Edworthy et al., 1991b), to vary criticality in the cues, to be discussed in the next section. Image taken from (Cao, van der Sluis, et al., 2010).

To conclude, this section has presented studies using warnings along the audio, visual and tactile modalities and some of their combinations. There are two factors that are not thoroughly examined in the presented work. Firstly, the cues have been combined in several cases and have shown some improved utility when this has been done. However, there has been no attempt to use all possible unimodal, bimodal and trimodal combinations of the cues and examine their effectiveness as driver alerts. This would result to truly multimodal cues, at least in these three modalities examined. Secondly, although the urgency of the event signified has been considered, there has been no attempt to systematically design urgency in all combinations of multimodal cues, varying levels of criticality to alert about situations of varying relevance to the main driving task. The first factor is what motivated the first part of this thesis' research question: *How do **multimodal** driver displays varying in urgency affect performance?* The second factor motivated the second part: *How do multimodal driver displays **varying in urgency** affect performance?* The next section will review available work on designing cues of varying urgency and further elaborate on how combining multimodal cues that vary in urgency is a relevant research direction.

2.3. Designing Urgency in Displays

As has been reported above, the driving task is a rich one, with a variety of events that may arise. However, these events are not always critical or even relevant to the task itself. For example, an imminent collision event is critical and can affect the driver substantially, a low level of fuel can have less imminent and less catastrophic consequences, while an incoming message may not affect the driving task at all or may even disrupt it when signified. Therefore, it is expected that events can be signified as more or less urgent in driving, by varying the warning design in a way that their perceived urgency will match the urgency of the event. Even from the earlier descriptions of alarms as a Human Factors strategy (see section 2.1.4 and (Stanton, 1994)), the importance of reflecting the criticality of the signified event in the alarms was highlighted. This section will review a set of studies with available guidelines on how to design urgency in audio, visual and tactile warnings. The absence of work in combining these guidelines multimodally to create warnings that convey different urgency levels in all combinations of audio, visual and tactile modalities motivated the second part of RQ-1: *How do multimodal driver displays* ***varying in urgency*** *affect performance?*

There have been numerous studies investigating how signal parameters of auditory warnings relate to their perceived urgency. These studies have been largely focused on abstract pulses, consisting of pure or alternating tones, since these signals have easily measurable qualities, which are also easy to vary. Such qualities are the sound level and loudness (the perceptual property of the sound level), the frequency and pitch (the perceptual property of the frequency), the timbre (the perceptual property of the sound as a whole, which distinguishes it from another sound), as well as the interpulse interval. Edworthy, Loxley & Dennis (Judy Edworthy et al., 1991b) & Hellier, Edworthy & Dennis (E. J. Hellier et al., 1993) showed that higher fundamental frequency, higher speed (an interpulse interval that decreases in time) and larger pitch range can increase the perceived urgency ratings of auditory warnings. Hass & Casali (E. C. Haas & Casali, 1995) reported that higher loudness and smaller interpulse interval can also achieve a higher perceived urgency effect, while they found quicker responses to stimuli of higher designed urgency. Haas & Edworthy (E.C. Haas & Edworthy, 1996) extended these results by suggesting high frequency, high speed, and high level of loudness as parameters that increase perceived urgency ratings and decrease response times. Edworthy *et al.* (J. Edworthy, Hellier, Walters, Weedon, & Adams, 2000) also observed significantly lower response times for highly urgent warnings designed according to the above guidelines, compared to warnings designed to be of medium and low urgency levels, confirming the successful urgency design. The results acquired from the above studies were based on subjective ratings of the warnings by participants, while the response times were to simple button pressing tasks. This was a valid technique in order to acquire results on how such cues were perceived out of a specific context, however the application of these guidelines in the automotive domain would further increase the efficacy of the cues as driving alerts.

Wiese & Lee (E. Wiese & Lee, 2001) & Marshall, Lee & Austria (D. Marshall, Lee, & Austria, 2001) used the above guidelines to design in-vehicle warnings for scenarios varying in urgency, *i.e.* an email alert and a collision avoidance alert. Confirming these guidelines on warning design, they also suggested that the annoyance of the alerts used is also an important factor to consider. They found that perceived urgency increased for alerts designed for more urgent scenarios, while reaction times to a simulated driving task decreased (E. Wiese & Lee, 2001). Perceived annoyance also increased but with a lower magnitude (D. Marshall et al., 2001). A similar result was observed by Wiese & Lee (E. E. Wiese & Lee, 2004), who found a relation of increased perceived urgency and increased perceived annoyance and perceived workload for these alerts. Reaction times to critical alerts were lower, however the authors advised against the overuse of alarms when not needed, so as to avoid high annoyance that could result in disabling of the alarms.

Marshall, Lee & Austria (D. C. Marshall, Lee, & Austria, 2007) also demonstrated how higher pulse duration and lower interpulse interval increased ratings of urgency of in-vehicle audio alerts. The sound cues investigated were similar to the above studies, namely impending collision, navigation and email messages. It was also found that ratings of urgency were positively correlated with ratings of appropriateness for critical warnings (collision alerts), while ratings of annoyance were negatively correlated with appropriateness for non-critical ones (e-mail alerts). This showed how annoyance needs to be considered when warning designed urgency decreases. More recently, Gonzalez *et al.* (Gonzalez, Lewis, Roberts, Pratt, & Baldwin, 2012) found that fundamental frequency, pulse rate and intensity of warning sounds positively influenced the ratings of urgency and annoyance of participants. However, pulse rate was suggested as the most suitable for conveying events of varying urgency, since it did not elicit such high ratings of annoyance. These studies show how the guidelines developed for auditory alerts can generalise in driving, but also how perceived annoyance is an important factor to consider when designing auditory warnings.

Other than abstract sounds, speech is another means that urgency can be conveyed with audio. It differs from abstract sounds in the sense that it conveys some semantic association to the signified event, which may be a useful alarm strategy. Studies that investigate how to design urgency into speech messages appeared later, so their adoption by the automotive warnings domain was quicker. Hellier *et al.* (E. Hellier et al., 2000; Elizabeth Hellier et al., 2002) found that acoustics and speaking style influence the ratings of urgency in speech messages. Signal words spoken urgently created higher ratings compared to non-urgently, which in turn were higher compared to words spoken in a monotone manner. Female speakers induced higher urgency ratings and a higher range in these ratings compared to males. Finally, the word *"Danger"* was perceived as highly urgent, matched only by the word *"Deadly"*. Additionally, it was found that an urgent utterance of the messages resulted in louder sounds, with higher pitch and pitch range. Edworthy *et al.* (Judy Edworthy, Hellier, Walters, Clift-Mathews, & Crowther, 2003) extended these findings, observing that signal words spoken urgently are perceived as more urgent, believable and appropriate. Comparing speech and non-speech warnings, Edworthy *et al.* (J. Edworthy, Walters, et al., 2000) found no difference between speech and non-speech cues of equal urgency, encouraging the use of both types of cues in warning design.

Applying these results in the driving context, Baldwin & Moore (C. L. Baldwin & Moore, 2002) investigated the signal words *"Danger"*, *"Warning"*, *"Caution"* and *"Notice"* when used together with different driving related speech messages (increase speed, decrease speed, following too close, close vehicle on right-left and vehicle tailgating). They found that the signal word *"Danger"* was perceived as more urgent compared to the words *"Warning"* and *"Caution"*, which in turn were perceived as more urgent compared to *"Notice"*. It was also found that a higher Signal to Noise (S/N) ratio (a louder presentation of the sound) when presenting warnings positively impacted ratings of urgency and alerting effectiveness, without strong impact in annoyance. Higher S/N ratio also positively affected the ratings of urgency, regardless of the semantic content of the collision avoidance messages. Similar effects were observed later by Baldwin (Carryl L Baldwin, 2011) in terms of reaction times, where participants responded quicker to urgent warnings, created by using urgent words and high signal intensity. Baldwin & May (Carryl L. Baldwin & May, 2011) found an interaction effect in terms of loudness between the critical signal word *"Danger"* and the non-critical one *"Notice"*. When artificially varying the intensity of the words, so that the low urgency warnings had high intensity and *vice versa*, the number of crashes signified by these words was reduced. Presenting the urgent word *"Danger"* in high intensity and the non-urgent word *"Notice"* in low intensity did not achieve this effect. This was an effect that pointed towards a startle effect of the high urgency word and therefore it was suggested that loudness should not be used as the sole characteristic of verbal warnings in vehicle alerts. The above studies extended the understanding of designing urgency in audio to verbal messages. As with the non-verbal ones, the findings of these studies were used in the experimental work of this thesis to vary the designed urgency of the designed audio cues.

Designing visual cues to convey different levels of urgency has been studied, mainly in the context of safety signs, but has not been as widely applied in driving. Specifically, the colour and text to be used has appeared in different national and international standards, see for example ANSI Z535 (ANSI, 2011). A general guideline in ANSI Z535 is that red denotes danger, orange denotes warning and yellow denote caution (see Figure 2-15). Chapanis (Chapanis, 1994) studied how colours were perceived in terms of urgency and found red, orange, yellow and white to denote decreasing levels of hazard. He also reported that words and colours for warnings signs were subjectively associated with a similar trend: red for danger, orange or yellow for warning, yellow or orange for caution. Braun & Silver (Braun & Silver, 1995) also observed a higher perceived urgency for colours red and orange, as well as that the effect of urgency of signal words was affected by the colour in which they were presented, for example the word "Danger" presented in green was not perceived as dangerous as when printed in red. Wogalter *et al.* (Wogalter, Kalsher, Frederick, Magurno, & Brewster, 1998) observed a different perception for yellow and orange, with yellow exceeding orange in perceived urgency of warning labels. More recently, Chan & Ng (Chan & Ng, 2009) observed an elevated perceived urgency of red flashing lights compared to yellow and blue. The above studies reveal a general consensus regarding the utility of colours to convey urgency, especially when combined with signal words. It has to be noted, however, that these studies are in the context of warning signs and labels and do not use a specific applied setting. Although the colour of stimuli has been used in vehicle studies, described in Section 2.2.2, the ones that vary colour to convey urgency are rare and mostly use multimodal signals (see below). This shows the opportunity to evaluate the usage of colour as a warning mechanism in different modality combinations, so as to enrich alarm design and create warnings suitable for scenarios of varying criticality.

Figure 2-15: ANSI Z535 signal word panels for hazard alerting labels. Image taken from: http://incompliancemag.com/article/designing-effective-product-safety-labels-how-to-convey-risk-severity-levels/.

In terms of conveying urgency through the tactile modality, literature is also limited and even more so in the context of a driving task. In general, though, there is consensus that higher intensity and lower interpulse interval are parameters that increase perceived urgency of vibration. White and White & Krausman (White, 2011; White & Krausman, 2012, 2015) studied the scenario of dismounted soldier movements using a tactile belt for navigational instructions. They compared a steady pulse (with interpulse interval of 0 *ms*) with a pulse having 500 *ms* interpulse interval at two intensity levels. They observed higher ratings of perceived urgency when the pulse was steady and intensity was high. Pratt *et al.* (Pratt et al., 2012) & Lewis, Eisert & Baldwin (Bridget A Lewis et al., 2014) reported a similar observation, where pulse rate was found to positively influence the ratings of perceived urgency and to have less impact on the ratings of perceived annoyance. In (Bridget A Lewis et al., 2014), the warnings were also used in a driving context and delivered in the wrist, the waist or the seat pan, which resulted in strikingly similar ratings according to the authors. In the context of mobile phones, Saket *et al.* (Saket, Prasojo, Huang, & Zhao, 2013) also observed that shorter interpulse interval and pulse duration contributed to higher ratings of urgency, while they reported that four different levels of urgency were clearly recognised by participants. Finally, Li & Burns (Y. Li, Burns, & Li, 2013) suggested that an increased number of tactile actuators can also consistently increase perceived urgency. These studies indicate that urgency can be designed in warnings also in the tactile modality. The fact that most of these studies are very recent reveals the relevance of this topic for future research. The key features of interpulse interval and intensity were widely used in the experimental work of this thesis. Further, the observed interchangeable character of warning locations in order to achieve the desired urgency effect support the use of the waist or the wrist as warning locations in the experiments conducted.

As a next step to designing urgency in a single modality, Baldwin *et al.* (C. L. Baldwin et al., 2012) and Lewis & Baldwin (B. A. Lewis & Baldwin, 2012) evaluated the perceived urgency of unimodal signals from audio, visual and tactile modalities. Urgency ratings were mapped in a single scale to decide on features that can be used across modalities to vary urgency. This initiated the creation of a crossmodal urgency scale (see Figure 2-16). Pulse rate (flash rate for visual signals) was suggested as an effective parameter to vary urgency across these three modalities. Intensity and frequency were additionally used for audio signals, and word choice and colours for visual ones. In these studies, Baldwin *et al.* used the colours red, orange, yellow and green in order to convey warnings of decreasing urgency, referring to previous guidelines. They also mentioned that there is limited information regarding the impact of presenting warnings of multiple modalities to drivers in varying urgency contexts. Similar results were presented by Baldwin & Lewis in the presence of a simulated driving task (Carryl L. Baldwin & Lewis, 2013), where manipulating tactile interpulse interval was found to be reflected stronger in urgency ratings and less so in annoyance ratings. Therefore, it was suggested as more suitable compared to varying colour for visuals and fundamental frequency for audio. More recently, van Erp, Toet & Janssen (van Erp, Toet, & Janssen, 2015) extended this finding in all combinations of audio, visual and tactile modalities and also found that interpulse interval affects perceived urgency.

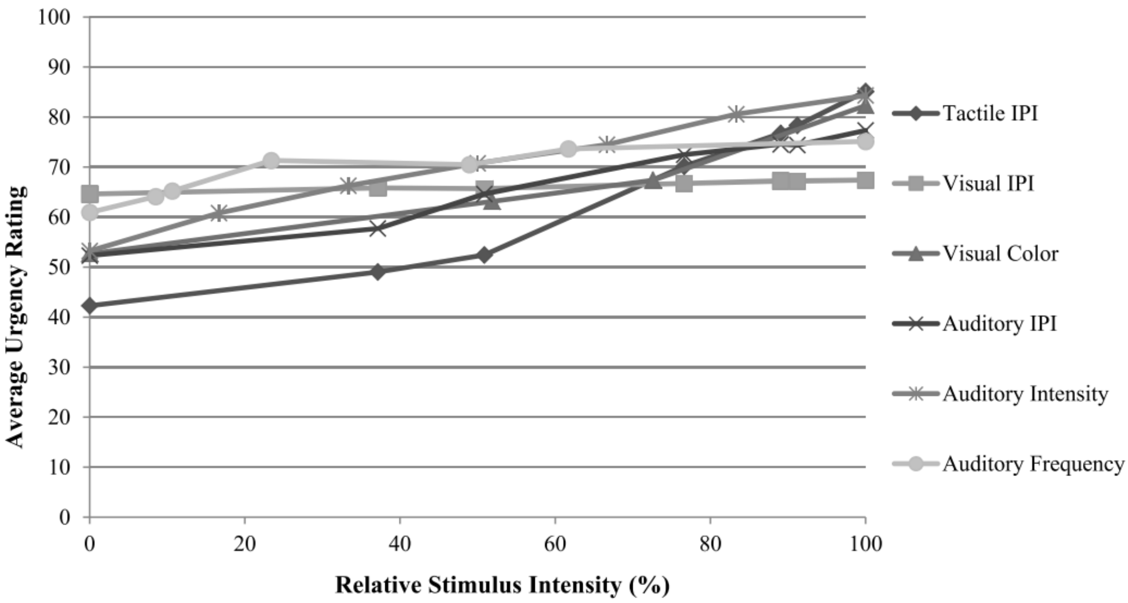


Figure 2-16: Urgency ratings for all levels of stimulus intensity (IPI stimuli are reversed meaning that the highest relative intensity is actually the lowest interpulse interval in milliseconds). Image taken from (B. A. Lewis & Baldwin, 2012).

In summary, although there have been several studies investigating the performance and perception of audio, visual or tactile warnings in the driving context, these have not been evaluated in all their multimodal combinations. Additionally, guidelines for designing messages of different urgency have not been applied multimodally in a driving simulator, so as to test their subjective and objective responses. Previous studies into the recognition of different levels of urgency of warnings have rarely used simulated driving as the main task. This indicates the relevance of investigating how all combinations of auditory, visual and tactile modalities will influence the ratings of urgency and annoyance in a simulated driving task. This will improve alarm design, by creating warnings that are perceived as appropriately urgent and do not annoy. Further, response times for the recognition of the different signals will extend the available results for audio and investigate how they apply multimodally. This will ensure that alarm reactions are as prompt as needed, especially for alarms signifying critical events. The research potential revealed in this and the previous section motivated the first research question of this thesis:

How do multimodal driver displays varying in urgency affect performance?

To address this question, two experiments were designed, investigating perceived urgency and annoyance, as well as recognition times of all multimodal combinations of audio, visual and tactile cues in a driving simulator. As shown in Section 2.2, combining the cues in all possible ways is a topic that has not been adequately investigated. This would reveal the utility of the cues in a much richer experimental context, allowing the creation of guidelines in a broader set of modality combinations. Additionally, assessing perceived urgency and annoyance of this set of multimodal cues would fill a significant research gap, as described above, providing insights into how effectively the above guidelines can be applied in multimodal cues, in a way that urgency in cues is effectively recognised and reacted to, while annoyance is kept at acceptable levels. The experiments designed to address RQ-1 are Experiments 1 and 2, described in Chapter 4.

2.4. Comparing Abstract and Language-based Multimodal Displays

Having presented the utility of multimodal warnings and of varying urgency in the warnings, a next meaningful step is to investigate the warning design in terms of its semantic content. As presented in Section 2.2, there is a variety of different semantic properties that can be used in warnings. Icons, pictures or text for the visual modality, Earcons, Auditory Icons, abstract sounds or speech for audio and varying tactile patterns for vibration are some of the cues that have been utilised, and have been discussed earlier. However, the semantic properties of these warnings varied in a way that was responding to the particular situation each study was signifying; there was less consideration on comparing information communicated in the warnings in fundamentally different ways, for example using speech versus abstract pulses. The properties of the warning signal were a concern when designing urgency in the cues, but again the semantics of the cues were not fundamentally different with the interventions; they were either more or less urgent abstract messages or more or less abstract speech messages. Therefore, an opportunity arises in comparing different ways of presenting messages that vary in urgency. This will expand the alarm design space, by presenting guidelines for warnings of varying message content, to be used depending on the demands of the driving context. For example, when the signified event is expected to be obvious to the driver (*e.g.* a threat approaching from the front) abstract cues can be a viable alarm strategy, while in cases where the error is unknown (e.g. a mechanical fault) language-based cues might provide useful information on the error. In this section, a set of studies comparing warnings of different semantic content will be presented, concluding with the next research question of this thesis, motivated by the limited available work on this topic.

Dingus *et al.* (Dingus et al., 1997), as mentioned earlier, used a car icon or bars increasing in size to denote the distance from a lead vehicle (see Figure 2-8). Both displays were equally effective in helping participants maintain a safe headway to the lead car. However, Hirst & Graham (Hirst & Graham, 1997) who used a similar visual setup and combined the visual cues with audio ones (an single tone or a voice saying *"Danger Ahead"*) found some different results. The abstract visual cue along with the speech message reduced breaking reaction time to a simulated critical event the most, while the abstract visual cue along with the abstract tone were preferred the most. This is one of the first studies detecting how speech and non-speech warnings augmented by visual cues can produce different reactions.

As reviewed earlier, Graham (Graham, 1999) found an advantage of Auditory Icons (a car horn and skidding tires) over a simple tone and a simple speech message saying "Ahead", although the auditory icons produced more inaccurate responses to assessing the criticality of a road video. This comparison showed the strength of such audio metaphors, however there was no urgency manipulation in the cues and the voice was uttered in calm manner, which contradicts the guidelines of Hellier *et al.* (Elizabeth Hellier et al., 2002) for critical speech warnings. The observed advantage of speech cues (the words *"front"* and *"back"*) over a car horn sound found by Ho & Spence (Cristy Ho & Spence, 2005) in a more ecologically valid threat detection task further highlighted the utility of speech as a warning modality. As stressed by the authors of that work, speech cues need to easily convey the intended message and when this is hard to achieve, non-speech cues should be used. An example illustrating this problem is Di Stasi's *et al.* work (Di Stasi et al., 2010) where emotional voices unrelated to the task at hand (a woman's scream and a baby's laugh) displayed poor performance compared to abstract pulses in warning motorcyclists of imminent collisions. Finally, Baldwin (Carryl L Baldwin & May, 2014) found that an abstract tone was equally effective as the word *"Danger"* to mitigate collisions.

A further study reporting rich comparisons between audio cues of different semantic content is the work from McKeown & Isherwood (Denis McKeown & Isherwood, 2007), reviewed in Section 2.2.1 (see Figure 2-17). There, Auditory Icons with association with driving (*e.g.* a car speeding past) and speech (*e.g. "Exceeding speed limit"*) outperformed abstract tones in terms of identification time and accuracy. Speech was also perceived as more pleasant and less urgent compared to the abstract sounds. McKeown, Isherwood & Conway (D. McKeown, Isherwood, & Conway, 2010) later compared repetitive pulses and a gunshot sound with the sound of screeching brakes and found shorter response times when participants reacted to the latter, which had a higher association with driving. The importance of associating the content of auditory warnings was therefore further highlighted.

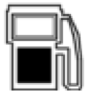
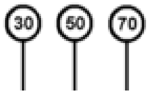
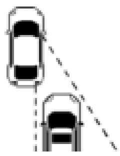
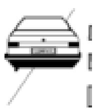
Driving Events	Pictures of Referent Events	Abstract Sounds	Auditory Icons	Nonspecific Environmental Sounds	Speech Messages
Petrol level is low		Single bell ding	Water pouring	Footsteps	"Petrol is low"
Oil level is low		Low-rate tapping	Steam and water sounds	Housemartin song	"Oil is low"
Tire pressure is low		Low-rate, low-pitched warbling tone	Air release blast	Seashore lapping	"Tire pressure is low"
Driver door is open		Moderate-rate siren	Car door shutting	Cockerel	"Driver door is open"
Speed limit is being exceeded		Moderate-rate fire bell	Car speeding past	Baby sneeze	"Exceeding speed limit"
Hand brake is on while driving		Moderate-rate tone alarm	Squeaking sound	Tram passing	"Hand brake is on"
A car is driving in the blind spot		High-rate, high-pitched warbling tone	Car horn blasts	Electric warble factory sound	"Car in blind spot"
The car is drifting off the road		High-rate, high-pitched tone alarm	Driving over "rumble strips"	Small glass smashed	"Drifting off road"
Headway to the vehicle in front is closing fast		High-rate, high-pitched zapping pulses	Car crashing	Textile factory loom	"Headway closing fast"

Figure 2-17: Driving Events, Referent Pictures, and Stimulus Descriptions used in McKeown & Isherwood (Denis McKeown & Isherwood, 2007). Image taken from (Denis McKeown & Isherwood, 2007).

Cao *et al*. (Cao, Mahr, et al., 2010; Cao, Theune, et al., 2010) utilised more modalities and used speech, abstract audio cues and visuals to present road obstacle warnings in a simulator. Speech combined with pictures led to high recall of the signified events when asked about them after the experiment, and low reaction times. The use of speech along with images was therefore suggested by the authors for tasks not requiring imminent responses, such as navigation. In more demanding situations, like low visibility and under fatigue, speech and images were also perceived as most useful. One of the limitations of these studies, as mentioned by the authors, is the relatively long utterances of the speech cues, some of which were as long as ten words.

The above work highlights a set of varying results that abstract and speech warnings can create. Often, speech displays particularly good results, however it suffers from the limitation of cue length, as also stated by Nees & Walker (Nees & Walker, 2011). Conversely, abstract cues are more flexible in conveying a variety of messages, however their semantic power is lower. Direct comparisons of the two are rare and have displayed little difference in warning effectiveness (Carryl L Baldwin & May, 2014; J. Edworthy, Walters, et al., 2000). A further limitation of current research is that abstract and language-based cues are mostly compared in relation to audio. No attempt has been made to design cues that transfer some features of speech to the tactile modality and compare speech, text and the new speech-related tactile cues with their abstract counterparts. Therefore, the opportunity presented is twofold: firstly, one can design and evaluate truly multimodal cues varying in message content (abstract and language-based cues) in the audio, visual and tactile modalities, using all the unimodal, bimodal and trimodal combinations of the cues. Secondly, one can compare these new cues with already available abstract cues and evaluate their effectiveness. In this way, the appropriate cue design for the appropriate driving context will be suggested, while the merits of combining cues of varying message content multimodally will be assessed. For consistency, the term "abstract" will be used for cues with a less semantically meaningful content, *i.e.* abstract pulses, colours or vibrations. The term "language-based" will be used for cues that are related or based on language, *i.e.* speech, text or a vibration that retains some features of speech. This motivates the next research question of this thesis:

How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?

To address this question, four experiments were designed, described in Chapters 6 and 7. In Experiments 4 and 5 (Chapter 6) a set of language-based tactile cues were created, the Speech Tactons, retaining some prosodic features of speech, based on ideas reviewed in Section 2.2.3. These were used in isolation or in combination with the speech cues they originated from and evaluated in terms of perceived urgency, annoyance and alerting effectiveness (Experiment 4) as well as recognition time and accuracy (Experiment 5), in line with prior work. In this way, a new warning mechanism would investigate how speech can be presented in a tactile way and investigate the implications of this new type of cue for driving. Obtaining a more complete picture of the utility of these new warnings, a comparison of the cues with abstract cues in a truly multimodal way would further enrich available knowledge on warning utility in the driving task, by comparing cues varying in message content. Any advantages of abstract or language-based cues would be highlighted and performance of the cues would be directly compared in a simulated driving task. Therefore, Experiments 6 and 7 (Chapter 7) would compare abstract versus language based cues in terms of reaction times and observed driving behaviour during a driving task with no critical events occurring (Experiment 6) and during driving while critical events would occur (Experiment 7). In this way the evaluation would be more complete, as it would assess these warnings in a broader set of situations, in line with prior studies evaluating driver displays.

2.5. The effect of Situational Urgency on Reactions

Warning design has been discussed in terms of modalities used, designed urgency, and semantic content. A next meaningful factor to consider is the situation in which the cues will be used. A well designed warning should improve reactions when delivered along with a critical event, this is why it is created. However, one might be less certain about whether adding the warning actually helps reactions in the context of the critical event unless one controls this event, at least in an experimental setting. If delivering the warnings without any event present (see also (D. Green & Swets, 1966)), presenting the event without any warnings, and both having an event and warnings all show no difference in terms of observed reactions, the warning utility may then be debated. Studies evaluating warnings often compare observed baseline reactions when warnings are absent versus present in order to conclude whether there was an improvement with the warnings (see for example (Chun et al., 2012)). Studies have also considered the influence of having warnings with no critical event present, which, as discussed, can also be described as a false alarm (see for example (Lees & Lee, 2007)). However, there has been no comparison of all possible situation combinations: warnings presented with no critical event (or a false alarm) / a critical event with no warnings (or an absence of alarm) / both a critical event and warnings (or a conventional use of alarm). This would enhance understanding on the influence of warnings; if and how their presence influences reactions compared to their absence when there is a threat, or to their redundant presence when there is no threat. Essentially, what would be varied in such an exercise would be the urgency of the situation the driver would have to face: the situational urgency. This section will review available studies that have tested alarms in various levels of situational urgency, and will conclude with the next research question of this thesis, motivated by the absence of work in testing the alarms in all levels of situational urgency. Many of these studies have been reviewed earlier, in which case only their general goal will be repeated, while the aspects related to situational urgency will be discussed more explicitly.

In a study on how mobile phones affect driver distraction, Alm & Nilsson (Alm & Nilsson, 1995) presented road hazards during a cognitive task with a mobile phone. When participants were judging whether a sentence uttered through the phone was sensible or not (*e.g. "The boy brushed his teeth"* versus *"The train bought a newspaper"*) a critical event was presented in half of the cases. In this way, distraction was evaluated in two situational urgency contexts. This variation was used to study how distraction affects driving in a variety of situations, however results were only focused on how the presence versus absence of conversation affected driving. There were no comparisons between critical and non-critical situations in this study. An aspect of situation severity was tested by Liebermann *et al.* (Liebermann et al., 1995), who measured quicker reaction times to a real versus a "dummy" braking event, where only the brake lights were activated with no deceleration of the lead vehicle. By varying further aspects of the situation tested, van der Hulst (Hulst, 1999) found that reactions were quicker when a lead car deceleration was fast as opposed to slow. Headway to the lead car was also longer when the deceleration was expected, *i.e.* there was another road user on a side road that the lead car might give way to, and shorter when the deceleration was unexpected, *i.e.* there was no obvious reason in the traffic to cause deceleration. A similar effect was observed by Warshawsky-Livne & Shinar (Warshawsky-Livne & Shinar, 2002), who found increased reactions to a braking event when it was happening in random as opposed to regular intervals after an alarm and also when there were false alarms about the event (alarms were verbal warnings by the experimenters). These studies showed the adaptability of driving behaviour to contextual cues, as well as how a high anticipation of an event can improve reactions.

Studying the role of driving complexity, Lee *et al.* (J D Lee, Caven, Haake, & Brown, 2001) observed slower reactions when participants were reacting to a complex as opposed to a simple email system (measured in number of menus and objects per menu) under complex as opposed to simple road conditions (measured in number of road objects encountered per minute). Lee, Hoffman & Hayes (John D. Lee, Hoffman, & Hayes, 2004) reported that reactions to a lead car braking were quicker as the severity of braking increased. Signifying the level of braking severity with tactile-visual or audio-visual alerts varying in intensity to match severity (graded alerts) led to safer minimum distances to the lead car, as opposed to only signifying severe braking events. Some further manipulations of the tested situation were performed by Kramer *et al.* (Kramer, Cassavaugh, Horrey, Becic, & Mayhugh, 2007), who tested an abstract sound and LED lights for threats originating from the front (a sudden appearance of a stationary vehicle in front) or the side (an appearance of a vehicle on the blind spot). They found that the multimodal combination of the cues led to quicker reaction for both types of events compared to the unimodal presentation, while there was no direct comparison of reactions between these events. Similarly, Lindgren *et al.* (Lindgren et al., 2009), as discussed earlier, examined collision avoidance, lane departure and curve speeding situations, mainly focusing on the nature of the warnings in each situation and without comparing results across situations. These studies show that road conditions can affect observed reactions when road complexity and types of threats are varied. This shows that responses are in line with the severity of the situation presented, however the types of situations used are either not compared to each other or to an absence of alarms.

Other than the severity of the driving situation, the alarm validity is another factor that can influence reactions. Lees & Lee (Lees & Lee, 2007) evaluated an abstract audio cue as a warning in a set of varying road events and varying alarm strategies. Audio warnings could be false alarms, occurring with no apparent reason on the road, unnecessary alarms, occurring when a manoeuvre of another road user could potentially create a threat, but the road user did not escalate that manoeuvre to a threat, or actual alarms when the manoeuvre did escalate to a threat. It was found that false alarms led to decreased trust and compliance to the system, while unnecessary alarms increased the number of unnecessary reactions to the simulated events. However, a system with unnecessary alarms helped to create quicker reactions to a subsequent critical event compared to one with accurate alarms or false alarms, making the drivers more cautious on the road. A further set of imperfect alerting systems was investigated by Maltz & Shinar (Maltz & Shinar, 2004, 2007). They found that auditory (speech or abstract tones) and visual (text) false alarms created more unnecessary reactions (Maltz & Shinar, 2004). Auditory tones were also preferred to visual cues or to multimodal combinations of the cues. In a follow-up study (Maltz & Shinar, 2007), participants showed overreliance to the system by maintaining shorter headways when alarms (abstract tones) were always true as opposed to sometimes false.

The above studies showed that the urgency of the situation can have an effect on the observed responses. Drivers adjust their responses to the criticality of the event presented, while the complexity of the situation presented is also reflected to reactions. Further, false alarms create a detrimental effect on driver reactions and damage their impression of the system usability. However, events were not directly compared in all variations of situational urgency. Further, when warnings were used in order to alert drivers, they were not multimodal using all combinations of audio, visual and tactile cues. These results point to a clear research potential in varying the event criticality to investigate reactions under varying levels of situational urgency signified by truly multimodal warnings. In this way, the utility of the designed alarms will be assessed, by evaluating their performance in the presence of situations of varying criticality. This motivated the next research question of this thesis:

How does situational urgency influence responses to multimodal driver displays varying in urgency?

To address this question, Experiment 3 was designed (Chapter 5), investigating the influence of situational urgency in reactions to multimodal warnings. Abstract designed warnings were evaluated in terms of reaction time, accuracy and observed driving behaviour when presented alone, or in combination with a critical event (a lead car braking), as well as when an event was presented with no warnings. In this way, all variations of situational urgency would be evaluated for this particular critical case, and the influence of multimodal cues to signify this event would be directly observable. Investigating cue modality would allow the comparison of reactions to unimodal, bimodal and trimodal cues in different variations of situational urgency and compare reactions to the absence of the cues. This would contribute to available knowledge by investigating both the influence of situational urgency to reactions, as well as the efficacy of multimodal cues to signify criticality. It is noted that this investigation used only abstract cues, since it was performed before the design of language-based cues (Experiments 4 and 5 presented in Chapter 6). However, the results did highlight the influence of situational urgency and their generalizability for language-based cues could be investigated in a straight forward way, by repeating the investigation with the new cues.

2.6. Designing Displays for Autonomous Vehicle Handovers

The previous section discussed the influence of situational urgency in responses to warnings. The situations examined were mainly critical or less critical events of a manually driven car, where the driver was required to constantly control the vehicle and always be attentive to driving. However, in the context of autonomous cars, this is less the case, since drivers are not expected to constantly monitor the road. A car that does not require a driver is not a new idea. Houdina Radio Control, a piece of radio equipment that could add autonomous capabilities to a car, was developed in 1925 ((“Science : Radio Auto,” 1925), see Figure 2-18.c). A major project in Ohio State University examining the feasibility of automated cars, which managed to solve several aspects of automation, was initiated in 1964 and concluded in 1980 ((Fenton & Mayhan, 1991), see Figure 2-18.a). However, the more recent attempts to create autonomous cars, like the one by Google ((Google, 2015b), see Figure 2-18.b), seem to have gained higher public acceptance (Schoettle & Sivak, 2014). This section will outline some implications related to operating an autonomous vehicle, present available work on warnings for the autonomous car driver and conclude with the next research question of this thesis, motivated by the limited work on warnings communicating transitions of control between car and driver.

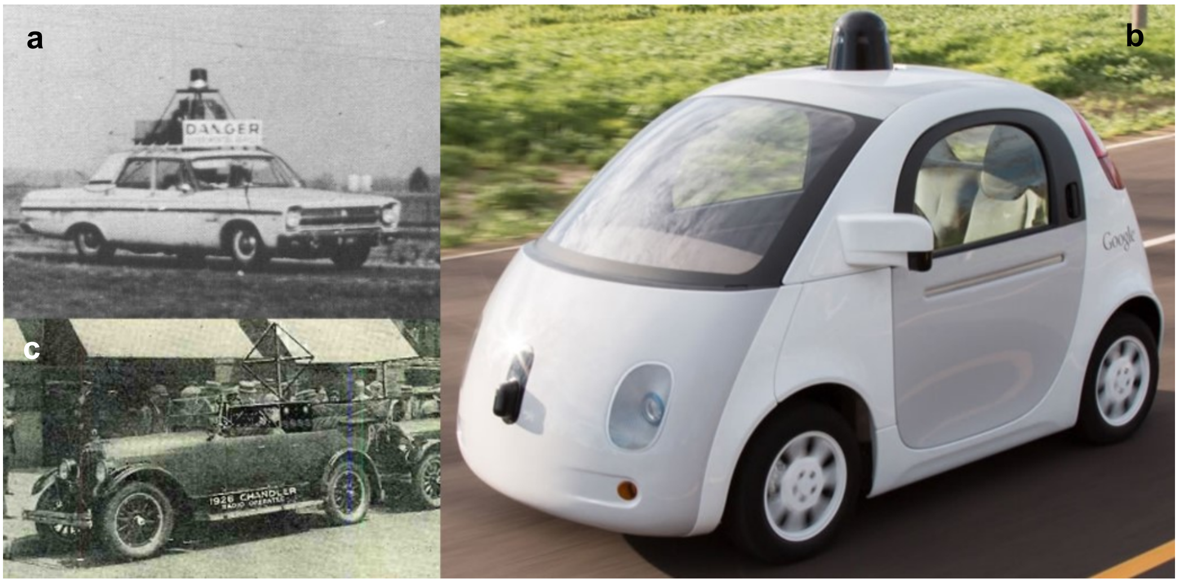


Figure 2-18: Three cars with autonomous capabilities from different eras. (a) Instrumented test vehicle from Ohio State University. Image taken from (Fenton, Cosgriff, Olson, & Blackwell, 1968). (b) The autonomous car developed by Google. Image taken from (Google, 2015b). (c) A car fitted with a Houdina Radio Control device, able to be radio controlled from a following car. Image taken from https://en.wikipedia.org/wiki/Houdina_Radio_Control#/media/File:Linrrican_Wonder.png.

The implications of the new wave of vehicle automation are exciting, which motivates rich discussion. Journalists predict massive changes in transportation over the next decade, including eco-friendly vehicles and improved safety but also potential job losses (Kanter, 2015). The human driver will become less involved in driving, or even prevented from driving altogether (Dredge, 2015). The autonomous car industry is gaining momentum, which is reflected in a plethora of patents related to the autonomous car interface, *e.g.* (Davidsson & Chen, 2014; A. C. Green, Salinger, Tsimhoni, & Raphael, 2013; Prokhorov & Uehara, 2015; Szybalski, Gomez, Nemec, Urmson, & Thrun, 2014). Aside from the Google car, other efforts in bringing autonomy to the road are currently nearing market release (Cruise, 2015). These advances are not without concerns from the public over the safety of this new technology (Kyriakidis et al., 2014), as well as possible security issues that might emerge when using these vehicles (Schoettle & Sivak, 2014). To address such worries, there is careful examination of road accidents involving autonomous vehicles from technology providers (Google, 2015c). This shows the importance of safety while automation is becoming more robust. Car autonomy is a staged rather than binary process, with levels of autonomy increasing as driver involvement decreases (National Highway Traffic Safety Administration, 2013; SAE J3016 & J3016, 2014). See Table 2-1 for the levels of autonomy according to NHTSA (National Highway Traffic Safety Administration, 2013) and their explanation. As highlighted in Level 3 of Table 2-1, when there is still necessity for the human driver to assume control of the vehicle, user interfaces are required that signify this event. The handover, the point of transition of control from the car to the driver, and *vice versa,* is therefore a critical part of this interaction. An effective warning mechanism for such a critical case is essential, as lack of clarity over who has control of the vehicle at a given moment can be catastrophic.

Level of automation	Summary of level characteristics
Level 4 **- Full Self-Driving Automation**	The vehicle is in complete and sole control.
Level 3 **- Limited Self-Driving Automation**	Automated vehicle functions operate in this level in combination. As a result, the driver can have both their hands and feet off the car controls. The driver is not solely responsible for safe operation and constant monitoring is not required. However, the driver is expected to be available for occasional control, signalled by the vehicle, with sufficiently comfortable transition time.
Level 2 **- Combined Function Automation**	Automated vehicle functions operate in this level in combination. As a result, the driver can have both their hands and feet off the car controls. The driver is still responsible for safe operation and constant monitoring is required for resumption of control on short notice.
Level 1 **- Function-Specific Automation**	Automated vehicle functions operate in this level, but not in combination. As a result, the driver can have their hands or feet off the car controls, but never both. The driver is solely responsible for safe operation and constant monitoring is required.
Level 0 **- No-Automation**	The driver is in complete and sole control of the vehicle.

Table 2-1: The levels of automation, as provided by NHTSA (National Highway Traffic Safety Administration, 2013), with the summary of each level's characteristics.

The differences between automated and manual driving have been highlighted in the past. De Winter *et al.* (de Winter et al., 2014) reported an increased occurrence of non-driving related tasks when automation was high. When participants were attentive to the road, situational awareness increased, however when engaged with non-driving tasks situational awareness decreased. Brandenburg & Skottke (Brandenburg & Skottke, 2014) observed more risky driving behaviours, *e.g.* shorter distances to the lead vehicle of a platoon when driving manually after having driven for ~33 *km* in automated mode, displaying a more risky behaviour after exposure to automation. These studies highlight the importance of maintaining the driver's situational awareness, at least when intervention is still expected. Interestingly, this is still a relevant topic in the field of aviation (Geiselman, Johnson, &

Buck, 2013). As discussed in previous sections, the use of warnings and cues provided by the interface is an effective mechanism to achieve this.

Young & Stanton (Young & Stanton, 2007) discussed the situation of an automation failure in a vehicle. They found that reaction times of unwarned participants were vastly higher compared to when the vehicle was manual, and concluded that warnings are required for such a case, since *"drivers using automation will have to be more attentive than ever before"*. A similar result was observed by Mahr *et al.* (Mahr et al., 2010). Merat *et al.* (Merat, Jamson, Lai, Daly, & Carsten, 2014) found that drivers were quicker to resume control from highly automated driving when automation was switched off regularly compared to when automation switch-off was triggered when drivers disengaged their attention from the road. In their study, the current driving mode was indicated through a text display lighting up when in automated driving. However, there was no emphasis on how to inform the driver about the transition from manual to automated modes. Gold *et al.* (C. Gold, Damböck, Lorenz, & Bengler, 2013) also investigated the behaviour of drivers when requested to return to driving due to automation failures. Drivers engaged in a tablet side task were warned through a pure tone and a visual icon that they needed to take over control due to an unexpected event and the required times for such transitions were investigated. There was no comparison of warning modalities in this study. It was found that when warnings were presented closer in time to an imminent handover (5 *sec* before) reactions were quicker but driving behaviour was less accurate compared to when they were presented earlier (7 *sec* before). Gold & Bengler (Christian Gold & Bengler, 2014) extended this discussion, reporting that during a handover of control, both time (how long it takes) as well as quality (driving performance during and after resumption of control) are important issues to be considered. Mok *et al.* (B. Mok et al., 2015; B. K. Mok et al., 2015) studied a similar case, where automation of a simulated car would fail before a steep curve and a voice alert would signify the failure. Drivers were watching a video while the car was in autonomous mode, when they were asked to take over control. They found that a 5 *sec* or 8 *sec* window for signifying this critical event created safer driving behaviour and higher likability ratings for the vehicle as opposed to signifying the event 2 *sec* before the driver would encounter the hazard. Pfromm *et al.* (Pfromm, Khan, Oppelt, Abendroth, & Bruder, 2015) evaluated this scenario on a test track and presented similar handover times for the drivers. The above studies investigate the appropriate timing to alert drivers of a handover, usually due to an automation failure. This is a valid approach, however it only covers one critical aspect of situational urgency. Less critical situations, such as loss of GPS signal or an incoming message are also a viable cause for a handover and should be investigated. Further, the above studies do not investigate the appropriate design of handover warnings.

Investigating the design of handover warnings, Koo *et al.* (Koo et al., 2014) found that speech warnings describing why (*e.g.* "Obstacle ahead") rather than how a car took over from the driver (*e.g.* "The car is braking") were preferred and led to better overall driving performance. Naujoks, Mai & Neukum (Naujoks et al., 2014) investigated a handover from the car to the driver due to an automation failure. A pure tone and a flashing icon on the dashboard created shorter handover times and better driving behaviour when combined compared to the icon alone. Through an online test, Eriksson *et al.* (Eriksson et al., 2015) found that the preferred type of information displayed during a handover was modulated by time constraints and traffic. More specifically, there were differences in the type of information the drivers expected to know (*e.g.* speed, automation status, road condition) depending on how much time was available for them to decide (15, 30 or 120 *sec*) and how busy the road was when resuming control (empty road, moderate traffic or rush-hour traffic). Telpaz et al. (Telpaz, Rhindress, Zelman, & Tsimhoni, 2015) used a haptic seat displaying the position of an approaching vehicle from the back using a mapping with tactile alerts. The use of the seat along with a simple audio and visual indication improved handover times and satisfaction compared to the absence of the seat. Walch *et al*. (Walch, Lange, Baumann, & Weber, 2015) used speech and text to alert drivers about handovers during a sharp curve, when a vehicle was blocking the road or when there was no hazard. In all cases except the sharp curve, 3 *sec* was an adequate time for a safe handover in terms of response times and comfortable in terms of participant responses. These studies show a growing momentum of work on signifying handovers, however, as with the work on automation failures, the warnings used are rarely multimodal and the situations addressed do not vary in urgency.

Before concluding with the research question motivated by the work discussed, a set of investigated secondary tasks in autonomous cars will be presented. When a car is partially or fully autonomous, the absence of a driving task allows the driver to engage in other activities. This expected behaviour can only increase driver distraction and decrease situational awareness. Carsten *et al.* (Carsten, Lai, Barnard, Jamson, & Merat, 2012) investigated the type of activities users engage in when in such vehicles. They found that participants primarily engaged in tasks such as listening to the radio or watching a DVD. Their engagement in secondary tasks increased as automation increased. It was also found that secondary task behaviour during lateral support (automated steering only) was very similar to full automation. Conversely, behaviour during longitudinal support (adaptive cruise control only) was closer to non-automated driving. Llaneras, Salinger & Green (Llaneras, Salinger, & Green, 2013) found similar results when comparing adaptive cruise control with full automation, with the most popular secondary tasks being listening to music and conversing with other passengers. Participants also spent less time looking at the road ahead when automation was high. Neubauer, Matthews & Saxby (Neubauer, Matthews, & Saxby, 2014) found that the engagement with a trivia game or a phone conversation during a drive that involved automated and manual parts reduced fatigue and improved driving metrics when participants had control of the vehicle. However, engagement with a secondary task created longer reaction times to an unexpected event. Miller *et al.* (Miller et al., 2015) also found that media consumption on a mobile device reduced fatigue of drivers and did not slow handovers when invited back to driving shortly before entering an area of increased hazard. Their handover warnings were visual and auditory, coming either from the tablet or the dashboard, but no comparison between locations was made.

The above studies reveal a distinct lack of research on how to design effective warnings for control handovers, particularly when the driver is engaged in a secondary task, a behaviour expected to increase as automation increases. This is a significant investigation, since the transfer of control between car and driver is expected to occur repeatedly in cars that are not fully autonomous, which is the next generation of cars expected to occupy the road. Some tasks that are likely to occur under autonomous mode have been identified, and reduced fatigue has been found when engaged with a subset of these tasks as opposed to merely supervising the autonomous car. However, using multimodal ways to alert drivers of imminent handovers is not explored, while interaction with games is also not frequently discussed, other than in cases where the car is fully autonomous and no intervention is expected (see for example (Krome, Goddard, Greuter, Walz, & Gerlicher, 2015; Terken, Haex, Beursgens, & Szostak, 2013)). Further, resuming control with the help of warnings originating from the area of the gaming interaction as opposed to the car has not been studied. This motivated the next research question of this thesis:

How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?

To address this question, a set of experiments investigating warnings for autonomous car handovers were designed, presented in Chapter 8. A set of language-based messages across all combinations of audio, tactile and visual modalities, taking into account the urgency of the takeover situation was created, according to guidelines developed in earlier studies of this thesis. Multimodal warnings addressing situations where control is handed from car to driver, as well as from driver to car were considered, since this distinction has important implications on warning content, which have not been investigated. Warnings were first evaluated to assess their perceived urgency, annoyance and alerting effectiveness in line with prior work (Experiment 8). The time required and driving behaviour observed when returning to driving from a tablet-based game, for all modalities and urgency levels, was then studied in a simulator, to gain a deeper understanding of effective warning design for autonomous car handovers (Experiment 9). Specifically focusing in the critical situation of an automation failure, a set of urgent multimodal warnings designed for such a failure, requiring immediate driver attention was finally investigated (Experiment 10). Warnings were delivered either from the simulator, which is the most common approach in the literature or from a tablet where the user would play a game as a secondary task that would increase distraction. Different warning designs were used, utilising abstract and language-based cues never before compared in this setting and reaction time, accuracy and driving behaviour when returning to driving were measured.

2.7. Conclusions

This chapter discussed available literature in order to motivate the research questions of this thesis. Section 2.1 presented the theoretical background of why alarms are a viable means of attracting attention, by first discussing why Situational Awareness is essential, since the driver needs to be informed on their environment, especially in a complex dynamic task such as driving. The fact that attentional resources during driving can be limited, and even competing across modalities, was also discussed, by presenting the Multiple Resources theory. This is an important consideration for alarm designers, revealing firstly that alerts will be presented in a demanding context, and secondly that there is opportunity to relieve workload in one modality by presenting information in another. A discussion on the properties of alerts as opposed to non-alerts (or noise) followed, by presenting the Signal Detection Theory. In this way, the importance of saliency of warnings was highlighted. Finally, a discussion on the way an alarm is detected and how it affects the operator of the system and their actions concluded this section, by presenting the Model of Alarm Initiated Activities. In that section, the above topics were brought together to discuss the steps taken when an alarm is presented and when the system operator reacts, and highlight the resulting considerations when designing alarms as a Human Factors strategy. The above discussions answered a primary research question, which motivated this thesis:

RQ-0: Why are driver warnings useful?

Sections 2.2 and 2.3 presented a set of multimodal displays that were used in order to alert drivers. The presentation followed studies using audio, visual and tactile modalities, as well as some of their combinations as alerts (Section 2.2). These studies demonstrated that the use multimodal warnings can be effective in attracting driver attention, however the degree of urgency of the signified event was not reflected in the cues. A set of guidelines on how to design warnings that vary in urgency along these modalities was also presented (Section 2.3). However, it was demonstrated that the warnings of varying urgency were rarely combined multimodally and never exhaustively in all modality combinations. Taken together, prior research revealed two opportunities. Firstly, the combination of all candidate modalities in displays would reveal their relative performance when used as alerts, enabling an exhaustive investigation on audio, visual and tactile cues as driver warnings. Secondly, combining multimodal cues that vary in urgency in all modalities would enable the presentation of guidelines on how to design urgency in richer alarms, to meet the demands of a dynamic set of situations varying in criticality, and study the effects of these two design factors (modality and designed urgency) in a systematic way. The evaluation of these factors will inform driver warning design, providing designers with a repertoire of cues to used, depending on the desired emphasis on how to deliver warnings (modality) and how to deliver warnings that are appropriately critical (urgency). This motivated the research question:

RQ-1: How do multimodal driver displays varying in urgency affect performance? (Experiments 1 and 2 – Chapter 4)

To answer this question, Experiment 1 investigated a set of multimodal warnings varying in urgency in terms of subjective measures, by identifying how urgent and annoying they were perceived as. This investigation with such an extensive set of multimodal warnings was absent from prior studies. The results would inform warning designs, where designed urgency matches perceived urgency, while perceived annoyance is kept at manageable levels. Experiment 2 investigated objective responses to the warnings, evaluating the speed and accuracy of participant responses, when identifying their urgency. This task was chosen, since warnings need to be identified effectively when used as alerts, so as to facilitate appropriate reactions. In Experiments 1 and 2 abstract warnings were used as an initial design of focus, to be later compared with language-based ones, and in richer situations. In combination, results of Experiments 1 and 2 would provide an initial but exhaustive investigation of responses to multimodal abstract warnings varying in urgency, so as to study the effects of combining the cues in all modality and designed urgency variations, and inform driver waring design with a novel investigation of these two factors combined.

Section 2.4 presented a set of studies where driver displays varying in message content were compared. A set of cues using different methods of message representation were used in these studies, however there was no systematic examination of how different message contents can affect responses multimodally and in varying degrees of urgency. This is important, since the ability to examine how the factors addressed in RQ-1 (urgency and modality) would be affected by a third factor, relating to the representation of the signified event (message content) would allow for guidelines on how to design warnings that effectively convey the desired message, using the right representation technique. The techniques of representation chosen to be examined in this thesis were two, namely an abstract and a language-based representation. They were chosen since they have distinctly different properties, one with no semantic association to the signified event and one using a language association. Further, since the factors of urgency and modality would still be in scope, the aforementioned representation would need to be evaluated multimodally and in varying degrees of urgency. Neither of these two factors (modality and urgency) had been studied in combination with message content in literature, making it difficult to draw conclusions on how multimodal displays varying in urgency and message content would affect performance. This would inform driver warning design on how to deliver warnings (modality) that are appropriately critical (urgency) and use the right representation of the message to be signified (message content). This motivated the research question:

RQ-3: How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance? (Experiments 4, 5 – Chapter 6, 6 and 7 – Chapter 7)

To answer this question, Experiments 4 and 5 focused on designing and evaluating a set of multimodal language-based driver warnings of varying urgency, and Experiments 6 and 7 focused on comparing these with abstract warnings. Multimodal language-based warnings of varying urgency were novel, primarily because tactile aspects of speech had rarely been investigated in the past. Therefore, Experiment 4 followed the subjective investigation of

Experiment 1, and studied the perceived urgency, annoyance, and alerting effectiveness of these cues, utilising a set of novel tactile patterns that were synchronous to speech, Speech Tactons. The goal was to create warnings that would be perceived as urgent as they were designed to be, that would not be annoying, and that would be considered effective. Experiment 5 investigated whether Speech Tactons would be effectively identified unimodally, so as to suggest their use or avoidance, without the use of speech. This would be a "stress test" of these novel cues, aiming to determine to what extent they could stand alone as driver warnings. Experiment 6 followed the design of Experiment 2, by using an identification task, but with both abstract and language-based multimodal warnings (designed in Experiment 4). As in Experiment 2, this would allow for guidelines on how effectively these warnings would be recognised by drivers, essential in order to facilitate appropriate reactions. Finally, Experiment 7 would evaluate these warnings in the presence of a critical event, and requiring quick reactions. This task would simulate the reaction to events requiring immediate attention on the road, and the results would inform warning design on the effectiveness of multimodal cues varying in urgency and message content.

The notion of situational urgency was presented in Section 2.5, by describing studies that varied the validity of alarms (true or false), and the situations where these alarms were delivered. However, the studies presented did not address the full spectrum of situations where an alarm and a critical event are involved. That is, a critical event can be accompanied by an alarm, an alarm can be present with no critical event, and a critical event can be present with no alarm. This is a useful investigation, since it would be able to detect the effect of the alarm in the presence or absence of a critical event, as well as the effect of the absence of an alarm. In this way, the utility of alarms, how much of a difference their presence would make, would be discovered. Furthermore, this investigation had not been performed in the presence of multimodal displays varying in urgency and message content. This would inform driver warning design on what difference does the presence of a critical event or an alarm make (situational urgency) when delivering warnings in different sensory channels (modality), that are appropriately critical (urgency). As described earlier, this topic was investigated earlier in this thesis' experimental work than the topic of RQ-3 (presented in the previous paragraph) but was discussed later in the literature review for presentation reasons, since the topic was seen to be more separate to the topics discussed in RQ-1 and RQ-3. The research question that was posed as a result of Section 2.5 was:

RQ-2: How does situational urgency influence responses to multimodal driver displays varying in urgency? (Experiment 3 – Chapter 5)

To answer this question, Experiment 3 investigated reactions to multimodal displays varying in urgency, in three variations of situational urgency (critical event plus alarms, alarms with no critical event, critical event with no alarms). Reactions to the above situations were evaluated, so as to investigate the influence of situational urgency on reactions, and be able to draw guidelines on the utility of the warnings (how much of a difference their presence makes). Further, when warnings would be present, the next step to the investigation performed in Experiment 2 would be taken, by using the warnings tested in that experiment in a more critical context, and evaluating their performance when warning modality and designed urgency are varied. Finally, in the absence of warnings, a baseline performance to a critical event when drivers would not be cued would be provided, to be able to compare with performance in the presence of warnings.

Finally, the context of autonomous cars was presented in Section 2.6, where an analysis of literature on using alerts in that context was provided. It was argued, that signifying multimodally a critical use case due to its safety implications, the handover of control, is not currently addressed in literature. This investigation is essential, since handovers are a major interaction that is expected to occur when cars are not fully self-driving, and it is critical to create effective alerts for this situation. Creating alerts that are multimodal, and vary in urgency and in message content would be valuable for the context of autonomous cars for reasons already described in this section, when discussing manual cars. Further, the non-driving related tasks that drivers are expected to engage in when a car is autonomous, would make the need for saliency in the warnings even more prominent, to enable an effective resumption to the driving task. Therefore, the experimental framework created in Experiments 1 and 4 in terms of subjective measures, and Experiments 2, 3, 6 and 7 in terms of objective measures was used, in order to repeat the investigation of modality, urgency and message content, described above, in the context of autonomous cars. This would inform driver warning design on how to deliver warnings (modality) that are appropriately critical (urgency), and use the right representation of the message to be signified (message content) in the context of autonomous car handovers. This motivated the research question:

RQ-4: How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car? (Experiments 8, 9 and 10 – Chapter 8)

To answer this question, Experiments 8 and 9 focused on designing and evaluating a set of multimodal language-based driver warnings varying in urgency, signifying critical and non-critical handovers of control, in terms of both subjective and objective measures. Specifically, Experiment 8 used the design of Experiments 1 and 4, to assess the perceived urgency, annoyance and alerting effectiveness of a set of language-based multimodal warnings varying in urgency, designed for handovers from the car to the driver and *vice versa*. As in these experiments, the goal was to create handover warnings that would be perceived as urgent as they were designed to be, that would not be annoying, and that would be considered effective. Experiment 9, using a variation of the tasks used in in Experiments 3 and 7, evaluated the reactions to the warnings when returning to driving in the presence of a critical event but only when the warnings were critical. This would allow for higher ecological validity of the task used, requiring different reactions when imminent attention was required (high urgency) and when simply an acknowledgement of the warning was needed (medium and low urgency). This task would simulate events requiring varying degrees of immediacy in reactions, and the results would inform warning design on the effectiveness of multimodal cues varying in urgency in the context of autonomous handovers. Further, the use of a distracting task (a tablet game), when the car was in autonomous mode, would ensure an increased workload of drivers when not performing a driving task, and assess the saliency of the designed cues when drivers would be called back to driving. Finally, Experiment 10, presented a variation of Experiment 9, focusing only on critical handovers during an automation failure, utilising both abstract and language-based cue designs, and enabling the delivery of cues in different locations. This would provide results on how different cue designs would compare to each other in this critical event and when delivered in the location of a distracting interaction (in this case a tablet) versus the location of the road in front. It would thus conclude the investigation on multimodal cues varying in urgency and message content, designed for autonomous car handovers, in the critical context of an automation failure, which would require the most imminent responses in a real driving scenario.

To conclude, this thesis will answer the above research questions, by presenting a set of experiments addressing the factors discussed, starting with Experiments 1 and 2 in Chapter

4, investigating how multimodal driver displays varying in urgency affect performance. Before doing so, to clarify some aspects of the experimental methodology of this thesis, Chapter 3 will discuss the experimental settings, the use of participants and the measures used in the Experiments performed.

3. Methodology

In this chapter, some methodological choices of this thesis will be discussed. Initially, the use of a driving simulator in the experiments performed will be justified, followed by a description of the metrics used to assess participant responses. The chapter will conclude with the methodology followed in the experiments of this thesis, which will be discussed together with the use of experimental participants, being part of the methodology followed.

3.1. Use of a Driving Simulator

This thesis' experimental work uses a driving simulator for the experiments performed. Interpreting the definition given by Kaptein, Theeuwes and Van Der Horst (Kaptein, Theeuwes, & Van Der Horst, 1996), the simulator used is a low-level simulator, consisting of a PC, a monitor and a cab with controls. Performing experimental tasks in such a setting has differences compared to actual driving. This section, as well as discussing limitations due to these differences, will describe how such a simulator can still produce useful experimental results.

Kaptein, Theeuwes and Van Der Horst (Kaptein et al., 1996) identify cost-effectiveness, environmental friendliness, optimal experimental control and safe reproduction of critical driving scenarios in a controlled setting as advantages of using driving simulators. As they mention, critical scenarios occur rarely in real road conditions, while their reproduction on the road may be impossible or unlawful, so there is an opportunity for close investigation of these scenarios in a simulator. However, one needs to be sure that results obtained in simulator experiments are not an effect of the simulator characteristics or a poor experimental design, which would limit their generalizability. For this purpose, Kaptein, Theeuwes and Van Der Horst propose simulator validity as a concept researchers need to be aware of when designing experiments using driving simulators.

According to (Kaptein et al., 1996), validity of a driving simulator to address a specific research question is the degree that the simulator creates the same behaviour as would be demonstrated in reality in similar circumstances. The authors posit that validity alone is not meaningful, unless tied to a specific research question related to a specific task, since not all available real-world information is utilised by people to perform a task in the simulator. As they mention: *"As long as the set of cues important to the aspect of driving that is the subject* *of investigation is available in a simulator, that simulator may be as valid as a field experiment"*. Further, they define absolute and relative validity, and internal and external validity as distinct terms that describe simulator validity.

Absolute validity with regard to a research question is whether the effect of an intervention (or measure, as referred to by Kaptein *et al.*) has an absolute size comparable to the effect in reality. Conversely, relative validity is whether the intervention affects the studied effect in the same direction as in reality. The example used in (Kaptein et al., 1996) to illustrate this, is speed reducing road measures. If, by applying a set of these measures in a simulator, it is found that the order in which the measures reduce speed is similar to reality (*i.e.* measure 1 is the best, measure 2 comes second, *etc.*), then the simulator has relative validity. If additionally, the amount by which they reduce speed is comparable to reality, then the simulator has also absolute validity.

Internal validity is the ability to attribute the effect obtained through an intervention tested with a simulator to the intervention itself and not to the design of the simulator. When a simulator has internal validity for a specific research question, then the environment is controlled for the attributes that may affect the results obtained. External validity refers to whether results obtained with a specific set of participants in a specific environment and in a specific period of time can be generalised to other sets of participants, other environments and other time periods. For example, a simulator that supports a small field of view would have low internal validity for research questions requiring a large field of view to be studied (*e.g.* abrupt turns or peripheral visual cues). Similarly, an experiment investigating responses of well-rested experienced drivers during simulated driving in a rural road would have low external validity when drawing conclusions for fatigued novice drivers in urban roads.

It is noted, that, since validity was not tested in the experiments performed in this thesis, in order to make definite claims on the simulator validity, further on-road studies would be needed, replicating the scenarios used and comparing the studies of this thesis with on-road studies. This would be challenging in terms of resources required, as well as liability issues in the case of critical scenarios, often used in the experiments. However, based on the above review, some arguments can be made on the use of a low-level simulator for this thesis, based on the definitions of validity presented.

It is argued that the simulator used had high relative validity and internal validity for the assessment of participant performance in presence of cues, *i.e.* it is expected that performance of different cues used in experiments would be similarly ranked if used in a real driving scenario. The reason is that, in research questions centred around cue performance, cues were the only aspect of driving varied in the studies. The cue design was of high fidelity and was delivered with high fidelity equipment, comparable to equipment that would be used in a car. Therefore, the intervention in question was simulated in an elaborate way, and it is expected that it captured the relative performance of cues in different modalities and designs. It is not expected that the simulator had high absolute validity or external validity, since it differed significantly to a real driving task, and many real-world parameters were not simulated (*e.g.* road motion, full field of view, car dynamics). Further, the scenarios simulated were simple, using roads of light traffic and low complexity of turns. It is therefore expected, that results might be different for more complex traffic scenarios. Another reason contributing to limitations of external validity, is that participants examined were typically young adults with no difficulties in performing driving tasks (see also Section 3.3). Therefore, results may have been different if a different population was tested.

Having stated the limitations of validity of the simulator used, it worth noting that prior studies have evaluated simulator validity and found close relation with results observed in on-road studies, in terms of hazard perception (*e.g.* (Underwood, Crundall, & Chapman, 2011)) and driving performance (*e.g.* (Mayhew et al., 2011)). These two parameters were closely examined in this thesis' experimental work. Driving performance was also assessed when comparing a low-cost desktop simulator to a higher fidelity mid-level simulator (Lemieux, Stinchcombe, Gagnon, & Bédard, 2014), and similar effects were found, highlighting its utility as a reliable replacement to a more expensive simulator. Finally, the simulator used in this thesis' experiments, has been extensively used in prior experimental studies with success, *e.g.* (D. Brumby & Seyedi, 2012; Zhao, Brumby, Chignell, Salvucci, & Goyal, 2013). Thus, it is argued that it is an appropriate tool for data collection with the research questions investigated, offering high relative validity and internal validity.

3.2. Metrics Used

The metrics used in the experiments of this thesis were in all cases taken from existing literature. Since both subjective and objective measures were examined, the metrics used will be presented below in this order. All metrics are also presented in the experimental chapters where they are used, as a reminder to the reader:

- **Subjective Measures**
  - **Perceived Urgency:** This measure assesses how urgent a cue is perceived by a participant. It is subjective, since the participant can freely state their opinion on the urgency of the cue. It is typically measured on a numeric scale (for example 0 to 100, *e.g.* in (D. C. Marshall et al., 2007)) or on a Likert scale (for example "Not at All" to "Extremely" *e.g.* in (Carryl L Baldwin, 2011)). It is desirable, that perceived urgency of a cue matches its designed urgency, so that it is perceived as critical as it is designed to be. This measure has been also used in studies like (C. L. Baldwin et al., 2012; C. L. Baldwin & Moore, 2002; Judy Edworthy et al., 1991b; Pratt et al., 2012).
  - **Perceived Annoyance:** This measure assesses how annoying a cue is perceived by a participant. It is subjective, since the participant can freely state their opinion on the annoyance of the cue. It is typically measured on a numeric scale (for example 0 to 100, *e.g.* in (D. C. Marshall et al., 2007)) or on a Likert scale (for example "Not at All" to "Extremely" *e.g.* in (Carryl L Baldwin, 2011)). It is desirable, that perceived annoyance of a cue is low, so that it is not disrupting the main driving task. This measure has been also used in studies like (C. L. Baldwin & Moore, 2002; Gonzalez et al., 2012).
  - **Perceived Alerting Effectiveness:** This measure assesses how effective as an alert a cue is perceived by a participant. It is subjective, since the participant can freely state their opinion on the effectiveness of the cue. It is typically measured on a Likert scale (for example "Not at All" to "Extremely" *e.g.* in (Carryl L Baldwin, 2011)). It is desirable, that perceived alerting effectiveness of a cue is high, so that it is considered as an effective alert. This measure has been also used in studies like (C. L. Baldwin & Moore, 2002).

- **Objective Measures**
  - **Response Time:** This measure assesses how quickly participants respond to a cue or an event. It can also be called Reaction Time, although this thesis uses the term Response Time. In this thesis' experiments, it is the time that passed from the onset of an event or cue until the participants' first response. Other studies may measure the time from the end of an event or cue or from some other point in the time of an event or cue, depending on the research question. Response Time is an objective

measure, since the time required is an absolute value measured in a time unit (typically seconds). The type of response can vary depending on the task. In this thesis' experiments, the response type could be stepping on a pedal or pushing a button. If the button was pushed in order to recognise the urgency of a cue, the measure was called Recognition Time, for easier discrimination. Lower Response Time denotes higher performance, at least when quick responses are desirable. Response Time is a very common measure to assess time performance in response to an event or cue, some prior studies that have used it are (Cao, van der Sluis, et al., 2010; Cristy Ho & Spence, 2005; Cristy Ho, Tan, et al., 2005, 2006; Scott & Gray, 2008). It is referred to as Response Time in SAE J2944 standard (*SAE J2944: Operational Definitions of Driving Performance Measures and Statistics*, 2015).

- **Lateral Deviation:** This is a driving performance metric. It is the Root Mean Squared Error of the lateral position of a vehicle for a period of time (the vehicle's centre distance from the centre of the lane it is situated on). Lateral Deviation is an objective measure, since the it is an absolute value measured in a distance unit (usually meters). It is referred to as Standard Deviation of Lane Position (SDLP) in in SAE J2944 standard (*SAE J2944: Operational Definitions of Driving Performance Measures and Statistics*, 2015). Higher Lateral Deviation denotes poorer driving performance. This measure has also been used in studies like (D. P. Brumby, Davies, Janssen, & Grace, 2011; Salvucci, 2006).
- **Steering Angle:** This is a driving performance metric. It is the Root Mean Squared Error of the position of the steering wheel for a period of time (the angular distance of the wheel from its central point). Steering Angle is an objective measure, since the it is an absolute value measured in an angular unit (typically radians). Higher Steering Angle denotes poorer driving performance. This measure has also been used in studies like (Salvucci, 2006).

3.3. Experimental Methodology and Use of Participants

The experimental designs used throughout this thesis were repeated measures within subjects designs (A. Field, 2009). The advantage of these designs is that all participants were exposed to all conditions, while any learning effects were ameliorated by always randomising the experimental conditions. The statistical analysis used was repeated measures Analysis of Variance (ANOVA), with number of factors varying from one to four, depending on the experimental design. This analysis allows for the study of multiple measures (dependent variables), as they are influenced by one or more factors (independent variables), as well as the study of the interaction between these factors, when they are more than one. The analysis described in (A. Field, 2009) was followed, and any corrections due to violations of sphericity or multiple comparisons were applied, mentioned in the respective sections of this thesis. Statistical power of the tests is always reported for completeness.

Due to the experimental design applied, a repeated exposure to warnings was necessary. This strategy can reduce ecological validity and limit the alerting effect of subsequent alert occurrences to the initial alert (see also (Aust, Engström, & Viström, 2013)). The limitations of delivering multiple cues have also been discussed in (Spence & Ho, 2008), where it is highlighted that the responses to such repeated tasks can become more automated, with less resemblance to a naturalistic delivery of cues. Conversely, the use of multiple alerts reduces drivers' willingness to engage in a secondary task, which is a positive effect for manual driving (Kidd, Nelson, & Baldwin, 2010). For this thesis, it was essential to follow this methodology, so as to deliver all the experimental conditions, consisting of multiple modalities, multiple levels of designed urgency and different warning designs (see experimental chapters 4 - 8). This technique has been followed extensively in earlier research, *e.g.* (E. J. Hellier et al., 1993; Cristy Ho et al., 2007; B. A. Lewis & Baldwin, 2012; Scott & Gray, 2008). As presented in the experimental chapters, this experimental design was still able to uncover a variety of effects, and highlight the relative performance of a large number of cue designs.

The use of experimental participants considered the suggestions provided in (A. P. Field & Hole, 2003; Purchase, 2012), with a desired number of participants that would allow for the discovery of effects of varying size, across multiple experimental factors, without merely struggling to achieve statistical significance. The numbers were generally adequate, at least fifteen, and as high as twenty-two in some cases. However, it is acknowledged that, resources permitting, a higher number of participants in all cases would have increased confidence in the generalisability of the presented findings.

As suggested in (A. P. Field & Hole, 2003; Purchase, 2012), incentives were always available to participants, so as to increase their commitment to the experiments. Further, ethical approval was always sought and granted for all experiments (ethics approval number: CSE01193). A limitation also mentioned in (A. P. Field & Hole, 2003; Purchase, 2012), was the use of mostly university students of high technical abilities, which was not possible to avoid in many cases of this thesis, and it limited the generalisability of the results. However, emphasis was put in recruiting participants of different ages and backgrounds where possible, as will be described in the experimental chapters.

Re-use of participants was minimal, with only a small number of participants taking part in more than one experiments, as again reported in the experimental chapters. An exception to this is when experiments were taking place on the same day one after another, and advertised as a single experiment with more than one parts. In these cases, using the same participants was unavoidable, since the experiments were grouped together, and using different participants would make them harder to conduct in the available time.

Finally, the structure all experiments accommodated a short discussion at the end, where participants were able to provide comments on their experience and openly discuss any topics related to the experiment. These discussions aimed in creating a more comfortable environment for participants, allowing them to express any comments or concerns they might have. These comments were not treated as experimental data, since there was no protocol, coding scheme or analysis applied. However, if they presented interest, they are reported as anecdotal evidence inside the experimental chapters, not contributing to the acceptance of rejection of the experimental hypotheses.

4. Investigating Abstract Multimodal Driver Displays

4.1. Introduction

As described in the previous chapter, there have been several studies evaluating the utility of audio, visual and tactile cues to alert drivers, *e.g.* (Erp & Veen, 2001; Cristy Ho et al., 2007; Scott & Gray, 2008). Such studies rarely consider the urgency of the signified event in order to vary the warning design multimodally and signify the appropriate urgency level. However, guidelines on how to achieve this have been developed for unimodal cues (Judy Edworthy et al., 1991b; B. A. Lewis & Baldwin, 2012; D. C. Marshall et al., 2007; Pratt et al., 2012). This allows for an investigation on how a set of multimodal cues designed according to these guidelines will perform when evaluated in a simulated driving task, motivating *RQ-1: How do multimodal driver displays varying in urgency affect performance?* This chapter addresses this question by designing a set of multimodal displays for drivers, in all unimodal, bimodal and trimodal combinations of audio, visual and tactile modalities. The cues were evaluated in terms of subjective and objective measures. Subjective measures included perceived urgency and annoyance, in line with (Judy Edworthy et al., 1991b; Gonzalez et al., 2012; D. C. Marshall et al., 2007), evaluated in Experiment 1. These were used to assess whether the desired effect of urgency was achieved in the participants' opinion, as well as how disruptive the resulting cues were perceived as. Objective measures were recognition time and accuracy, in line with (Cao, van der Sluis, et al., 2010; J. Edworthy, Hellier, et al., 2000), evaluated in Experiment 2. These were used to identify how effectively the cue urgency would be identified, and investigate the utility of the cues initially during a non-critical simulated driving scenario.

Section 4.2 describes the warning design of the cues for Experiments 1 and 2. As an initial investigation, abstract pulses were used in the warnings, to be later compared with language-based cues (Experiments 6 and 7). Section 4.3 describes Experiment 1 and Section 4.4 describes Experiment 2. Since an influence of the number of modalities used in the cues (one, two or three) was found in the results of both experiments, this effect is described in Section 4.5. Finally, Section 4.6 discusses the first two experiments, and Section 4.7 provides conclusions and a set of guidelines for abstract multimodal cues varying in urgency.

4.2. Warning Design

In line with (D. C. Marshall et al., 2007), three levels of urgency were designed to signify situations of different importance. Level High (L_H) signified situations of high urgency, such as an impending collision, Level Medium (L_M) of medium urgency, such as low fuel and Level Low (L_L) of low urgency, such as an incoming message. Contrary to (D. C. Marshall et al., 2007), where a navigational message was used as the scenario of intermediate urgency, low fuel was chosen in this study, since a navigational message would require specific steering actions in a real driving setting, and might have created confusion in responses. A similar message ("Low oil pressure") was used as an intermediate alert in (J.D. Lee, Bricker, & Hoffman, 2008), where a set of in-vehicle messages were prioritised in terms of urgency according to SAE J2395 (SAE, 2002). Audio, visual and tactile modalities, as well as all of their combinations were used for the warnings (Audio (A), Tactile (T), Visual (V), Audio + Tactile (AT), Audio + Visual (AV), Tactile + Visual (TV), Audio + Tactile + Visual (ATV)). This resulted in 21 different warning signals, 7 signals with the above modalities × 3 levels of designed urgency.

The warnings consisted of pure tones, colours or vibrations, delivered as pulses to the participants, with a varying pulse rate depending on the level of urgency. Using such simple parameters allowed the warning design to be as similar as possible across all modalities. In line with (C. L. Baldwin et al., 2012; B. A. Lewis & Baldwin, 2012), pulse rate was varied to signify escalating urgency. Warnings of the same level had common characteristics of pulse rate regardless of modality. There were 8 pulses with 0.1 *sec* single pulse duration and 0.1 *sec* interpulse interval for L_H, 5 pulses with 0.17 *sec* single pulse duration and 0.17 *sec* interpulse interval for L_M and 2 pulses with 0.5 *sec* single pulse duration and 0.5 *sec* interpulse interval for L_L. All warnings lasted 1.5 *sec*.

For auditory warnings, base frequency was also varied, in line with (C. L. Baldwin et al., 2012; Judy Edworthy et al., 1991b; B. A. Lewis & Baldwin, 2012; D. C. Marshall et al., 2007) (1000 *Hz* for L_H, 700 *Hz* for L_M and 400 *Hz* for L_L). For visual warnings colour was also varied, in line with (C. L. Baldwin et al., 2012) (Red for L_H, Orange for L_M and Yellow for L_L[1]). A C2 Tactor from Engineering Acoustics[2] was used for the tactile stimuli, as is

[1] Red was *RGB*(255,0,0), Orange was *RGB*(255,127,0) and Yellow was *RGB*(255,255,0).
[2] http://www.atactech.com/PR_tactors.html

common in studies investigating tactile feedback, *e.g.* (E. E. Hoggan & Brewster, 2006; E. Hoggan, Raisamo, & Brewster, 2009). The frequency of the tactile stimuli was kept constant at 250 *Hz*, which is the nominal centre frequency of the C2 and the frequency at which the skin is most sensitive. Multimodal signals consisted of simultaneous delivery of unimodal ones to achieve a synchronous effect of sound, vibration, visuals or any of their combinations. Stimulus intensity was not varied in any of the modalities, to avoid causing discomfort to the participants, as suggested in studies of both Earcons and Tactons (S. A. Brewster et al., 1993; Gonzalez et al., 2012; E. E. Hoggan & Brewster, 2006; E. Hoggan et al., 2009).

To evaluate the warnings created, two experiments were designed. The goal of Experiment 1 was to acquire participants' subjective ratings of perceived urgency and annoyance when exposed to the warnings, without being given any information about their designed urgency level. In Experiment 2, participants would identify the level of urgency of the same set of warnings with performance measured in terms of recognition time and accuracy.

4.3. Experiment 1: Perceived Urgency and Annoyance of Abstract Cues

4.3.1. Motivation

As discussed in Chapter 2, subjective measures of driver warnings can be an important indicator of how they are perceived, providing insights on their acceptance, alerting effectiveness and eventual use rather than disuse (D. C. Marshall et al., 2007). Without the use of a driving task, there can be initial assessment of the cues' perception, helping understanding of the effectiveness of the chosen design. Experiment 1, being the very first of this thesis, explored this subjective quality of the warnings, aiming to evaluate whether a design based on a variety of guidelines would succeed in conveying the appropriate degree of urgency (thus resulting to effective alerts), without a high degree of annoyance (thus aiding cue acceptance). If the cue designs were successful in this respect, they would provide a safe baseline to create abstract multimodal alerts for the rest of the thesis, by varying a set of signal parameters in a systematic way (see previous section, 4.2).

4.3.2. Design

Experiment 1 investigated the subjective responses of participants in terms of perceived urgency and perceived annoyance. In line with (D. C. Marshall et al., 2007), it was hypothesized that the different levels of urgency designed in the warnings would influence the ratings of urgency and annoyance by the participants. Ratings of urgency and annoyance would also differ depending on the modalities used. To investigate the robustness of the warnings across different contexts, all responses were acquired both in the presence and in the absence of a driving simulator. The expectation was that if the participants became immersed in the context of driving, this would influence their ratings. A 7×3×2 within subjects design was followed for this experiment, with Modality, Level of Designed Urgency (LDU) and Context as the independent variables and Perceived Urgency (PU) and Perceived Annoyance (PA) as the dependent variables. Modality had 7 levels: A, T, V, AT, AV, TV, ATV. LDU had 3 levels: L_H (High Urgency), L_M (Medium Urgency) and L_L (Low Urgency). Context had 2 levels: Driving Simulator and No Driving Simulator.

In line with (Gonzalez et al., 2012; D. C. Marshall et al., 2007), it was hypothesized that LDU would be recognised by participants, since the above guidelines were applied in cue design. Annoyance was expected to follow the ratings of urgency, as in (Gonzalez et al., 2012; D. C. Marshall et al., 2007). Since the audio visual and tactile modalities can create different combinations of cues not investigated before, there were also effects in terms of modality expected. One expectation arose from the differential alerting character of audio and vibration, demonstrated in (Cristy Ho, Tan, et al., 2006). There, audio was effective in capturing attention and facilitated an interpretation of a visual event, while tactile managed to capture attention as well but did not influence visual interpretation. This led to an expectation of different ratings for cues including these modalities. Further, the visual modality has been shown to influence ratings less than the other two (Carryl L. Baldwin & Lewis, 2013), but has not been combined exhaustively with the other modalities to investigate responses. An increased number of modalities was also expected to increase ratings, since it has been shown to affect responses in previous studies, *e.g.* (Cristy Ho et al., 2007). Finally, the addition of a visual presentation of the road was expected to increase workload and has been shown to be a valid subject of investigation in previous work (Carryl L. Baldwin & Lewis, 2013). The expectation was therefore that a richer Context (the presence of the simulator) would also increase ratings of urgency and annoyance. As a result, there were the following hypotheses:

- The ratings of PU will be influenced by Modality (H_{1a}), LDU (H_{1b}) and Context (H_{1c});
  - Specifically, PU was expected to increase in multimodal as opposed to unimodal cues, in higher levels of LDU, and in the presence of the simulator.
- The ratings of PA will be influenced by Modality (H_{2a}), LDU (H_{2b}) and Context (H_{2c});
  - Specifically, PA was expected to increase in multimodal as opposed to unimodal cues, in higher levels of LDU, and in the presence of the simulator.

4.3.3. Participants

Twenty participants (6 female) aged between 19 and 32 years old ($M = 22.4$, $SD = 4.3$) took part in this experiment. They all held a valid driving license and had between 1 and 13 years of driving experience ($M = 3.45$, $SD = 3.31$). They were either undergraduate or postgraduate University students. All participants reported either normal or corrected to normal vision and no injuries around the abdominal area, where vibrations were delivered. One participant reported moderately reduced hearing from one ear, which however did not hinder everyday activities. Therefore, data from this participant were retained.

4.3.4. Equipment

The experiment took place in a dedicated University room, where the participants sat on a padded chair in front of a desk with a 27-inch computer screen (Dell 2709W). The computer was running driving simulator software, depicting a three lane road in a rural area and a front car maintaining a steady speed. The simulator has been previously used in many research studies, for example (D. Brumby & Seyedi, 2012). As in (D. Brumby & Seyedi, 2012), safety cones were placed on either side of the central lane, to reinforce lane keeping. Sound was delivered through a set of headphones (Sennheiser HD 25-1). Vibration was delivered through a C2 Tactor from Engineering Acoustics, attached on an adjustable waist belt. The belt was placed by the participants in the middle of the abdominal area, to simulate a vibrating seat belt, similar to (Scott & Gray, 2008). Visuals were delivered through coloured circles that flashed in the top centre of the screen, sized 400×400 pixels (12×12 *cm*). The circles did not obstruct the front car and were designed to simulate the feedback of a Head-Up Display (HUD). The location chosen for visuals was similar to (Doshi et al., 2009) to give salient visual cues (see Figure 4-1.b). Participants used a mouse to submit their ratings. Figure 4-1.a depicts the setup of Experiment 1, 1.c the waist belt and Tactor and 1.e a screenshot of the simulator with a L_M visual cue being displayed.

4.3.5. Procedure

Participants were welcomed and provided an introduction to the experiment. In order to cover any noise from the Tactor, car sound was played through the headphones throughout the experiment. The car sound was an extract from a recording of a vehicle idling, retrieved from the Internet[3].

Before beginning the ratings, all 21 signals were played once to the participants, always in the following order: A → V → T → AV → AT → TV → ATV for L_H, then the same order for L_M and then for L_L. If needed, sound and vibration were adjusted so as to achieve comfortable levels. No information about the levels of designed urgency was given to the participants. Next, the warning signals were played to the participants in a random order and with a random interval of any integral value between (and including) 8 – 20 *sec*, similarly to (Cristy Ho et al., 2007; Cristy Ho & Spence, 2005; Cristy Ho, Tan, et al., 2005). Each stimulus was played 3 times. This resulted in a total of 63 stimuli played to the participants. After each stimulus was played, participants were asked to rate the perceived urgency and annoyance of the stimulus on a scale of 0 to 100, in line with (D. C. Marshall et al., 2007) (0 for lowest, 100 for highest). This was done by manipulating the value of a slider in a window that appeared on the screen, after each stimulus had finished playing.

---

[3] http://soundfxcenter.com/transport/car/020ff2_Compact_Car_Idle_Sound_Effect.mp3

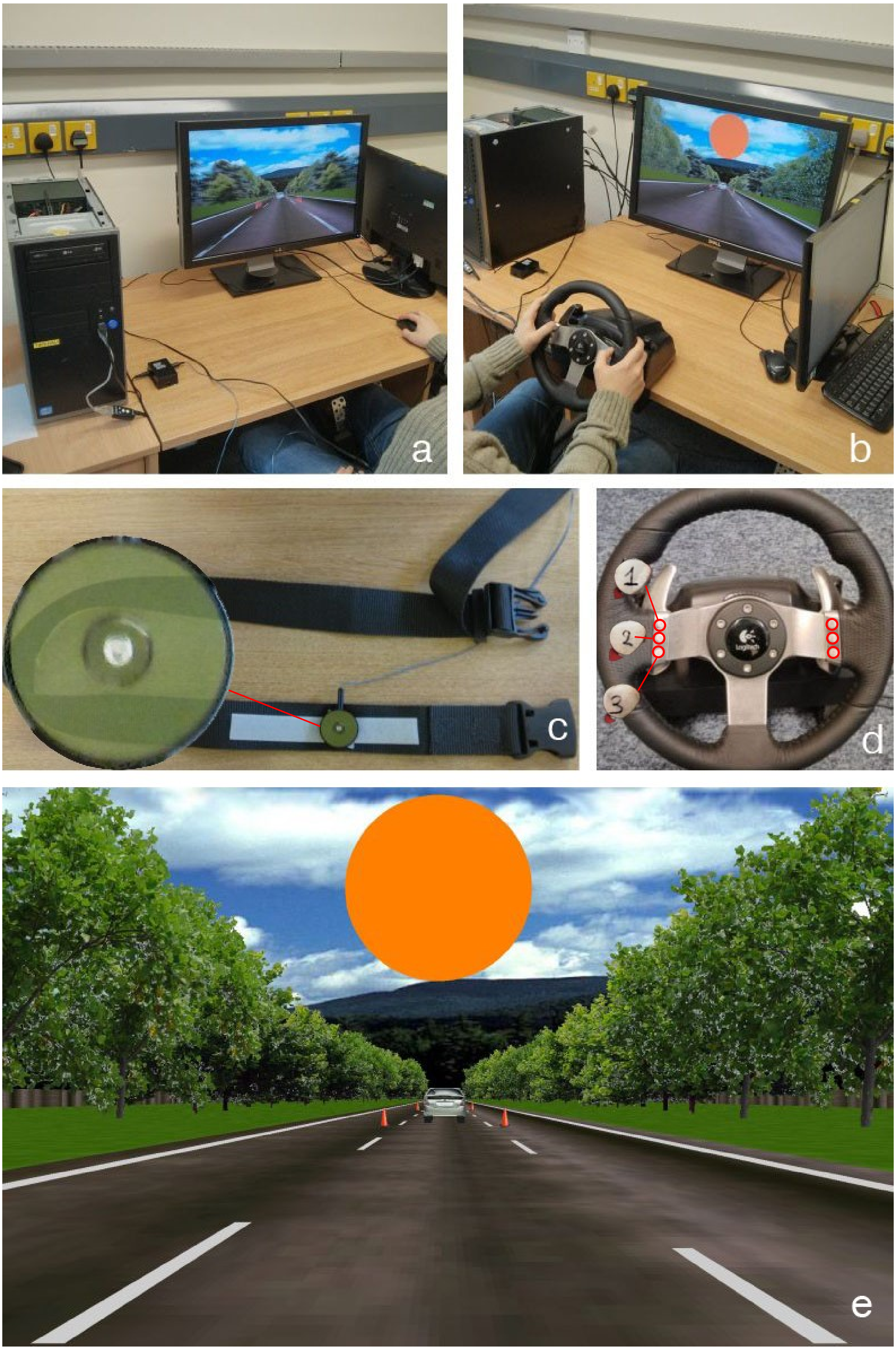


Figure 4-1: The setup of Experiments 1 (a) and 2 (b), the waist belt with the Tactor (c), the steering wheel used in Experiment 2, with the response buttons highlighted (d), and a screenshot of the simulator displaying a L_M visual cue (e).

The above procedure was repeated in two contexts. In the first context, participants rated the stimuli in front of a blank computer screen and in the second context they were looking at the driving simulator with a car that was accelerating and then maintaining a speed of about 60 *mph*. It was chosen not to let participants control the vehicle in this case because pilot tests showed that it was not practical to manipulate the steering wheel and then switch to the mouse rating for this many repetitions. However, they were asked to imagine they were driving the car on the simulator. The above contexts were balanced across participants. After rating the stimuli both in front of the simulator and in front of a blank screen, the experiment was concluded and participants were debriefed about the purpose of the study. The experiment lasted about 45 minutes in total and participants received £6.

4.3.6. Results

4.3.6.1. Perceived Urgency

Data for perceived urgency were analysed using a three-way repeated measures ANOVA, with Context, Modality and Level as factors. Mauchly's test showed that the assumption of sphericity had been violated for Modality and Level, therefore Degrees of Freedom were corrected with Greenhouse–Geisser sphericity estimates. ***Hypothesis H_{1a}:*** The main effect of Modality was found to be significant ($F(3.43,202.07) = 73.64$, $p < 0.001$). Contrasts revealed that ATV warnings were rated as significantly more urgent than TV ones ($F(1,59) = 34.28$, $r = 0.61$, $p < 0.001$), AV as significantly more urgent than AT ($F(1,59) = 31.17$, $r = 0.59$, $p < 0.001$) and V was significantly more urgent than A ($F(1,59) = 16.19$, $r = 0.46$, $p < 0.001$). ***Hypothesis H_{1b}:*** The main effect of LDU was found to be significant ($F(1.26,74.06) = 213.41$, $p < 0.001$). Contrasts revealed that warnings of L_H were rated as significantly more urgent than warnings of L_M ($F(1,59) = 293.88$, $r = 0.91$, $p < 0.001$), which in turn were rated as significantly more urgent than warnings of L_L ($F(1,59) = 92.15$, $r = 0.69$, $p < 0.001$). The mean ratings of perceived urgency across levels can be found in Figure 4-2. ***Hypothesis H_{1c}:*** Finally, there was no significant main effect of Context ($F(1,59) = 2.341$, $p = 0.131$). See Figure 4-3 for the mean ratings of perceived urgency across modalities and Table 4-1 for the pairwise comparisons between modalities for PU.

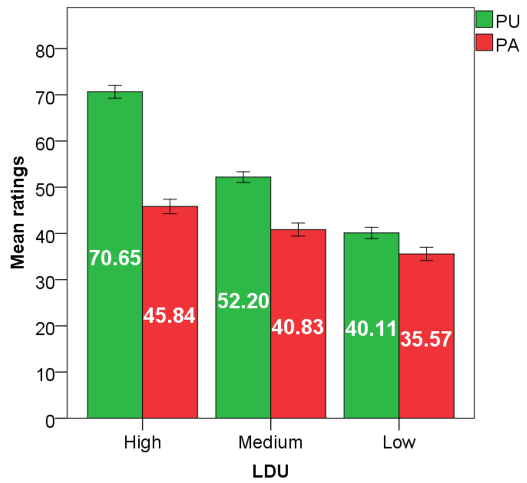


Figure 4-2: The mean ratings of Perceived Urgency (PU) and Perceived Annoyance (PA) across levels for Experiment 1 (hypotheses H_{1b}, H_{2b}). For all graphs, error bars show 95% Confidence Intervals.

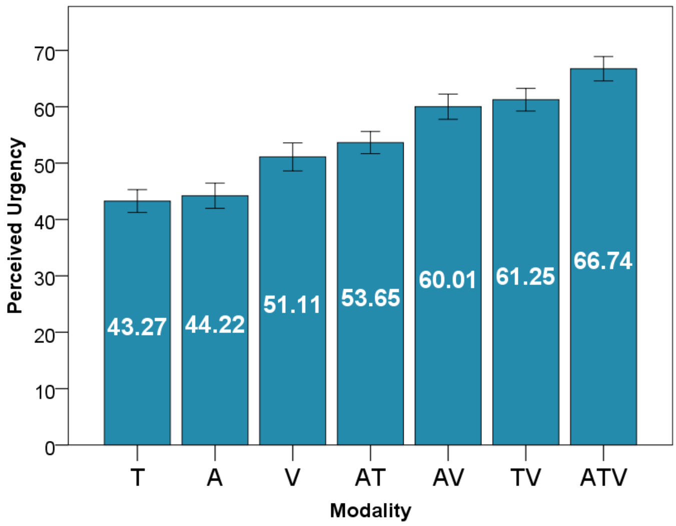


Figure 4-3: The mean ratings of perceived urgency across modalities for Experiment 1 (hypothesis H_{1a}), sorted by their mean values.

	T	A	V	AT	AV	TV	ATV
T		.579	**.000**	**.000**	**.000**	**.000**	**.000**
A	.579		**.000**	**.000**	**.000**	**.000**	**.000**
V	**.000**	**.000**		.092	**.000**	**.000**	**.000**
AT	**.000**	**.000**	.092		**.000**	**.000**	**.000**
AV	**.000**	**.000**	**.000**	**.000**		.124	**.000**
TV	**.000**	**.000**	**.000**	**.000**	.124		**.000**
ATV	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 4-1: Pairwise comparisons between modalities for Perceived Urgency (hypothesis H_{1a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Level and Modality ($F(6.56,387.14) = 16.4$, $p < 0.001$). Contrasts revealed that the observed effect of increased PU was even more pronounced in L_H compared to L_M, for T compared to A ($F(1,59) = 5.08$, $r = 0.28$, $p < 0.05$), A compared to V ($F(1,59) = 17.39$, $r = 0.48$, $p < 0.001$), AT compared to AV ($F(1,59) = 13.58$, $r = 0.19$, $p < 0.001$) and TV compared to ATV ($F(1,59) = 5.31$, $r = 0.29$, $p < 0.05$). Conversely, V was rated higher than AT in L_H as opposed to L_M where this effect was reversed ($F(1,59) = 20.37$, $r = 0.51$, p < 0.001). Comparing L_M with L_L, T was rated contrary to the main effect (higher compared to A) in L_L ($F(1,59) = 14.92$, $r = 0.44$, $p < 0.001$), and the same happened for AT, which was rated higher than AV in L_L ($F(1,59) = 7.93$, $r = 0.34$, $p < 0.01$). Conversely, the main effect was even more pronounced in L_L for AV, which was rated lower than TV ($F(1,59) = 5.94$, $r = 0.30$, $p < 0.05$). Finally, there was a significant interaction effect between Context and Modality ($F(4.51,265.99) = 2.34$, $p < 0.05$), revealing that when the simulator was absent T was rated higher in PU compared to A, and this effect was reversed when the simulator was present ($F(1,59) = 4.78$, $r = 0.27$, $p < 0.05$). Further, AT was rated lower than AV when the simulator was absent and this effect was more pronounced when it was present ($F(1,59) = 7.07$, $r = 0.33$, $p < 0.05$), while AV was rated lower than TV in the absence of the simulator, and this effect was reversed when the simulator was present ($F(1,59) = 5.77$, $r = 0.30$, $p < 0.05$). Therefore, H_{1a} and H_{1b} were accepted, while H_{1c} was rejected. See Figure 4-4 for interactions between LDU and Modality and Figure 4-5 for interactions between Context and Modality.

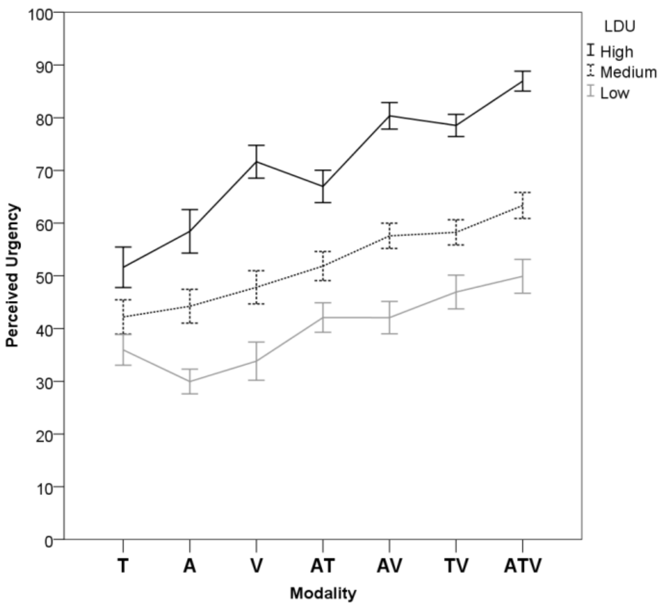


Figure 4-4: The interaction between LDU and Modality for Perceived Urgency in Experiment 1 (H_{1a} – H_{1b}). Modalities are sorted by their mean values of PU.

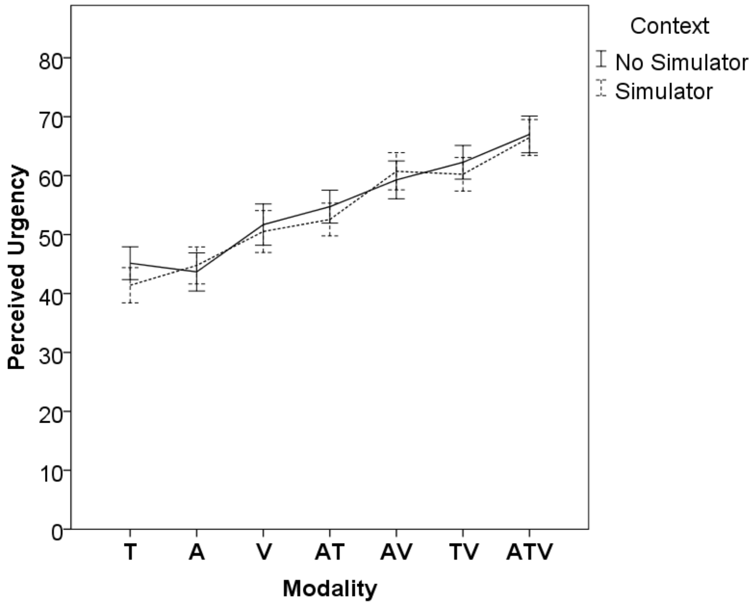


Figure 4-5: The interaction between Context and Modality for Perceived Urgency in Experiment 1 (H_{1a} – H_{1c}). Modalities are sorted by their mean values of PU.

4.3.6.2. Perceived Annoyance

Data for perceived annoyance were also analysed using a three-way repeated measures ANOVA, with Context, Modality and Level as factors. Mauchly's test showed that the assumption of sphericity had been violated for Modality and Level, therefore Degrees of Freedom were corrected with Greenhouse–Geisser sphericity estimates. ***Hypothesis H_{2a}:*** The effect of Modality was found to be significant ($F(2.87,169.37) = 27.54$, $p < 0.001$). Contrasts revealed that ATV warnings were rated as significantly more annoying than AT ones ($F(1,59) = 5.49$, $r = 0.29$, $p < 0.05$), TV warnings as significantly more annoying than T ($F(1,59) = 20.56$, $r = 0.51$, $p < 0.001$), AV warnings as significantly more annoying than A ($F(1,59) = 6.93$, $r = 0.32$, $p < 0.05$) and A warnings as significantly more annoying than V ($F(1,59) = 8.81$, $r = 0.36$, $p < 0.05$). The mean ratings of annoyance across modalities can be found on Figure 4-6 and the pairwise comparisons for PA in Table 4-2. ***Hypothesis H_{2b}:*** The main effect of LDU was found to be significant ($F(1.65,97.56) = 37.76$, $p < 0.001$). Contrasts revealed that warnings of L_H were rated as significantly more annoying than warnings of L_M ($F(1,59) = 16.42$, $r = 0.47$, $p < 0.001$), which in turn were rated as significantly more annoying than warnings of L_L ($F(1,59) = 34.74$, $r = 0.61$, $p < 0.001$). ***Hypothesis H_{2c}:*** Finally, there was no significant main effect of Context ($F(1,59) = 0.84$, $p = 0.36$). The mean ratings of annoyance across levels can be found on Figure 4-2.

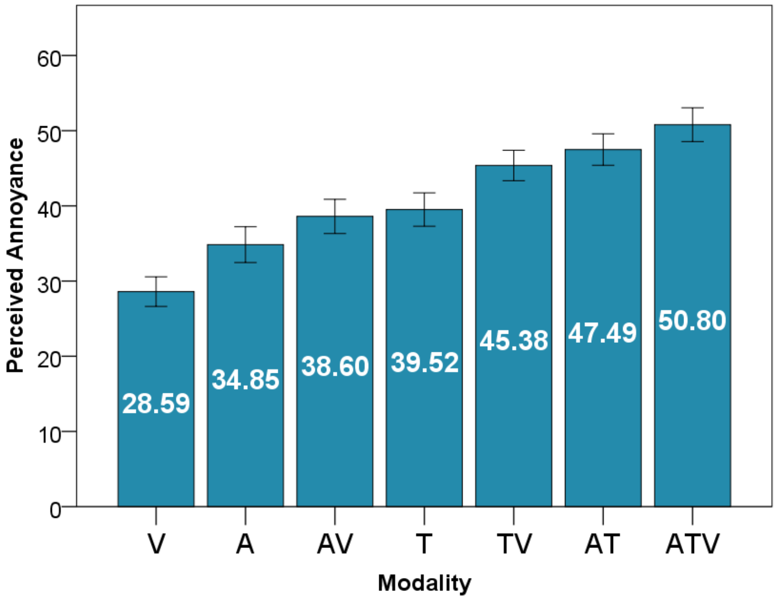


Figure 4-6: The mean ratings of annoyance across modalities for Experiment 1 (hypothesis H_{2a}), sorted by their mean values.

	V	A	AV	T	TV	AT	ATV
V		**.004**	**.000**	**.000**	**.000**	**.000**	**.000**
A	**.004**		**.011**	.105	**.000**	**.000**	**.000**
AV	**.000**	**.011**		.743	**.007**	**.000**	**.000**
T	**.000**	.105	.743		**.000**	**.000**	**.000**
TV	**.000**	**.000**	**.007**	**.000**		.160	**.002**
AT	**.000**	**.000**	**.000**	**.000**	.160		**.023**
ATV	**.000**	**.000**	**.000**	**.000**	**.002**	**.023**	

Table 4-2: Pairwise comparisons between modalities for Perceived Annoyance (hypothesis H_{2a}). The significance (*p)* values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

There was a significant interaction effect between LDU and Modality ($F(8.11,478.29) = 3.74$, $p < 0.001$). Contrasts revealed that the T warnings elicited significantly lower ratings of annoyance compared to the AV ones for L_H, an effect that was reversed for L_M ($F(1,59) = 5.72$, $r = 0.30$, $p < 0.05$). See Figure 4-7 for the interaction between LDU and Modality for PA. There was also a significant interaction between Context and Modality ($F(4.51,265.97) = 5.12$, $p < 0.001$), revealing that when the simulator was present T was rated lower for annoyance than AV, which was reversed when the simulator was absent ($F(1,59) = 6.07$, $r = 0.31$, $p < 0.05$). Therefore, H_{2a} and H_{2b} were accepted, while H_{2c} was rejected. See Figure 4-8 for the interaction between Context and Modality for PA.

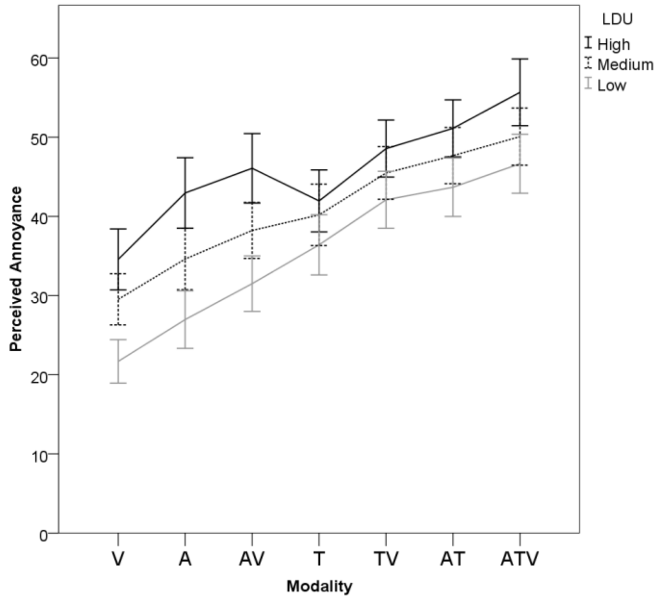


Figure 4-7: The interaction between LDU and Modality for Perceived Annoyance in Experiment 1 (H_{2a} – H_{2b}). Modalities are sorted by their mean values of PA.

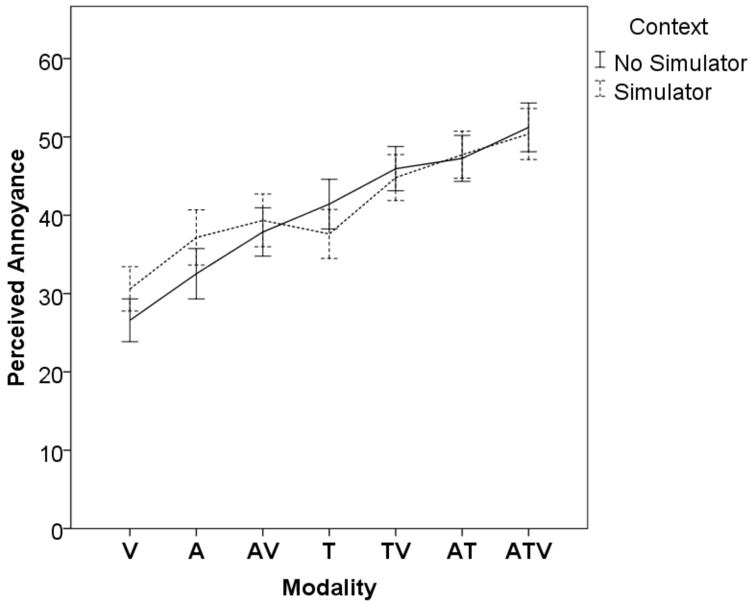


Figure 4-8: The interaction between Context and Modality for Perceived Annoyance in Experiment 1 (H_{2a} – H_{2c}). Modalities are sorted by their mean values of PA.

As evident from Experiment 1, the perceived urgency of the warnings matched their designed urgency, since there was a clear difference of participants' ratings along the three urgency levels. This means that the urgency of the cues designed was clearly identified, even without any training, suggesting that the design was effective. Although perceived annoyance did increase when warnings became more urgent, this effect was not as strong as urgency. In terms of modalities, the urgency ratings increased as more modalities were used, and the ratings of annoyance were higher for signals using the tactile modality. To further investigate the performance of the cues while in a simulated driving task, Experiment 2 was conducted, where speeded recognition of the cues was required by participants.

4.4. Experiment 2: Recognition Time and Accuracy of Abstract Cues

4.4.1. Motivation

Experiment 1 investigated the subjective responses of warnings, looking into perceived urgency and annoyance. Although useful, these results would provide little information regarding the immediacy of responses these warnings could achieve. In a driving scenario, recognition of the cues delivered to the driver is essential, to increase situational awareness and facilitate appropriate responses to the signified situation (Endsley, 1995). This was investigated with Experiment 2, where objective responses to the cues were sought, by introducing a recognition task to participants, where they would be able to identify the level of urgency of the warnings. In this manner, the cues' effectiveness would be assessed in the context of a time based task, simulating realistic conditions, where cue events need to be recognised and acted upon. It should be noted that Experiment 2 did not aim to simulate critical events, where quick responses would be essential, as in Experiments 3, 7, 9 and 10. The experimental goal was rather to focus on identification tasks, which, although not critical, are performed very often while driving.

4.4.2. Design

For Experiment 2, the same warnings as Experiment 1 were evaluated in terms of recognition time and accuracy. In line with (Cao, van der Sluis, et al., 2010; J. Edworthy, Hellier, et al., 2000), it was hypothesized that the designed urgency and modality of warnings would influence their recognition time. As in Experiment 1, all responses were acquired both in the presence and in the absence of a driving simulator. The expectation was that if participants were engaged in a primary driving task, this would influence their responses. A 7×3×2 within subjects design was followed for this experiment, with Modality, LDU and Context as the independent variables and Recognition Time (RecT) and Recognition Accuracy (RecA) as the dependent variables. All participants from Experiment 1, except one, participated in Experiment 2 over the period of a week. This resulted in nineteen participants (6 female) aged between 19 and 32 years (M = 22.52, SD = 4.38). The only difference in equipment between the experiments was that instead of a mouse, participants used a Logitech G27[4] gaming steering wheel to control the simulator and to provide their responses. The simulator logged participants' inputs at a frequency of 50 *Hz*. Figure 4-1.b depicts the setup of the experiment and 1.d the steering wheel.

Similarly to (Cao, van der Sluis, et al., 2010; J. Edworthy, Hellier, et al., 2000), it was expected that the urgency of the cues would be reflected in responses, leading to more rapid reactions for cues of higher urgency. Further, the multimodal character of the cues of this study was further expected to accelerate reactions. Finally, as with Experiment 1, the influence of Context was expected to be observable due to the increase of visual workload in the presence of the simulator. Due to the increased saliency of multimodal cues, and cues delivered with the presence of the simulator, the accuracy of the responses was expected to increase in its presence, while the elevated alertness achieved by highly urgent cues was expected to improve accuracy of responses in higher levels of urgency. As a result, there were the following hypotheses:

- The observed values of RecT will be influenced by Modality (H_{3a}), LDU (H_{3b}) and Context (H_{3c});
    - Specifically, RecT was expected to decrease in multimodal as opposed to unimodal cues, in higher levels of LDU, and in the presence of the simulator.
- The observed values of RecA will be influenced by Modality (H_{4a}), LDU (H_{4b}) and Context (H_{4c}).
    - Specifically, RecA was expected to increase in multimodal as opposed to unimodal cues, in higher levels of LDU, and in the presence of the simulator.

4.4.3. Procedure

Participants were welcomed and provided an introduction to the experiment. As in Experiment 1, car sound was heard through the headphones throughout the experiment to cover any sound from the Tactor. Before beginning the task, a training session was provided, where all 21 signals were played once to the participants. A label with the text "Level 1: Signals of HIGH urgency *e.g.* Impending Collision" was presented and then all signals of

[4] http://support.logitech.com/en_gb/product/g27-racing-wheel

L_H, (A → V → T → AV → AT → TV → ATV). This was followed by a label with the text "Level 2: Signals of MEDIUM urgency *e.g.* Low Fuel" and the signals of L_M in the same order. Finally, a label with the text "Level 3: Signals of LOW urgency *e.g.* Incoming Message" was shown, followed by the signals of L_L. The whole training lasted about 80 *sec* in total. Any adjustments to sound or vibration were also performed at this part to ensure participants were comfortable.

For the main experiment, the warning signals were played to the participants in a random order and with a random interval of any integral value between (and including) 8 – 20 *sec*, as in Experiment 1. Each stimulus was played 3 times. This resulted in a total of 63 stimuli. Participants were asked to identify the level of urgency of each stimulus by pressing one of three buttons on the steering wheel as quickly as possible. Buttons were labelled with numbers (1, 2 or 3) according to the urgency levels – topmost for L_H, middle for L_M, bottom for L_L (see Figure 4-1.d).

The above procedure was repeated in two Contexts, balanced across participants. In the first, participants responded to the stimuli in front of a blank screen and in the second they were steering a car in the simulator, which maintained a speed of about 60 *mph*. Participants were instructed to keep a central position in the lane. The accelerator and brake pedals were not used. Similar to (D. Brumby & Seyedi, 2012), noise was added to the vehicle dynamics so that steering was required to keep the vehicle in the centre of the road and create a realistic driving task. The experiment lasted about 45 minutes and participants received £6.

4.4.4. Results

4.4.4.1. Recognition Time

Data from two participants were discarded because, in the first case, the participant showed obvious signs of fatigue during the experiment and, in the second case, the participant rested the hands on top of the wheel after each response, which led to considerably slower responses. The rest of the data for recognition time were analysed using a three-way repeated measures ANOVA, with Context, Modality and Level as factors. Mauchly's test showed that the assumption of sphericity had been violated for Modality, therefore Degrees of Freedom were corrected with Greenhouse–Geisser sphericity estimates. ***Hypothesis H_{3a}:*** There was a significant main effect of Modality ($F(3.56,170.77) = 71.00$, $p < 0.001$). Contrasts revealed that T warnings elicited significantly slower responses compared to AT warnings ($F(1,48) = 52.60$, $r = 0.72$, $p < 0.001$) and responses to A warnings were significantly slower compared to V warnings ($F(1,48) = 18.71$, $r = 0.53$, $p < 0.001$). The mean response time across levels can be found in Figure 4-9 and across modalities in Figure 4-10, while the pairwise comparisons in RecT between modalities can be found on Table 4-3. ***Hypothesis H_{3b}:*** There was a significant main effect of LDU ($F(1.87,89.65) = 147.65$, $p < 0.001$). Contrasts revealed that warnings at L_H elicited significantly quicker responses than warnings at L_M ($F(1,48) = 284.63$, $r = 0.93$, $p < 0.001$) and at L_L ($F(1,48) = 210.23$, $r = 0.90$, $p < 0.001$). There was no significant difference in recognition times between L_M and L_L ($F(1,48) = 0.554$, $p = 0.46$). ***Hypothesis H_{3c}:*** There was no significant main effect of Context ($F(1,48) = 0.241$, $p = 0.63$).

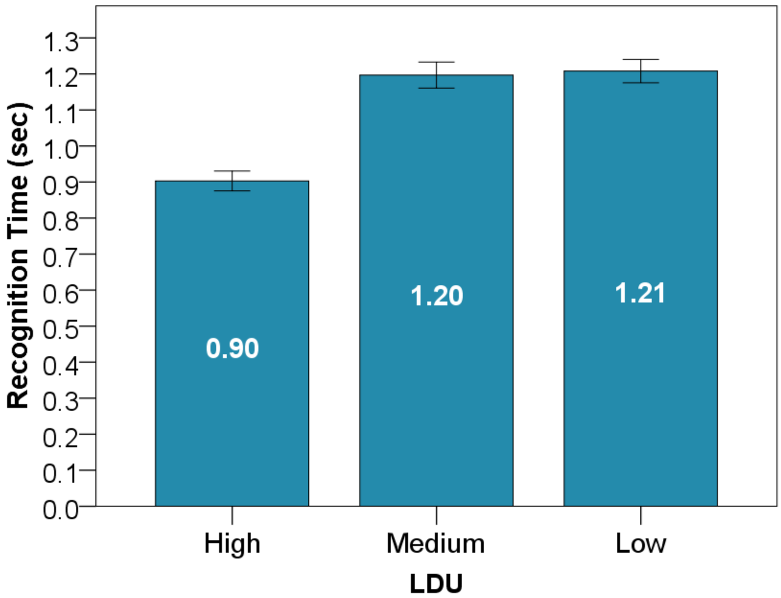


Figure 4-9: The mean recognition times across levels for Experiment 2 (hypothesis H_{3b}).

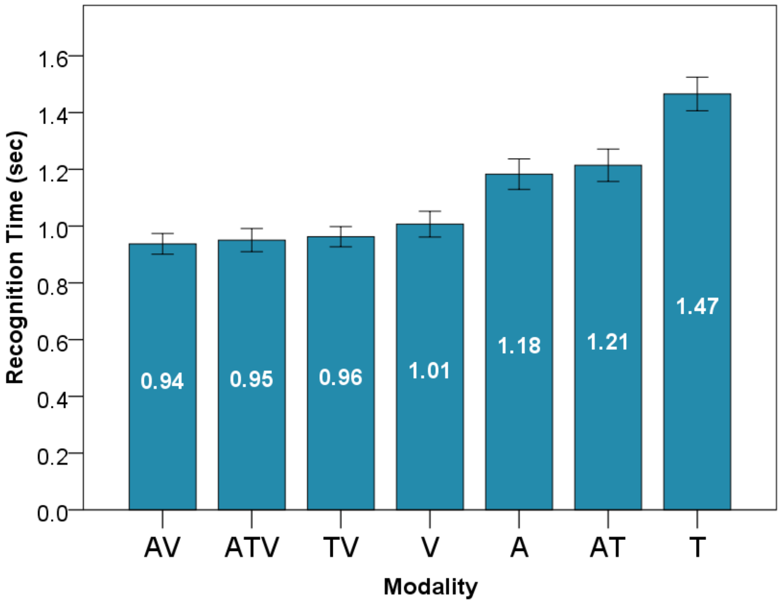


Figure 4-10: The mean recognition times across modalities for Experiment 2 (hypothesis H_{3a}), sorted by their mean values.

	AV	ATV	TV	V	A	AT	T
AV		.903	.273	**.014**	**.000**	**.000**	**.000**
ATV	.903		.225	**.012**	**.000**	**.000**	**.000**
TV	.273	.225		.121	**.000**	**.000**	**.000**
V	**.014**	**.012**	.121		**.000**	**.000**	**.000**
A	**.000**	**.000**	**.000**	**.000**		.112	**.000**
AT	**.000**	**.000**	**.000**	**.000**	.112		**.000**
T	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 4-3: Pairwise comparisons between modalities for Response Time (hypothesis H_{3a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Level and Modality ($F(7.81,374.94) = 3.09$, $p < 0.001$), indicating that the observed differences between V and A were mainly found in L_M ($F(1,48) = 4.68$, $r = 0.30$, $p < 0.05$), where A cues elicited slower responses in compared to L_L ($F(1,48) = 5.61$, $r = 0.30$, $p < 0.05$) and the same happened for AT cues ($F(1,48) = 4.11$, $r = 0.30$, $p < 0.05$). Therefore, H_{3a} and H_{3b} were accepted, while H_{3c} was rejected. See Figure 4-11 for the interaction between Modality and LDU.

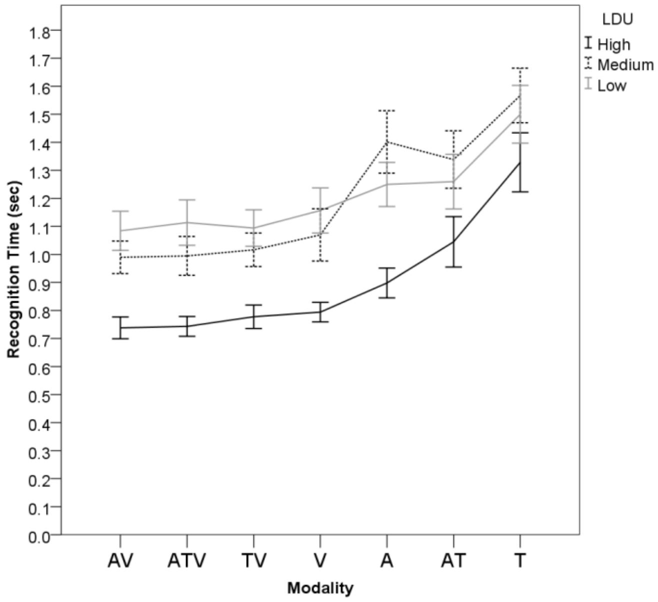


Figure 4-11: The interaction between LDU and Modality for Response Time in Experiment 2 (H_{3a} – H_{3b}). Modalities are sorted by their mean values of RecT.

4.4.4.2. Recognition Accuracy

In all, there were 2390 participant responses and only 4 cases where the participants failed to respond. Nine responses were excluded from the analysis, since the Tactor was audible by the participant. This was due to a misplacement of the Tactor and was reported by the participant during the trial, both when it started and when it stopped being audible. For the rest of the responses, 2255 were valid (94.7%) and 126 incorrect (5.3%). Data for recognition accuracy were treated as dichotomous (with values "correct" or "incorrect") and analysed with Cochran's Q tests. ***Hypothesis H_{4a}:*** Cochran's Q tests revealed that participants made significantly more mistakes in the T modality compared to A ($Q(1) = 17.04$, $p < 0.001$), V ($Q(1) = 8.32$, $p < 0.01$), AT ($Q(1) = 12.76$, $p < 0.001$), AV ($Q(1) = 12.00$, $p < 0.01$), TV ($Q(1) = 7.69$, $p < 0.01$) and ATV ($Q(1) = 7.08$, $p < 0.01$). ***Hypothesis H_{4b}:*** Significantly more misrecognitions were made for L_M compared to L_H ($Q(1) = 16.07$, $p < 0.001$) and significantly more misrecognitions were made for L_L compared to L_M ($Q(1) = 11.78$, $p < 0.001$). ***Hypothesis H_{4c}:*** Finally, it was found that participants made significantly more mistakes when the simulator was present ($Q(1) = 23.02$, $p < 0.001$). Therefore, H_{4a} and H_{4b} were accepted, while H_{4c} was rejected. Table 4-4 shows the RecA across factors for Experiment 2 and Figure 4-12 shows the percentages of correct and incorrect responses per Level of Designed Urgency.

Context		LDU			Modality						
NoSim	Sim	L_H	L_M	L_L	A	T	V	AT	AV	TV	ATV
96%	**92%**	**98%**	**95%**	**90%**	**97%**	**86%**	**95%**	**96%**	**96%**	**94%**	**94%**

Table 4-4: Response Accuracy for Experiment 2 across factors (hypotheses H_{4a}, H_{4b}, H_{4c}).

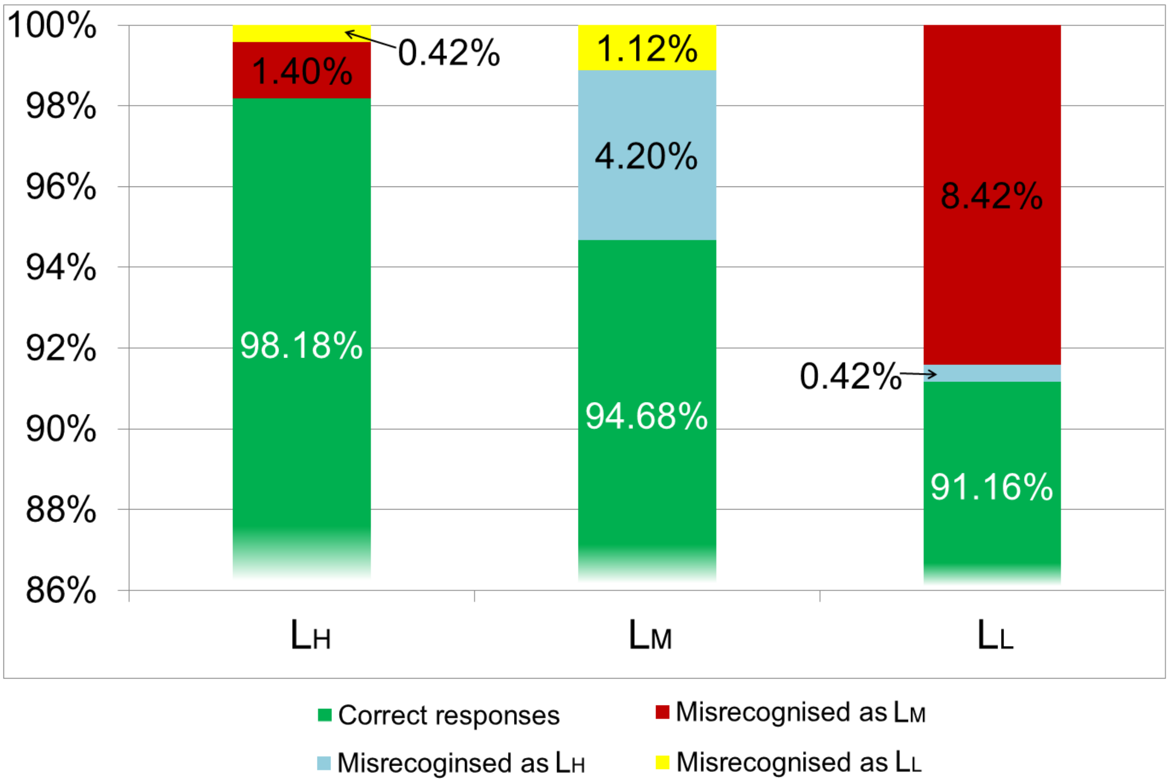


Figure 4-12: The percentages of correct and incorrect responses per Level of Designed Urgency for Experiment 2 (hypothesis H_{4b}).

The results of Experiment 2 suggest that there was a clear advantage for warnings of L_H in terms of recognition time and accuracy, whereas there was no strong difference when comparing L_M and L_L. AV and ATV warnings were the quickest to be recognised, and tactile warnings were the slowest and the least accurate in terms of recognition.

4.5. Number of Modalities

Several participants anecdotally commented that their responses were influenced by how many modalities were present in the signals, namely one modality (A, T, V), two modalities (AT, AV, TV) or three modalities (ATV). To investigate this further, a separate analysis was performed on the number of modalities for both experiments.

Data for Perceived Urgency from Experiment 1 were analysed using a one-way ANOVA, with Number of Modalities (NoM) as factor. It was found that the effect of NoM was significant ($F(2,1077) = 83.65$, $p < 0.001$, $\omega = 0.26$), with planned contrasts revealing that warnings with three modalities elicited significantly higher ratings of urgency compared to warnings with two modalities ($t(717) = 5.31$, $p < 0.001$, $r = 0.19$), which in turn were rated as significantly more urgent than warnings with one modality ($t(709) = 7.67$, $p < 0.001$, $r = 0.28$). A one-way ANOVA was also performed for data of Perceived Annoyance. It was found that the effect of NoM was significant ($F(2,1077) = 50.29$, $p < 0.001$, $\omega = 0.20$), while planned contrasts revealed that warnings with three modalities elicited significantly higher ratings of annoyance compared to warnings with two modalities ($t(1077) = 4.79$, $p < 0.001$, $r = 0.14$), which in turn were rated as significantly more annoying than warnings with one modality ($t(1077) = 5.24$, $p < 0.001$, $r = 0.16$). Figure 4-13 shows the ratings of urgency and annoyance in experiment 1, for different NoM.

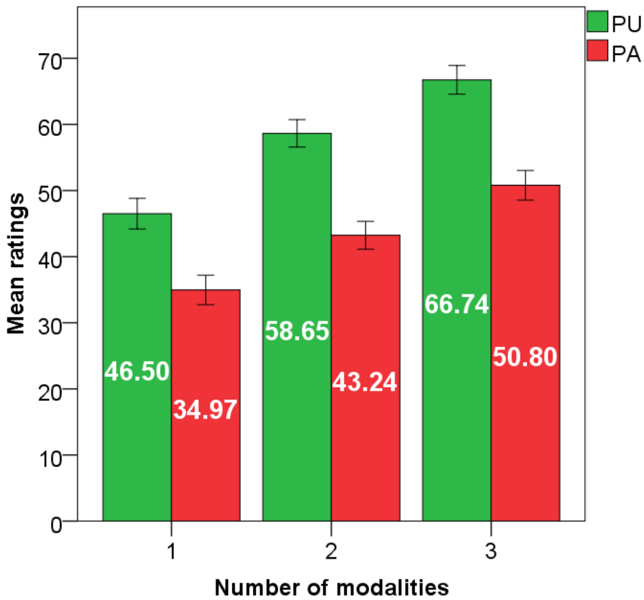


Figure 4-13: The mean ratings of urgency and annoyance for different NoM in Experiment 1.

Data for recognition time from Experiment 2 were analysed using a one-way ANOVA, with NoM as factor. The main effect of NoM was found to be significant ($F(2,1021) = 25.92$, $p < 0.001$, $\omega = 0.22$), with planned contrasts revealing that warnings with three modalities elicited significantly quicker responses compared to warnings with two modalities ($t(663) = -3.09$, $p < 0.001$, $r = 0.12$), which in turn elicited significantly quicker responses than warnings with one modality ($t(680) = -3.97$, $p < 0.001$, $r = 0.15$). Finally, there was no significant difference in terms of recognition accuracy between warnings with one, two or three modalities. Figure 4-14 shows the recognition times in Experiment 2, for different NoM.

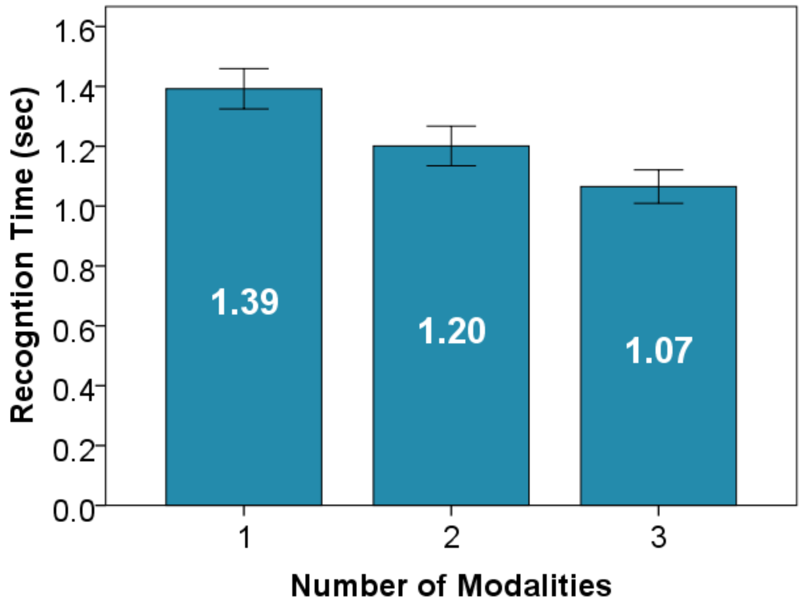


Figure 4-14: The mean recognition times for different NoM in Experiment 2.

4.6. Discussion

As found in both experiments, there is evidence that the warnings succeeded in clearly conveying three different urgency levels to participants, since it was found that ratings of perceived urgency were in accordance with the urgency designed in the warnings (H_{1b} was accepted). This result extends existing work by testing perceived urgency in all combinations of the Audio, Tactile and Visual modalities. This was also done in the context of a driving simulator. As a guideline, it is argued that car warning designers can utilise interpulse interval to vary urgency across any of the above modalities or their combinations. Frequency for audio and colour for visual signals can also be used as a means to manipulate urgency of warnings. Although warnings were rated as more annoying as their designed urgency increased (H_{2b} was accepted), the values and effect sizes of observed differences in perceived annoyance across levels were lower compared to their perceived urgency. This finding is in line with (Gonzalez et al., 2012; D. C. Marshall et al., 2007; Pratt et al., 2012) and indicates that urgency was a more decisive factor than annoyance in the ratings. This is arguably important, since warnings that annoy the driver can be less effective or may even be disabled (see also (Denis McKeown & Isherwood, 2007)). This supports the view that as long as warning designers are cautious not to overload the drivers with alarms of low importance, annoyance can be kept at manageable levels as urgency increases. Chapters 6 and 8 explicitly investigated whether higher annoyance is actually acceptable when cues are critical, by investigating perceived alerting effectiveness of the designed warnings. This can further inform the decision to vary urgency in warnings, potentially creating annoying alerts.

In terms of recognition times, warnings of high urgency were found to be recognised both more quickly and more accurately compared to warnings of medium and low urgency (H_{3b} and H_{4b} were accepted). This result extends existing work like (Cao, van der Sluis, et al., 2010; J. Edworthy, Hellier, et al., 2000) by testing in the context of a driving simulator and using all combinations of Audio, Visual and Tactile modalities. More urgent warnings created quicker and more accurate responses in the present study. These results can suggest that high urgency warnings were effectively associated with a high urgency situation (impending collision) and performed better compared to warnings associated with medium or low urgency situations (low fuel, incoming message). This indicates the robustness of the warnings and their appropriateness for use in a critical driving task.

The modalities used affected the results in two main ways. Firstly, warnings involving visual cues were perceived as more urgent in all signals with equal numbers of modalities (H_{1a} was accepted). Namely, ATV warnings were rated as the most urgent, TV and AV as more urgent than AT, and V as more urgent than A and T. These results were all statistically significant. This can illustrate the strength of the visual modality as a means to design urgent messages. However, it could be argued that in the experiments presented here little visual attention was required for the driving task. Chapter 5 investigates the performance of all of these cues in a more complex simulator driving situation to see if the benefits of the visual cues still hold. Secondly, warnings using the tactile modality ware rated as more annoying (H_{2a} was accepted). This could be partly supported by anecdotal evidence, since some participants mentioned after the end of both experiments that Tactile was often not liked.

There were several interactions found in the results, all with small effect sizes, mainly pointing towards the increased saliency of V and AV cues in L_H (interactions between H_{1a} and H_{1b}, H_{2a} and H_{2b}, H_{3a} and H_{3b}), and the lower perceived urgency and annoyance of unimodal tactile when the simulator was present (interactions between H_{1a} and H_{1c}, H_{2a} and H_{2c}). Further, unimodal T cues did not differ largely in annoyance across levels of designed urgency (interactions between H_{1a} and H_{1b}, H_{2a} and H_{2b}). These effects show that vibration can be well integrated in the driving context without disrupting the main driving task, while maintaining a manageable level of annoyance. However, as more modalities are added to vibration in the cues, annoyance increases, as described above.

Looking at recognition times, findings showed again that signals using the Visual modality lead to quicker responses (H_{3a} was accepted), while signals involving Tactile lead to slower responses (also hypothesized by H_{3a}) and more mistakes (H_{4a} was accepted), a result that was in line with (Cao, van der Sluis, et al., 2010). Taken together, these results can highlight that Visual, used in isolation or combined with Audio or Tactile is a promising modality for conveying urgency both quickly and accurately. Tactile, on the other hand, should be used with caution, as it may create higher annoyance and slower and less accurate responses. Chapter 6 investigates the wrist as a location for tactile feedback and use other features like roughness (see also (E. Hoggan et al., 2009)) to see if this limitation can be ameliorated.

Looking at the interactions for recognition times, again the effects were small, and indicated a poor performance of A and AV cues in L_M compared to L_L (interactions between H_{3a} and H_{3b}). This might have been caused by prolonged decision times over which level to select in this case. It has to be noted however that L_M and L_L do not signify urgent events, therefore such a selection would not be under time pressure in a real driving situation. The increased performance of L_H is reassuring in this case and confirms the effectiveness of the cue design for critical situations.

An interesting effect observed was the influence of NoM in ratings as well as recognition times. Warnings with three modalities were found to be rated as more urgent than warnings with two modalities and the latter more urgent than warnings with one modality. The effect of annoyance was also present as more modalities were used, but with an effect not as strong. Finally, warnings with three modalities caused quicker responses compared to warnings with two modalities, which in turn were quicker compared to warnings with one. Further, there was no difference in terms of recognition accuracy as more modalities were used. In studies like (Erp & Veen, 2001; Cristy Ho et al., 2007; Kern, Marshall, Hornecker, & Rogers, 2009), modality combinations presented better results than the modalities in isolation, but no study until the time Experiments 1 and 2 were conducted had found as clear results in terms of how the number of modalities used affects responses for driver displays. Later studies also found consistent results with this one (Bridget A. Lewis, Penaranda, Roberts, & Baldwin, 2013; van Erp et al., 2015). A clear guideline for warning design is that NoM can be used to convey urgency without sacrificing recognition accuracy. They can also create responses that vary according to the NoM used. Later studies of this thesis investigated whether this effect can be replicated when using richer cue designs in the driving context, including speech, text or tactile roughness.

Additionally, the effect of whether the simulator was present or absent in this study (H_{1c}, H_{2c}, H_{3c}, H_{4c}) was existent but not strong, since the only effect found was that participants made significantly more mistakes in recognizing modalities when the simulator was present. This result could be justified by the higher cognitive load required with the driving task. Although the driving task used was admittedly simple, this result illustrates the robustness of the cues across contexts. Finally, as also mentioned in Chapter 3, it is noted that the results of this study were acquired using a simulated driving task, thus the degree to which they can be generalized to a real driving situation should be investigated (see also Chapter 9.6 on limitations & future work).

4.7. Conclusions and Statement of Findings

In Experiments 1 and 2, the research question *"How do multimodal driver displays varying in urgency affect performance?"* was investigated. All multimodal combinations of Audio, Visual and Tactile modalities were used to alert drivers to events with varying urgency. This contributed to available literature by providing an exhaustive investigation of modality combinations, both in terms of subjective (Experiment 1) as well as objective measures (Experiment 2). It was found that the cues were clearly identified both in terms of perceived urgency in Experiment 1, as well as recognition time in Experiment 2. Perceived annoyance in Experiment 1 was not as high as perceived urgency, indicating the appropriateness of the cues for the driving context. These findings extend available results in all modality combinations used and in the context of a driving simulator. The strength of cues involving visuals in conveying messages quickly and accurately in (Experiment 2), as well as some limitations of utilising tactile cues for warnings (in both experiments) were also highlighted. In addition, more modalities meant quicker and more accurate responses (Experiment 2), as well as higher perceived urgency, without a large increase in perceived annoyance (Experiment 1). The potential of using the number of modalities to convey urgency is a new result for automotive warning design.

Experiments 1 and 2 provided an initial set of results regarding both subjective and objective measures for abstract multimodal warnings, as a baseline knowledge for the rest of the thesis.

The repeated measures experimental design showed potential of successfully highlighting the influence of a number of factors Modality, LDU and Context) to both subjective and objective measures, and was therefore used in the rest of the experiments of this thesis as well. The effective cue design created, conveying appropriate degrees of urgency, was used in all following experiments with abstract cues (Experiments 3, 6, 7 and 10). The mechanism of recognising the urgency level of cues, highlighting the differences between the levels also proved effective, and was used in Experiment 6, where abstract versus language-based cues were investigated in terms of the resulting recognition times. Further, the investigation of Context, whereby the simulator was either present or absent in the tasks performed, showed little difference in observed responses. This increased confidence in the experimental tasks designed, by showing that the effects observed were not an artefact of the presence of the simulator. Since this comparison between the presence and the absence of the simulator was satisfactory, it was not repeated in the rest of this thesis. Finally, the increased annoyance of vibration on the abdomen led to a change of location for the later studies of this thesis, *i.e.* the wrist, to avoid negative perception of the cues (see also Chapter 6). The statement of findings for Experiments 1 and 2 follows:

- Interpulse interval is an effective way to vary urgency in multimodal audio, visual and tactile cues. Varying this feature in all cues resulted to clear distinctions in urgency ratings and recognition time, enabling the design of cues with different levels of urgency;
- Perceived annoyance increases with perceived urgency, since results of subjective measures of urgency and annoyance were increasing as designed urgency was increasing. Therefore, highly urgent cues should not be used unless signifying critical events, order to avoid unnecessary annoyance that could result to alarm disuse;
- More urgent cues lead to quicker reactions when recognising the urgency of warnings in a simulated driving task, since the higher the level of designed urgency the quicker the observed recognition of the cues. This is a positive result, indicating that the cue design used was effective, enabling responses that matched the intended cue urgency;
- Multimodal cues including visuals lead to an increased perception of urgency. This result can be used in warning design, by adding visual elements to cues when urgency needs to be increased.
- Tactile cues on the abdomen and multimodal cues including tactile lead to slower recognition and higher ratings of annoyance, therefore they should be used with caution in non-critical contexts. As described above, creating unnecessary annoyance can have

detrimental effects and lead to disuse, while slow recognition can have negative safety implications when a message needs to be conveyed quickly.

- The number of modalities influences subjective and objective responses to warnings and can be used as a parameter to vary urgency in cue design. Using the number of modalities as a design parameter can manipulate responses to cues, by introducing higher saliency for more critical messages.

The next experiment (Experiment 3) addressed two areas Experiments 1 and 2 did not investigate. Firstly, all cues in Experiments 1 and 2 were investigated in a fixed level of situational urgency, without varying the criticality of road conditions. This was investigated in Experiment 3, where the cues designed in Experiments 1 and 2 were tested both in the presence and absence of a critical event. Secondly, the responses to the cues in Experiments 1 and 2 were acquired using a recognition task, i.e. recognising what the message was, which is a task not typically associated with critical events. Therefore, the response task was modified in Experiment 3, where all responses were under a reaction task, *i.e.* reacting to a critical event, which provided an investigation of the cues in a more critical scenario. In this way, the combination of Experiments 2 and 3 would provide a set of objective measures for abstract cues in varying levels of situational urgency and using tasks of both low (Experiment 2) and high criticality (Experiment 3). Finally, the elevated perceived urgency observed in Experiment 1, and the low recognition time of cues including visuals observed in Experiment 2 would be put to a further test in Experiment 3, by adding more visual workload to the driving task. This was achieved through a visual critical event (a car braking in front), which would increase visual workload and still require quick reactions.

5. Investigating the Influence of Situational Urgency on Abstract Multimodal Driver Displays

5.1. Introduction

As described in Section 2.5, aspects of how the urgency of the driving situation can affect responses to driver displays have been explored. For example, when the brake lights of the lead vehicle are activated, responses can be quicker if there is an actual deceleration of that vehicle (real braking), as opposed to no deceleration ("dummy" braking) (Liebermann et al., 1995). Further, the braking severity of the lead vehicle has been shown to affect reactions (Hulst, 1999; John D. Lee et al., 2004). When there is an alarm without a critical event present (a false alarm) there is further degradation of overall driving performance, since the system's perceived reliability decreases (Lees & Lee, 2007) and unnecessary reactions increase (Maltz & Shinar, 2004). However, there have been no studies using multimodal displays to compare all possible variations of situational urgency, at least for a simple but critical situation, a car in front braking. This event could be signified by alarms, alarms could be absent when this event occurs or finally alarms could be present, but without the occurrence of the event. This motivated the research question: *How does situational urgency influence responses to multimodal driver displays varying in urgency?* This chapter attempts to answer this question with Experiment 3, by evaluating the warnings designed in Experiments 1 and 2 in the above contexts of situational urgency. In this way, the cues design in Experiments 1 and 2, and tested in the absence of a critical task, would be evaluated in a more demanding critical context, and guidelines on their effectiveness when criticality of road condition is increased would be provided.

To assess the influence of warnings on driving, reaction time to the warnings is measured, in line with numerous studies evaluating multimodal displays (*e.g.* (Cao, van der Sluis, et al., 2010; Cristy Ho et al., 2007; Hogema et al., 2009)). Further, lateral deviation and steering angle are used as driving metrics, so as to assess the disruption to the driving task caused by exposure to the cues (as in (D. P. Brumby et al., 2011; Lindgren et al., 2009)). In this way, the influence of modality, designed urgency and situation are assessed in a simulated driving task, and the added value of providing multimodal warnings varying in urgency to signify a critical event is identified. The remainder of this chapter describes the experiment designed to evaluate the influence of situational urgency in responses (Section 5.2). The observed results are then discussed (Section 5.3) and a set of guidelines for using warnings in a context of varying situational urgency are provided (Section 5.4).

5.2. Experiment 3: Evaluating the Effect of Situational Urgency of Abstract Cues

5.2.1. Motivation

As found in Chapter 4, multimodal displays varying in urgency can have advantages both in subjective measures, *i.e.* how people perceive them, as well as objective measures they create, *i.e.* how people respond to them. Experiments 1 and 2 presented a set of results in both types of measures, and showed how varying signal parameters can modulate perceived urgency multimodally and improve cue recognition. However, the task used in Experiments 1 and 2 did not involve the simulation of a critical event, which can have a strong effect in reactions (Hulst, 1999; John D. Lee et al., 2004; Lees & Lee, 2007; Liebermann et al., 1995; Maltz & Shinar, 2004). As mentioned above, such experiments investigating false or unnecessary alarms have not used a systematic variations of multimodal displays. This motivated Experiment 3, where a critical event was simulated in all its variations. As a continuation to Experiments 1 and 2, the abstract multimodal cues designed in the previous were evaluated in Experiment 3, so as to assess their robustness across contexts of situational urgency. This is an important exercise, which can provide clear insights on the effectiveness of the cues, and increase confidence that they are well suited for critical contexts.

The set of multimodal warnings used in Experiments 1 and 2 was also used in Experiment 3, to represent three different levels of urgency and tested in a driving simulator. As in Experiments 1 and 2, the signals were designed across three different urgency levels, according to existing guidelines (Judy Edworthy et al., 1991b; Gonzalez et al., 2012; D. C. Marshall et al., 2007; Pratt et al., 2012). They were tested across three levels of situational urgency: a lead car braking and a warning presented, a lead car braking with no warning presented and just a warning presented. The goal was to investigate the effect of the situation simulated on driver responses, so as to be able to assess the added value of multimodal driver warnings. As described above, several studies have reported how designed urgency and modality affected response times, for example (J. Edworthy, Hellier, et al., 2000; Cristy Ho et al., 2007). The influence of modalities used in warnings on lateral deviation and steering angle has also been shown in the past, for example in (Lindgren et al., 2009). Therefore, it was hypothesized that response times and driving behaviour would be influenced by the modalities used in the warnings, the level of designed urgency of the warnings, as well as the situational urgency of the simulated event.

5.2.2. Warning Design

The set of warnings used in this study were identical to the ones used in Experiments 1 and 2. As in these experiments, three Levels of Designed Urgency (LDU) were created to indicate conditions varying in importance. L_H (Level High) was designed to signify situations of high urgency, such as an impending collision, L_M (Level Medium) situations of medium urgency, such as low fuel and L_L (Level Low) situations of low urgency, such as an incoming message. All unimodal and multimodal combinations of the audio, visual and tactile modalities were used in the warnings: Audio (A), Visual (V), Tactile (T), Audio + Visual (AV), Audio + Tactile (AT), Tactile + Visual (TV), Audio + Tactile + Visual (ATV). The result was 21 different signals: 7 signals with the above modalities × 3 levels of designed urgency.

Pure tones, colours or vibrations were used in the warnings and were delivered repeatedly as pulses to the participants. Depending on the level of urgency, pulse rate varied, increasing as signals became more urgent, as in (C. L. Baldwin et al., 2012; B. A. Lewis & Baldwin, 2012). Independent of modality, warnings of the same urgency level had the same pulse rate. 8 pulses with 0.1 *sec* single pulse duration and 0.1 *sec* interpulse interval were used for L_H, 5 pulses with 0.17 *sec* single pulse duration and 0.17 *sec* interpulse interval for L_M and 2 pulses with 0.5 *sec* single pulse duration and 0.5 *sec* interpulse interval for L_L. All warnings lasted 1.5 *sec* each.

Auditory warnings were additionally varied in base frequency, as suggested in (C. L. Baldwin et al., 2012; Judy Edworthy et al., 1991b; B. A. Lewis & Baldwin, 2012; D. C. Marshall et al., 2007) (1000 *Hz* for L_H, 700 *Hz* for L_M and 400 *Hz* for L_L).Visual warnings were also varied in colour, in line with (C. L. Baldwin et al., 2012) (Red for L_H, Orange for L_M and Yellow for L_L[5]). A C2 Tactor from Engineering Acoustics[6] was used for the tactile

[5] Red was *RGB*(255,0,0), Orange was *RGB*(255,127,0) and Yellow was *RGB*(255,255,0).
[6] http://www.atactech.com/PR_tactors.html

stimuli, a common device in studies of tactile feedback, *e.g.* (E. E. Hoggan & Brewster, 2006; E. Hoggan et al., 2009). Tactile stimuli had a constant frequency of 250 *Hz*, the nominal centre frequency of the C2 - the frequency at which the skin is most sensitive. Stimulus intensity was kept constant in all modalities, to avoid discomfort, a common practice in studies of both Earcons and Tactons (E. E. Hoggan & Brewster, 2006; E. Hoggan et al., 2009). Simultaneous delivery of unimodal signals was used for multimodal ones, to create a synchronous effect of sound, vibration, visuals and their combinations.

5.2.3. Driving Metrics

In addition to measuring the response times of drivers to warnings, lateral deviation and variation of steering angle were also measured to give a complete picture of performance. Lower lateral deviation and variation of the steering angle indicate lower driver distraction (Lindgren et al., 2009; Y. C. Liu, 2001). As in (D. P. Brumby et al., 2011; Salvucci, 2006), the Root Mean Square Error (RMSE) of the vehicle's lateral deviation and steering angle were used as metrics of driver distraction. The effect of presenting multimodal warnings in the presence and the absence of a critical event on these driving metrics has not been investigated in this level of detail in the past.

5.2.4. Experiment Design

A 7×3×3 within subjects design was used for this experiment, with Modality, LDU and Situational Urgency as the Independent Variables. Response Time (ResT), RMSE of Lateral Deviation (LatDev) and RMSE of Steering Angle (SteAng) as the Dependent Variables. Modality had 7 levels: A, T, V, AT, AV, TV, ATV. LDU had 3 levels: L_H (High Urgency), L_M (Medium Urgency) and L_L (Low Urgency). Situational Urgency had 3 levels: Car Braking + No Warning Presented, No Car Braking + Warning Presented and Car Braking + Warning Presented. In line with (Hulst, 1999; John D. Lee et al., 2004; Liebermann et al., 1995), it was expected that when there was an increased urgency of the situation, reactions would be quicker. Further, during false alarms (warnings without a braking event), performance was expected to be poorer, in line with (Lees & Lee, 2007; Maltz & Shinar, 2004). Finally, lane keeping behaviour was expected to be affected by the warnings, since they would pose an additional load to the driving task, as observed in studies like (D. P. Brumby et al., 2011; Salvucci, 2006). As a result, there were the following hypotheses:

- The observed values of ResT will be affected by the Situational Urgency simulated (H_{1a}), the LDU of the warnings (H_{1b}) and the Modality of the warnings (H_{1c}).
    - Specifically, ResT was expected to decrease when Situational Urgency was higher (a critical event was present along with cues), in higher levels of LDU, and when multimodal as opposed to unimodal cues were used as warnings.
- The observed values of LatDev will be affected by the Situational Urgency simulated (H_{2a}), the LDU of the warnings (H_{2b}) and the Modality of the warnings (H_{2c}).
    - Specifically, LatDev was expected to increase when Situational Urgency was higher (a critical event was present along with cues), in higher levels of LDU, and when multimodal as opposed to unimodal cues were used as warnings.
- The observed values of SteAng will be affected by the Situational Urgency simulated (H_{3a}), the LDU of the warnings (H_{3b}) and the Modality of the warnings (H_{3c}).
    - Specifically, SteAng was expected to increase when Situational Urgency was higher (a critical event was present along with cues), in higher levels of LDU, and when multimodal as opposed to unimodal cues were used as warnings.

5.2.5. Participants

Fifteen participants (10 female) aged between 19 and 28 years ($M = 22.67$, $SD = 2.66$) took part. They had not participated in the previous experiments. They all held a valid driving license and had between 1.5 and 8 years of driving experience ($M = 4.5$, $SD = 2.02$). There were 14 university students and one private employee. They reported normal or corrected to normal vision and hearing and no injuries around the abdominal area where vibrations were delivered.

5.2.6. Equipment

The experiment took place in a usability lab, where participants sat on a chair in front of a desk with a 27-inch Dell 2709W monitor and a PC running the driving simulator software, also used in Experiments 1 and 2. In the software, a three lane road in a rural area with a lead car was depicted, maintaining a steady speed (see Figure 5-1 for the setup of the experiment and Figure 5-2.a for a screenshot of the simulator). This simulator has been used in several previous studies, *e.g.* (D. P. Brumby et al., 2011). As in (D. P. Brumby et al., 2011), safety cones were placed on either side of the central lane, to reinforce lane keeping. Participants used the Logitech G27[7] gaming wheel to steer the simulated vehicle and brake. Participants' inputs were logged with a frequency of 50 *Hz*. Sound was delivered through a set of Sennheiser HD 25-1 headphones. Tactile cues were delivered through a C2 Tactor attached to an adjustable waist belt. The belt was placed by the participants in the middle of the abdominal area and was designed to simulate a vibrating seat belt, similar to (Scott & Gray, 2008). Visual cues were delivered through coloured circles that flashed in the top central area of the screen, and were sized 400×400 pixels (12×12 *cm*). The circles did not obstruct the lead car and were designed to simulate the feedback of a HUD. Figure 5-1 depicts the setup of the experiment, Figure 5-2.a a screen from the simulator with the car braking and a visual signal presented and Figure 5-2.b the waist belt and Tactor.

Figure 5-1: The setup of the experiment. Headphones, tactile belt and computer screen were used to deliver the multimodal signals. The second screen, mouse and keyboard were used by the experimenter to control the driving simulator.

[7] http://support.logitech.com/en_gb/product/g27-racing-wheel

5.2.7. Procedure

Participants were welcomed and provided with a brief introduction. To cover any noise from the Tactor, car sound was heard through the headphones throughout the experiment. The car sound was an extract from a recording of a vehicle idling, retrieved from the Internet.

Before beginning the procedure, all 21 signals were played once to the participants, always in the following order: A → V → T → AV → AT → TV → ATV for L_H, then the same order for L_M and then for L_L. Where necessary, sound and vibration were slightly adjusted to maintain comfortable intensities. No specific information about the levels of designed urgency was given to the participants. The only information provided was that the signals presented were not always designed to convey the same level of urgency. Next, participants were asked to drive with the simulator for 90 *sec*, to get accustomed to the experimental setup.

In the main part of the experiment, participants were presented with a driving scene, where they drove a simulated vehicle along a straight rural road and followed a car in front. Participants were able to steer the vehicle and brake, but did not use the accelerator pedal, since the vehicle maintained a constant speed of about 80 *mph*. This speed was chosen so as to exceed the UK motorway speed limit (70 *mph*) and create a hazardous driving situation requiring the drivers' attention. The participants encountered three possible situations during one session. The first involved the front car braking and a warning presented at the same time (Car & Stimulus: CarStim). The second situation involved only the Car braking (Car) and the third only the warning presented (Stim). There were 21 trials for each of the CarStim, Car and Stim conditions (one for each type of multimodal warning). This resulted in 63 trials, which happened in a random order and were separated by a random time interval of any integral value of 8 – 20 *sec*. These values were chosen to be similar to previous studies investigating a repeated occurrence of critical events in the driving task, such as (Cristy Ho et al., 2007; Cristy Ho & Spence, 2005; Cristy Ho, Tan, et al., 2005) and gave the driver time to settle back into driving before receiving another warning.

Participants were asked to maintain a central position in the lane and press the brake whenever they saw the front car braking, or felt a stimulus presented or both of the above. Their ResT was calculated from the onset of the stimulus and / or the start of the braking event of the lead car, until the participant first pressed the brake pedal. Their LatDev and

SteAng were logged from 4 seconds to 1 second before any situation arose, forming their baseline value for driving performance. They were logged again for 3 seconds immediately after the event to assess the effects on driving. For both LatDev and SteAng, the RMSE values were then computed from the logged values. As a result, for each of the 63 trials of one condition, there was one value for each participant's ResT, two values for their LatDev (baseline value and value after the situation arose) and two values for their SteAng (baseline value and value after the situation arose). Each participant repeated the above procedure twice during the course of a week. After the second session the experiment was concluded and participants were debriefed. The experiment lasted about 120 minutes (60 min per session) and participants received payment of £12.

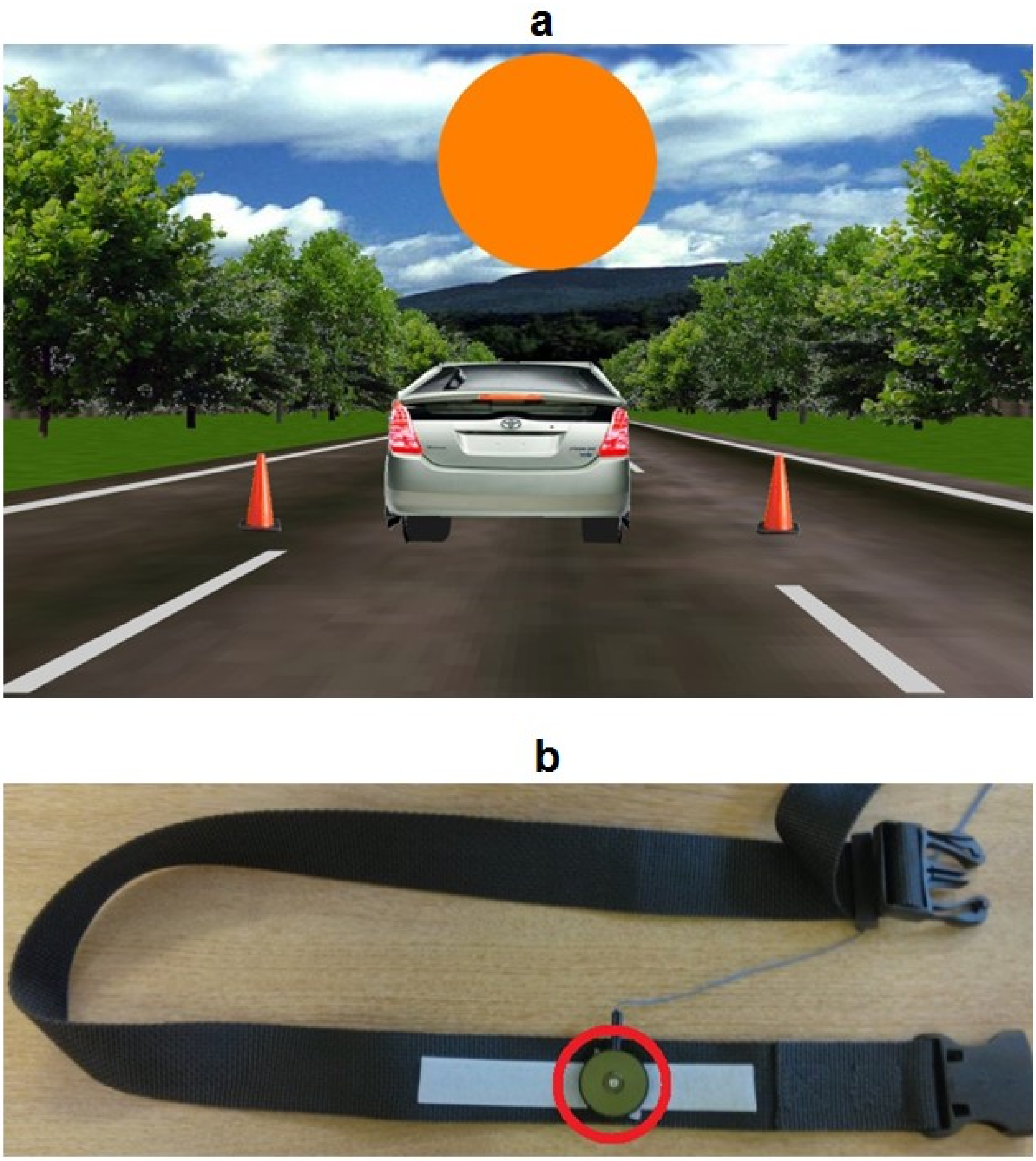


Figure 5-2: (a) A screen from the simulator software, depicting the front car braking and a visual stimulus of medium LDU presented. (b) The waist belt used to provide tactile stimuli, Tactor is highlighted.

5.2.8. Results

5.2.8.1. Response Time

Data for response time were first analysed using a one-way ANOVA with Situational Urgency as a factor. ***Hypothesis H_{1a}:*** There was a significant effect of situation on ResT ($F(2,1883) = 48.56$, $p < 0.001$, $\omega = 0.20$). Planned contrasts revealed that situation CarStim induced significantly shorter ResT compared to situations Car and Stim ($t(1569) = 10.73$, $p < 0.001$, $r = 0.26$), while situations Car and Stim did not differ. As a result H_{1a} was accepted. See Figure 5-3 for the mean response times across situations.

Data for situations Stim and CarStim, where there was a signal present, were analysed using a three-way repeated measures ANOVA, with Situational Urgency, LDU and Modality as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Modality, therefore Degrees of Freedom for Modality were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{1a}:*** There was a significant main effect of Situational Urgency ($F(1,27) = 59.34$, $p < 0.001$). Contrasts revealed, as expected, that situation CarStim induced quicker responses compared to Stim ($F(1,27) = 59.34$, $r = 0.83$, $p < 0.001$).

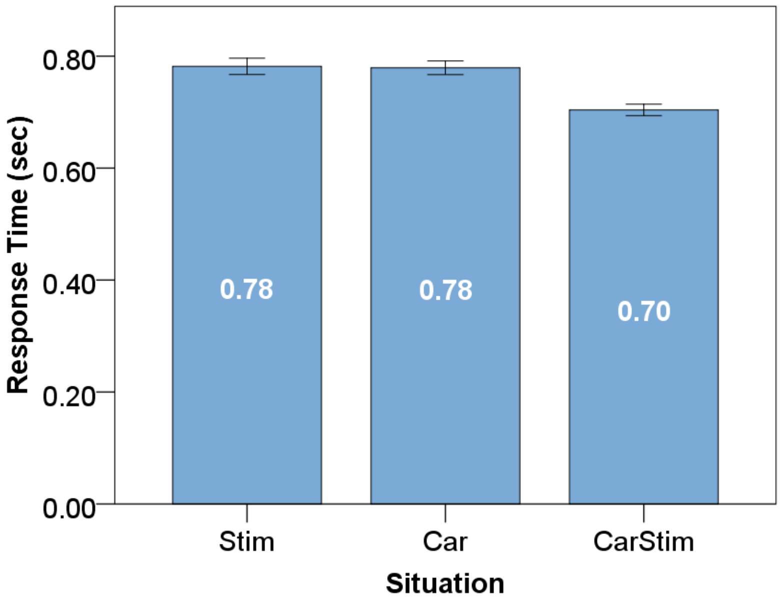


Figure 5-3: The response times across situations (hypothesis H_{1a}). For all graphs, error bars represent 95% Confidence Interval.

Hypothesis H_{1b}: There was a significant main effect of LDU ($F(2,54) = 12.88$, $p < 0.001$). Contrasts revealed that warnings of L_H induced significantly quicker reactions compared to L_M ($F(1,27) = 10.33$, $r = 0.53$, $p < 0.01$), while the difference between levels medium and low did not reach significance ($F(1,27) = 3.87$, $p = 0.06$). Thus H_{1b} was accepted. See Figure 5-4 for the mean response times across levels of designed urgency.

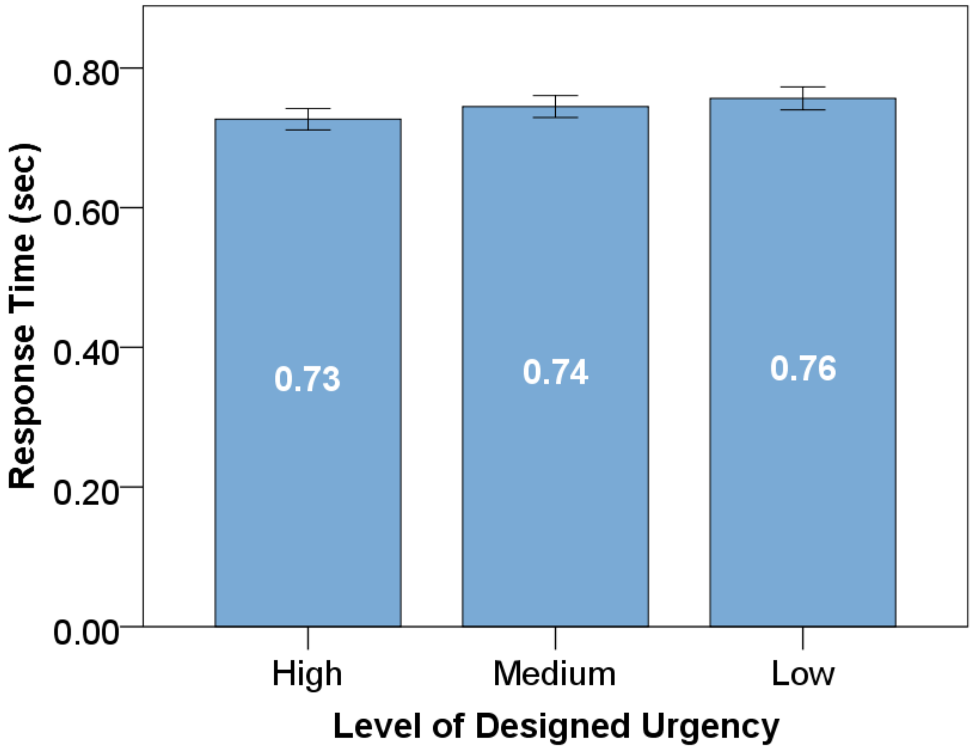


Figure 5-4: The response times across levels of designed urgency (hypothesis H_{1b}).

Hypothesis H_{1c}: There was also a significant main effect of Modality ($F(4.12,111.16) = 23.39$, $p < 0.001$). Contrasts revealed that warnings of the AV, ATV, AT and TV modality all created quicker responses compared to A, V and T warnings ($F(1,27) = 28.18$, $r = 0.71$, $p < 0.001$). As a result H_{1c} was accepted. See Figure 5-5 for the mean response times across modalities and Table 5-1 for the pairwise comparisons in ResT between modalities.

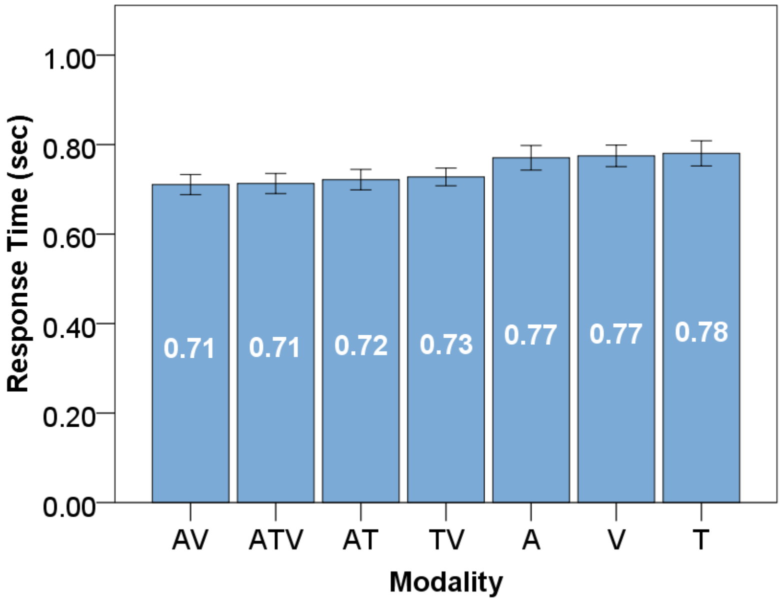


Figure 5-5: The response times across modalities, sorted by their mean values (hypothesis H_{1c}).

	AV	ATV	AT	TV	A	V	T
AV		.743	.063	**.007**	**.000**	**.000**	**.000**
ATV	.743		.278	.050	**.000**	**.000**	**.000**
AT	.063	.278		.587	**.000**	**.000**	**.000**
TV	**.007**	.050	.587		**.000**	**.000**	**.000**
A	**.000**	**.000**	**.000**	**.000**		.863	.345
V	**.000**	**.000**	**.000**	**.000**	.863		.295
T	**.000**	**.000**	**.000**	**.000**	.345	.295	

Table 5-1: Pairwise comparisons between modalities for Response Time (hypothesis H_{1c}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Situation and Modality for situations Stim and CarStim, where there were modalities present ($F(4.83,130.27) = 22.48$, $p < 0.001$). Contrasts revealed that while in situation Stim, ATV warnings created significantly quicker responses than AT ones, this effect was reversed for situation CarStim ($F(1,27) = 9.04$, $r = 0.50$, $p < 0.05$). Further, AT warnings created significantly slower responses compared to TV ones in situation Stim, but this effect was again reversed in situation CarStim ($F(1,27) = 7.43$, $r = 0.46$, $p < 0.05$). Finally, A warnings had significantly slower response times than V in situation Stim, but this effect was reversed in situation CarStim ($F(1,27) = 32.03$, $r = 0.74$, $p < 0.001$). See Figure 5-6 for the interaction between Situation and Modality. There was a significant interaction between LDU and Modality ($F(7.41,200.03) = 2.52$, $p < 0.05$), indicating that T cues elicited quicker responses compared to V ones for L_L, which was reversed in L_M ($F(1,27) = 8.98$, $r = 0.50$, $p < 0.05$).

Finally, there was a significant interaction between Situation, LDU and Modality ($F(6.58,177.63) = 2.11$, $p < 0.05$), indicating that the above described interaction between LDU and Modality was observed in situation CarStim ($F(1,27) = 5.62$, $r = 0.42$, $p < 0.05$).

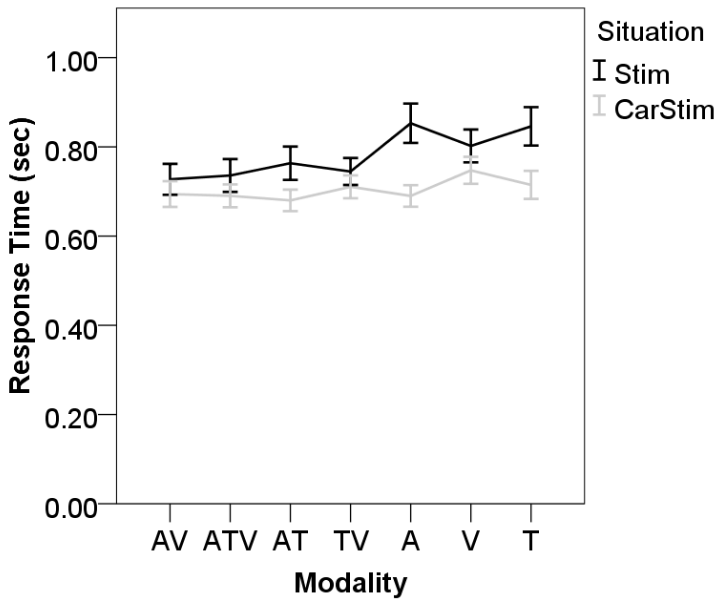


Figure 5-6: The interaction between Situation and Modality (H_{1a} – H_{1c}) with modalities sorted by their mean ResT values.

These results indicate that Situational Urgency, LDU and Modality all influenced driver responses. They also show that warnings including visuals did not create as quick responses in situation CarStim.

5.2.8.2. Lateral Deviation

Data for LatDev were first analysed using a two-way repeated measures ANOVA, with Situation and Time as factors. Situation had three levels: Stim, Car and CarStim. Time had two levels: Before Situation (baseline data) and After Situation (data after the situation arose). Mauchly's test revealed that the assumption of sphericity had been violated for Situation, therefore Degrees of Freedom for Situation were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{2a}:*** There was a significant main effect of Situation ($F(1.93,1215.88) = 59.17$, $p < 0.001$). Contrasts revealed that situation CarStim induced higher values of LatDev compared to Car ($F(1,629) = 81.04$, $r = 0.34$, $p < 0.001$), while values of LatDev did not differ among situations Stim and Car ($F(1,629) = 0.56$, $p = 0.81$). There was a significant main effect of Time ($F(1,629) = 258.22$, $p < 0.001$). Contrasts revealed that LatDev was significantly lower after any situation arose compared to before ($F(1,629) = 258.22$, $r = 0.54$, $p < 0.001$).

There was a significant interaction between Situation and Time ($F(1.97,1240.66) = 63.74$, $p < 0.001$). Contrasts revealed that while in situation Car values of LatDev were lower after the event compared to before it, but there was no such difference for situation CarStim ($F(1,629) = 103.52$, $r = 0.38$, $p < 0.001$). As a result H_{2a} was accepted. See Figure 5-7 for the interaction between Situation and Time for LatDev values. ***Hypotheses H_{2b} and H_{2c}:*** A separate four-way ANOVA test for situations Stim and CarStim (where warnings were present) with Situation, Time, LDU and Modality as factors showed no significant results of LDU and Modality, so H_{2b} and H_{2c} were rejected.

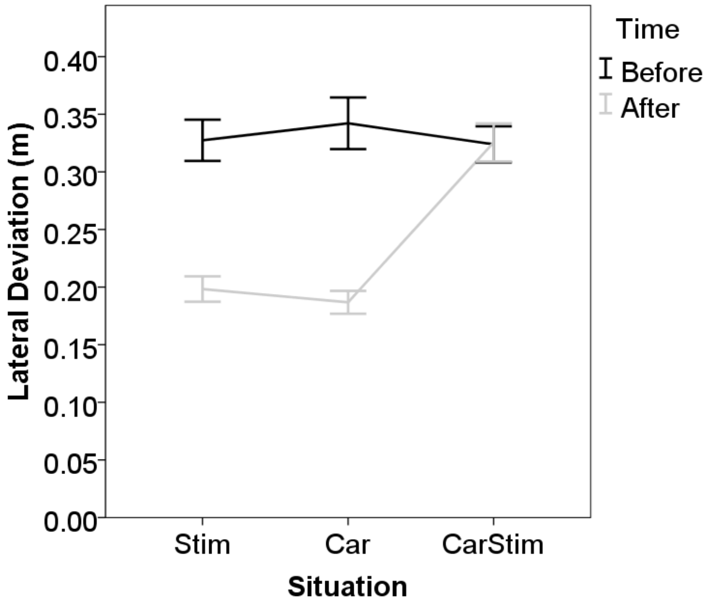


Figure 5-7: The interaction between Situation and Time for Lateral Deviation (H_{2a}).

5.2.8.3. Steering Angle

Data for SteAng were first analysed using a two-way repeated measures ANOVA, with Situation and Time as factors as above. Mauchly's test revealed that the assumption of sphericity had been violated for Situation, therefore Degrees of Freedom for Situation were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{3a}:*** There was a significant main effect of Situation ($F(1.98,1244.49) = 196.07$, $p < 0.001$). Contrasts revealed that situation CarStim induced higher values of SteAng compared to Car ($F(1,629) = 297.35$, $r = 0.57$, $p < 0.001$), while values of SteAng did not differ among situations Stim and Car ($F(1,629) = 0.68$, $p = 0.79$). There was a significant main effect of Time ($F(1,629) = 601.05$, $p < 0.001$). Contrasts revealed that SteAng was lower after any situation arose compared to before ($F(1,629) = 601.05$, $r = 0.70$, $p < 0.001$).

There was a significant interaction between Situation and Time ($F(1.93,1216.11) = 317.76$, $p < 0.001$). Contrasts revealed that while in situation Car values of SteAng were lower after the event compared to before it, this effect was reversed for situation CarStim ($F(1,629) = 421.21$, $r = 0.63$, $p < 0.001$). As a result H_{3a} was accepted. See Figure 5-8 for the interaction between Situation and Time for SteAng.

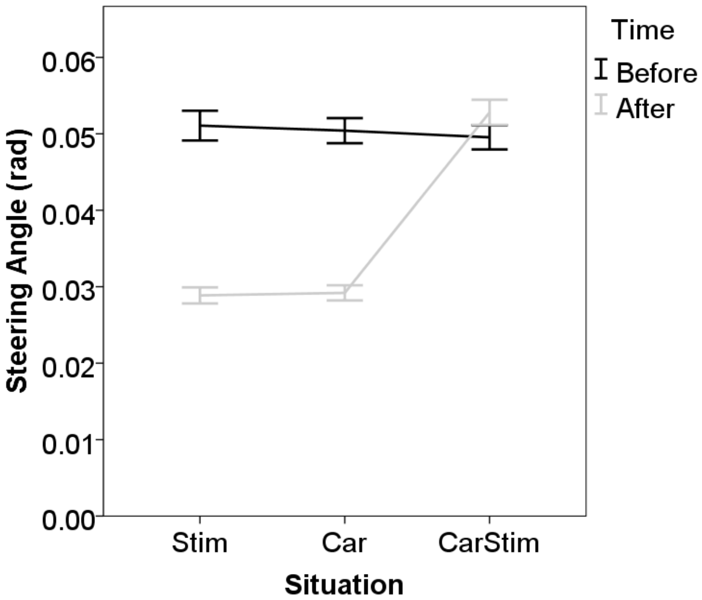


Figure 5-8: The interaction between Situation and Time for Steering Angle (H_{3a}).

Hypotheses H_{3b} and H_{3c}: A separate four-way ANOVA for situations Stim and CarStim (where warnings were present) with Situation, Time, LDU and Modality as factors showed a significant interaction between Situation and LDU ($F(1.93,55.99) = 4.94$, $p < 0.05$). Contrasts revealed that in situation CarStim, the SteAng was significantly higher for L_L compared to L_M, which was not the case in situation Stim ($F(1,29) = 8.99$, $r = 0.49$, $p < 0.05$). No other significant findings related to LDU or Modality were present. As a result H_{3b} was accepted and H_{3c} was rejected.

Results for LatDev and SteAng both show a differential effect of Situation and Time in the driving metrics. While in situations Car and Stim the metrics improved after a situation arose, this was not the case for CarStim.

5.3. Discussion

5.3.1. Response Times

The results for response times indicate a clear advantage of using warnings in synergy with a critical event in the driving task, since ResT was reduced in this manner (H_{1a} was accepted). This result addresses the research space highlighted by (C. L. Baldwin et al., 2012) and also identified as this chapter's motivation. It provides evidence that there is an influence of situational urgency in driver warnings. While there were no differences in terms of ResT for the simple Car and Stim conditions, when these events occurred together in CarStim, there was a pronounced effect in how quickly participants reacted. This also extends the results of Ho & Spence and Ho, Tan & Spence, where spatially predictive audio (Cristy Ho & Spence, 2005) and vibrotactile cues (Cristy Ho, Tan, et al., 2005), meaning cues that correctly predicted the direction of an approaching threat, resulted in lower reaction times compared to their non-predictive variants. A similar result was also found when combining multimodal audio and visual cues (Y.-C. Liu & Jhuang, 2012). Lieberman *et al.* (Liebermann et al., 1995) also found an improvement in reactions when situational urgency was increased, when the lead vehicle braking lights were activated and a deceleration of the vehicle followed, as opposed to no deceleration. In the present study, richer cues were used and it became clear that the advantages of providing combinations of audio, visual and tactile cues hold not only when they predict the direction but also the existence of a critical event.

In terms of the modalities used, there was an advantage of multimodal warnings over unimodal ones in terms of ResT. This is because A, T and V warnings were all slower than AT, AV, TV and ATV ones (H_{1c} was accepted). This advantage of using more than one modality to alert drivers has been discussed in several previous studies (Cristy Ho et al., 2007; Bridget A. Lewis et al., 2013; Oskarsson, Eriksson, & Carlander, 2011). However, never before has this effect been shown in all modality combinations and with a braking task, rather than just a button pressing task as in Experiments 1 and 2. It is argued, that, tn the driving context, there seems to be an additive effect of conveying the same information across more than one sensory channel. As will be discussed later, this advantage in ResT does not necessarily come with a similar advantage in other metrics, such as LatDev and SteAng. Even so, the benefit of using multimodal signals in the driving task, especially when signifying critical situations, is clear, since a quick response to critical events indicates higher cue saliency.

The Level of Designed Urgency of warnings was another factor that influenced responses (H_{1b} was accepted). It was observed, that warnings of high designed urgency elicited significantly quicker responses, even with no prior information related to the type or content of the message given to participants[8]. This result extends prior work like (C. L. Baldwin et al., 2012; Bridget A. Lewis et al., 2013) and findings of Experiments 1 and 2, by evaluating reaction time across unimodal, bimodal and trimodal combinations of warnings and in varying contexts of situational urgency. Guidelines related to fundamental frequency of sounds, colour for visuals and interpulse interval for all three modalities used (Judy Edworthy et al., 1991b; B. A. Lewis & Baldwin, 2012; D. C. Marshall et al., 2007) seem to apply uniformly in the driving task. These should be considered when designing driver displays, since this study shows that the resulting warnings elicit quicker responses when designed to be highly urgent. This is an important finding, especially as the cues used in this study provided no information on the event they signified. The next chapters will explore the influence of using richer multimodal cues than those used in this study and evaluate whether these benefits will hold also in that case.

Finally, there was a significant decrease in performance when encountering warnings involving the visual modality in situation CarStim (interaction between Situation, H_{1a}, and Modality, H_{1c}). It was found that none of the advantages of ATV warnings over AT ones, TV ones over AT ones and V warnings over A and T ones were present in situation CarStim. This is an indication that the benefits of visual signals as driver displays can be limited when there is high visual load in the task at hand. The presentation of a car braking was visual, and in combination with visual signals, it seemed to damage rather than benefit the response times. A similar disadvantage of the visual modality was found in (Murata, Kanbayashi, & Hayami, 2012). In (Y.-C. Liu & Jhuang, 2012), there was also an advantage of audio over visual displays when a visual indicator to a critical situation was provided. This result extends the findings of Experiments 1 and 2, where multimodal signals involving visuals created quicker responses, but in absence of any visually demanding events in the driving task. Horrey & Wickens (Horrey & Wickens, 2004) also found that response times to a critical event degraded when voice dialling was aided by a head-down display. Although no side task was used in this study, these results also suggest a cluttering of the visual modality

[8] This could be partly supported by anecdotal evidence, as follows: During unstructured discussions after some experimental sessions regarding which properties of the signals in their opinion affected the perceived urgency of a stimulus, participants often identified interpulse interval, colour and frequency.

during a visual critical event. As a derived guideline, visual warnings should be avoided in road events of high situational urgency, and signals involving audio or tactile modalities should be preferred, as they reduce the visual load of driving.

5.3.2. Lateral Deviation and Steering Angle

The results of LatDev and SteAng showed a differential effect of Situation on the driving metrics. Situations Stim and Car both led to improved lane keeping behaviour and to less variation in the steering angle (H_{2a} and H_{3a} were accepted). However, this effect was not present in situation CarStim. For SteAng, values were significantly higher after situation CarStim arose. However, the disturbance to the driving behaviour reflected in SteAng was not high enough to increase values of LatDev (see Figures 5-7 and 5-8). The fact is that for both LatDev and SteAng there was no improvement in situation CarStim.

It is argued, that this result can be accounted to the increased workload created by situation CarStim. The simultaneous onset of warnings and a critical event may have created a startle effect, similar to the one observed in (Bliss & Acton, 2003), where participants' control over the simulated vehicle was poorer when critical warnings were delivered. Some participants' spontaneous comments during the experiment, mentioning that situation CarStim was startling, could also be anecdotal evidence to the proposed increased workload. Along with the observed increase of reaction times to signals including the visual modality (see previous section), this observation provides evidence on how the increased amount of visual information can affect driving performance. Lindgren *et al.* (Lindgren et al., 2009) and Liu (Y. C. Liu, 2001) also observed poorer lane keeping and steering behaviours when using visual as opposed to audio displays to aid non-critical tasks (list selection (Lindgren et al., 2009) and navigation (Y. C. Liu, 2001)). Although no differences in terms of modalities were found in this study, the findings of Lindgren *et al.* and Liu also add to the argument that visual load is increased during driving. The addition of a critical visual event as CarStim in the present study could only have added to this load.

From the results of LatDev and SteAng several conclusions can arguably be derived. When there is no critical situation demanding attention, multimodal warnings seem to improve drivers' alertness and lead to a better driving behaviour. This is also true in situation Car, where the onset of the lead brake lights and the deceleration of the lead vehicle seems to have acted as an environmental cue that improved lane keeping behaviour. The benefit of these effects seems to disappear when there is a visual task demanding immediate action, such as situation CarStim. Although response times can improve when a multimodal signal is presented in situation CarStim, lane keeping behaviour is neither improved nor worsened by the cues. Quicker reactions are essential in more critical situations, so the benefits of multimodal cues are valuable in this context. However, lane keeping performance is also essential when there is no imminent critical event, so the benefit of multimodal cues in this case is arguably still present.

Finally, there was marginally better performance in terms of SteAng for warnings at L_M and in situation CarStim (H_{3b} was accepted). It appears that warnings of L_M aided driving behaviour in terms of SteAng more than the ones of L_L or L_H. Combined with the result of intermediate response times achieved by these warnings, they seem a good option to facilitate overall alertness for drivers in situations that require quick but not immediate responses. Interestingly, these situations, for example low fuel, were the ones that these warnings were designed to address.

5.4. Conclusions & Statement of Findings

In Experiment 3 the research question of *"How does situational urgency influence responses to multimodal driver displays varying in urgency?"* was addressed. The warnings designed in Experiments 1 and 2 were used in a critical situation, where a car in front was braking towards the driver. This allowed their investigation in a more demanding driving scenario, requiring speeded driver reactions. Further, the use of the warnings both in the presence and absence of a critical event highlighted their utility as alerts, since drivers' best performance in terms of reaction times was observed when the warnings were delivered along with the critical event. It was found that when situational urgency was increased the responses were quicker. This was also true when cues were multimodal as opposed to unimodal. The limitation of using visual warnings was identified in the most critical, and also visually richer, situation. This was also observed in lane keeping behaviour, which did not improve in this situation, as opposed to the other situations, where it improved. These results extend knowledge of in car warning design by identifying the effect of situational urgency on participant response times as well as driving metrics. They also verify the benefit of using multimodal displays of varying designed urgency to alert drivers in a context of varying situational urgency, a case not previously simulated. The evidence of high visual load during a critical event highlights the limitation of the visual modality when encountering critical events in the driving scene. A unique feature of this study is that it investigated the effect of multimodal displays on driving metrics in detail, evaluating driver responses to each combination of modality and situation. Assessing these metrics in such detail showed the differential effect of providing warnings on the lane keeping and steering behaviours. These results indicate the utility of multimodal driver displays when requiring immediate responses and the potential of non-visual warnings to decrease driving workload. As a result, the following guidelines can be derived from this chapter:

- Using bimodal and trimodal warnings rather than unimodal ones can cause faster reaction times to critical events of high situational urgency. The increased number of modalities in cues increased saliency, resulting to shorter response times during a critical event.
- Using warnings of high designed urgency can speed up reactions critical situations. Applying the critical warning design created in Experiments 1 and 2 in a critical situation in Experiment 3 improved reactions.
- Using warnings of medium designed urgency can provide an overall alertness, as well as improved lane keeping and steering behaviour when no critical event is present. The driving metrics improved when using warnings of medium designed urgency with no critical event present, showing evidence of their benefit in non-critical situations.
- Non-visual signals are more effective in visually demanding situations. The cues including visuals led to slower reactions when signifying the visually rich critical situation presented in Experiment 3.

The next experiments (Experiments 4 and 5 – Chapter 6) extend the research of Experiments 1, 2 and 3, in the area of language-based warnings. Although Experiments 1, 2 and 3 provided an extensive investigation in multimodal cues to alert drivers, the designed warnings did not have semantic association with the signified events, they were abstract. Providing some more detailed information on the events to be signified can have additional benefits, by clarifying the situation at hand to the driver. As a first approach, the cues designed in Experiments 4 and 5 provided information on the signified event in the form of language, they were language-based. Since the focus of this thesis is assessing the saliency and therefore the benefit of multimodal cues, Experiments 4 and 5 attempted to create multimodal language-based cues. Language in the auditory modality commonly takes the form of speech, which was therefore used in Experiments 4 and 5. However there is less investigation on how to create language-based cues for vibration. Therefore, a set of tactile cues transferring some properties of speech into vibration was designed in Experiments 4 and 5. These cues were tested alone or in combination with speech, in terms of subjective and objective measures, in line with the experimental designs of Chapters 4 and 5. In this way, new guidelines on multimodal language-based cues as driver alerts were presented.

6. Investigating Language-based Multimodal Driver Displays

6.1. Introduction

As described in Sections 2.2 and 2.3, there have been studies investigating the efficacy of auditory and tactile warnings in the car, signifying events of varying urgency, *e.g.* (Cristy Ho et al., 2007; Serrano et al., 2011). Auditory warnings used can either be abstract signals (Cristy Ho, Reed, et al., 2006), sounds associated with the events (D. McKeown et al., 2010) or speech (Cristy Ho & Spence, 2005). However, the potential of tactile messages to convey some aspects of speech has been much less investigated in general, and never before in the context of driving. This is interesting to investigate since transferring features of speech to vibration has provided good results in the past (Salminen et al., 2012; Spens et al., 1997). Also, as described in Section 2.4, comparisons between abstract and language-based multimodal warnings for driving are lacking. This motivated the research question: *How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?* In order to answer this question, one needs to design both variations of multimodal driver displays.

Abstract displays were designed and evaluated in Chapters 4 and 5. An elaborate set of guidelines regarding their utility were presented, and subjective and objective measures of the displays were tested. In order to investigate cues with higher semantic associations to the signified events, this chapter describes the design of language-based multimodal driver displays, evaluated with subjective measures in Experiment 4, and with objective measures in Experiment 5. In this way, comparisons with abstract cues will be possible in Chapter 7. Thus, it answers a part of the above research question, namely how language-based multimodal driver displays affect performance. It presents a first investigation of how tactile messages based on speech, called Speech Tactons, perform as warnings for drivers. A set of audio and tactile messages related to different driving events was designed and evaluated in terms of subjective responses and recognition accuracy, in line with (Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002). These cues will be further combined with visual cues in Chapter 7, designing truly multimodal language-based cues, to be compared with multimodal abstract ones.

Section 6.2 describes the design of the Speech Tactons. Section 6.3 describes Experiment 4, evaluating the designed warnings in terms of perceived urgency, annoyance and alerting effectiveness, in order to assess subjective responses to the cues. Section 6.4 describes Experiment 5, evaluating the recognition accuracy of the Speech Tactons when delivered without audio cues, to evaluate cue performance. Section 6.5 discusses the findings of this chapter and finally Section 6.6 describes the derived conclusions and guidelines.

6.2. Warning Design

Six speech messages relating to various driving events were recorded, designed to convey three different urgency levels, Level High (L_H), Level Medium (L_M) and Level Low (L_L), in line with Experiments 1, 2 and 3. The messages used were chosen from (J.D. Lee et al., 2008), where a set of in-vehicle messages were prioritized according to the SAE J2395 standard (SAE, 2002). Messages of highest priority in (J.D. Lee et al., 2008) were mapped to L_H in this study, messages of intermediate priority to L_M and messages of lowest priority to L_L. The word *"Danger!"* was also added before each L_H message, *"Warning!"* before each L_M and *"Notice!"* before each L_L, since these words have shown to provide distinctively different urgency ratings in previous studies (C. L. Baldwin & Moore, 2002; Elizabeth Hellier et al., 2002). The resulting messages can be found on Table 6-1. W_1, N_1 and N_2 were slightly adjusted from their original text in (J.D. Lee et al., 2008) so that no messages would resemble each other in terms of rhythm and number of syllables. All messages were recorded by a female voice actor using a Rode NT2-A[9] condenser microphone. Female speakers have been found to produce messages with higher variation in ratings of urgency (Elizabeth Hellier et al., 2002). In line with (Elizabeth Hellier et al., 2002), the actor was instructed to speak messages of L_H in an urgent manner, as if a loved one was in imminent danger. Accordingly, L_M messages were spoken non-urgently, as if in a friendly conversation with nothing interesting about the situation and L_L messages were spoken in a monotone, deadpan manner.

For the Speech Tactons, all stimuli used were auditory, designed for a C2 Tactor[10]. Initially, the fundamental frequency F_0 (pitch) of each sample of the speech recordings was obtained, resulting in alternating pure tones for each utterance. Then, the changes in intensity of the

[9] http://www.rodemic.com/microphones/nt2-a
[10] http://www.atactech.com/PR_tactors.html

original sound files were added to the tones. This resulted in tactile design *P* (Pitch). As described in Section 2.2.3, roughness and intensity have been effectively used in previous studies as a means to create richer tactile cues. These parameters have not been investigated in a driving setting, therefore they were added in the cue design. In order to investigate the effect of roughness in the resulting cues, an amplitude modulation of 30 *Hz* was added over the P messages, as in (L. M. Brown et al., 2006). This resulted in design *PR* (Pitch-Roughness). Designs P and PR maintained the intensity levels of the original audio recording, *i.e.* they had the same peak levels as the respective audio cues. Finally, to investigate the use of intensity in the cues, two more tactile designs were created, where the maximum possible intensity was used in the cues, while still avoiding clipping. This modification to design P provided design *PI* (Pitch and maximum Intensity) and the same modification to PR provided *PRI* (Pitch-Roughness and maximum Intensity). Designs PI and PRI had peak levels of 0.0 *dBFS*. All tactile cues retained the rhythm and intensity variations of the original utterances. Further, the resulting values of average frequency of all tactile cues never differed to the average frequency of the audio more than ±10*Hz*.

Urgency		
L_H	**D_1:** Danger! Collision Imminent **D:** 1.7 *sec* **P:** -1.9 *dBFS* **AF:** 377 *Hz*	**D_2:** Danger! Tire pressure falling **D:** 1.7 *sec* **P:** -1.9 *dBFS* **AF:** 372 *Hz*
L_M	**W_1:** Warning! Active fog lamps **D:** 2.6 *sec* **P:** -9.5 *dBFS* **AF:** 310 *Hz*	**W_2:** Warning! Left side headlamp out **D:** 2.7 *sec* **P:** -11.1 *dBFS* **AF:** 285 *Hz*
L_L	**N_1:** Notice! Rest area 17 miles **D:** 3.4 *sec* **P:** -15.2 *dBFS* **AF:** 198 *Hz*	**N_2:** Notice! Call and win free tickets **D:** 3.7 *sec* **P:** -16.5 *dBFS* **AF:** 202 *Hz*

Table 6-1: The messages designed, using High (L_H), Medium (L_M) and Low urgency (L_L) levels. For each message the duration (D), peak (P) and average frequency (AF) are reported. The values were acquired by only presenting the actor with the verbal instructions described and using no other intervention. All the resulting messages can be found in http://soundcloud.com/idpolitis.

Overall, 54 different cues were created: 6 Audio (A), 24 Tactile (T), *i.e.* 6 cues × 4 designs and 24 audio and tactile (AT), *i.e.* A cues together with the equivalent T ones. As an example, see Figure 6-1 for the waveforms of N_2. For all modifications, Praat[11] and Audacity[12] software was used.

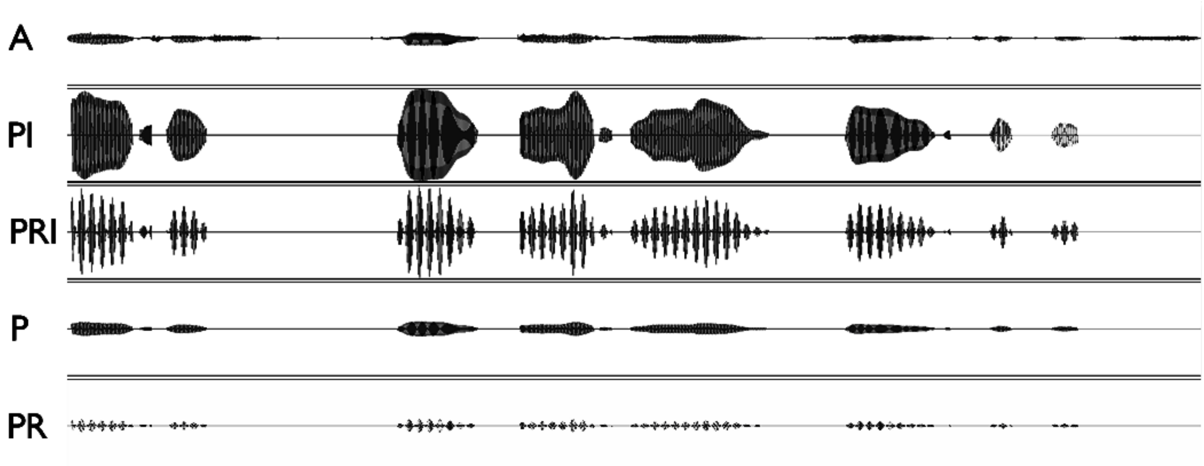


Figure 6-1: The waveforms of message N_2: "Notice! Call and win free tickets."

6.3. Experiment 4: Perceived Urgency, Annoyance and Alerting Effectiveness of Language-based Cues

6.3.1. Motivation

As discussed in Chapters 4 and 5 (Experiments 1, 2 and 3) abstract warnings can create effective alerts, able to modulate perceived urgency multimodally. The investigation described in these experiments provided a set of guidelines for such warnings, however, the semantic association of these warnings with the signified event was low. Therefore, Experiments 4 and 5 were designed in order to investigate the utility of language-based warnings (described in the previous section), using both subjective and objective measures. Designing alerts with higher semantic association with the signified event has provided good results in the past (*e.g.* (Carryl L Baldwin, 2011; Cristy Ho & Spence, 2005; D. McKeown et al., 2010)), however there has been no attempt to evaluate such alerts multimodally, and especially utilising the tactile modality. Experiment 4 used subjective measures of urgency,

---

[11] http://www.fon.hum.uva.nl/praat/
[12] http://audacity.sourceforge.net/

annoyance and alerting effectiveness, so as to present an initial set of results on how these new alerts were perceived.

6.3.2. Design

Experiment 4 investigated the subjective responses provided by participants when exposed to the warnings. A 6×3×4 within subjects design was used with Message, Modality and Design as the independent variables and Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) as the dependent ones. Message had 6 levels (D_1, D_2, W_1, W_2, N_1, N_2), Modality had 3 levels (A, T, AT) and Design had 4 levels (P, PR, PI, PRI). In line with Experiment 1 and (Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002), it was hypothesized that increased urgency of the cues would increase their perceived urgency. Perceived annoyance and alerting effectiveness were also expected to increase with cue urgency, as also observed in (Carryl L Baldwin, 2011). As a result, there were the following hypotheses:

- The ratings of PU will be influenced by Message (H_{1a}), Modality (H_{1b}) and Design (H_{1c});
  - Specifically, PU was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in designs involving higher cue intensity.
- The ratings of PA will be influenced by Message (H_{2a}), Modality (H_{2b}) and Design (H_{2c});
  - Specifically, PA was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in designs involving higher cue intensity.
- The ratings of PAE will be influenced by Message (H_{3a}), Modality (H_{3b}) and Design (H_{3c}).
  - Specifically, PAE was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in designs involving higher cue intensity.

6.3.3. Procedure

Twenty-two participants (9 female) aged between 18 and 44 years ($M = 25.04$, $SD = 5.95$) took part in this experiment. They had not participated in the previous experiments, except one, who had participated in Experiments 1 and 2. They all held a valid driving licence and had between 1 and 27 years of driving experience (*M* = 5.79, *SD* = 5.85). Participants were all right handed and reported normal hearing. They were either University students or employees. The experiment took place in a University room, where participants sat in front of 27-inch Dell 2709W monitor and a PC running the experimental software. They wore a set of Sennheiser HD 25-1 headphones and a wristband on their left hand with a C2 Tactor attached on the inside of the band (see Figure 6-2), in line with (Pratt et al., 2012; Tuuri et al., 2010). Participants provided all responses using a mouse with their right hand and were asked to rest their left hand on the desk. To cover the Tactor noise, car sound was played throughout the experiment, as in Experiments 1,2 and 3.

After being welcomed and explained the experimental procedure, participants were exposed to the 54 cues (6 A, 24 T, 24 AT) in a random order, to familiarize them with the signals. For each cue, they had the option to repeat it or to proceed to the next one when they felt familiar with it. Afterwards, they were again presented with the cues and asked to rate them all in terms of PA, PU and PAE, by completing a 5-point Likert scale for each rating, in line with (Carryl L Baldwin, 2011). In all ratings, the scale was: Not at all (1), Slightly (2), Moderately (3), Very (4) and Extremely (5). Participants were asked to imagine they were driving and wearing a wrist mounted device like a smart watch for vibration, while also listening to their car speakers for sound. The wrist was selected, since previous studies (Pratt et al., 2012; Tuuri et al., 2010) have shown good recognition of vibration on this area, while using the abdomen created higher ratings of perceived annoyance in Experiment 1. The wrist has also shown good results in previous studies investigating different locations for vibration, including the waist (C. Ho & Spence, 2009) and fingertips (Summers et al., 2005). Each cue was presented twice, resulting to 108 trials. The experiment lasted about 30 minutes and participants were then prepared for Experiment 5 in the same session.

Figure 6-2: The wristband and the C2 Tactor used in the experiments.

6.3.4. Results

6.3.4.1. Perceived Urgency

Data for PU were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Modality and Modality × Message, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{1a}:*** There was a significant effect of Message ($F(5,215) = 223.21$, $p < 0.001$). Contrasts revealed that D_1 was perceived as more urgent that D_2 ($F(1,43) = 7.36$, $r = 0.38$, $p < 0.05$), D_2 more urgent than W_1 ($F(1,43) = 124.39$, $r = 0.86$, $p < 0.001$), W_2 more urgent than N_1 ($F(1,43) = 112.37$, $r = 0.85$, $p < 0.001$) and N_1 more urgent than N_2 ($F(1,43) = 9.67$, $r = 0.43$, $p < 0.05$). Therefore, H_{1a} was accepted. ***Hypothesis H_{1b}:*** There was a significant main effect of Modality ($F(1.40,60.26) = 6.27$, $p < 0.01$). Contrasts revealed that modality AT created higher ratings of PU compared to A and T ($F(1,43) = 15.34$, $r = 0.51$, $p < 0.001$). Therefore, H_{1b} was accepted. See Figure 6-3 for mean ratings of PU across modalities, Figure 6-4 for mean ratings across messages and Table 6-2 for the pairwise comparisons of PU between messages.

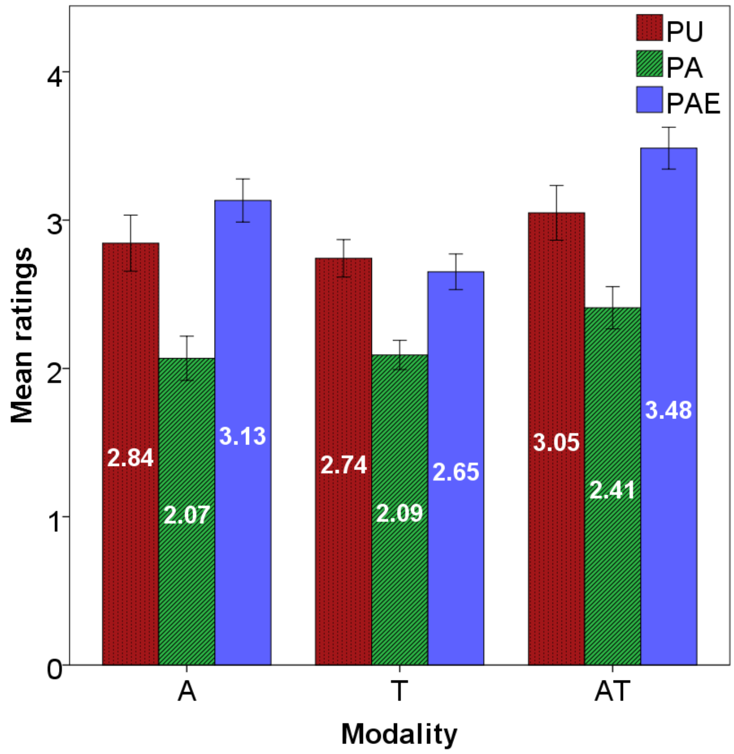


Figure 6-3: Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across modalities (hypotheses H_{1b}, H_{2b}, H_{3b}). Error bars indicate 95% confidence intervals.

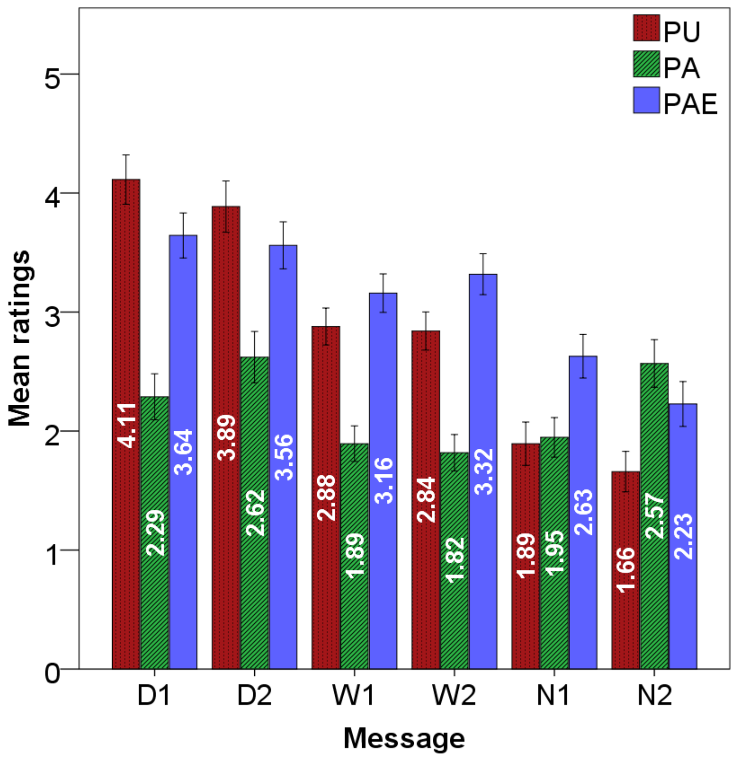


Figure 6-4: Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across messages (hypotheses H_{1a}, H_{2a}, H_{3a}).

	N_2	N_1	W_2	W_1	D_2	D_1
N_2		**.003**	**.000**	**.000**	**.000**	**.000**
N_1	**.003**		**.000**	**.000**	**.000**	**.000**
W_2	**.000**	**.000**		.640	**.000**	**.000**
W_1	**.000**	**.000**	.640		**.000**	**.000**
D_2	**.000**	**.000**	**.000**	**.000**		**.010**
D_1	**.000**	**.000**	**.000**	**.000**	**.010**	

Table 6-2: Pairwise comparisons between messages for Perceived Urgency (H_{1a}). Messages are sorted by their mean values of PU. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects (H_{1a}, H_{1b}): There was a significant interaction between Modality and Message ($F(6.34,272.84) = 68.25$, $p < 0.001$). Contrasts revealed that the significant differences in ratings of PU described above were not present in modality T, when comparing D_2 with W_1 ($F(1,43) = 70.18$, $r = 0.79$, $p < 0.001$) and W_2 with N_1 ($F(1,43) = 41.79$, $r = 0.70$, $p < 0.001$). They also revealed that message D_2 did not differ in rating of PU in modalities A and AT ($F(1,43) = 4.24$, $r = 0.30$, $p < 0.05$). See Figure 6-5 for the interaction between Modality and Message.

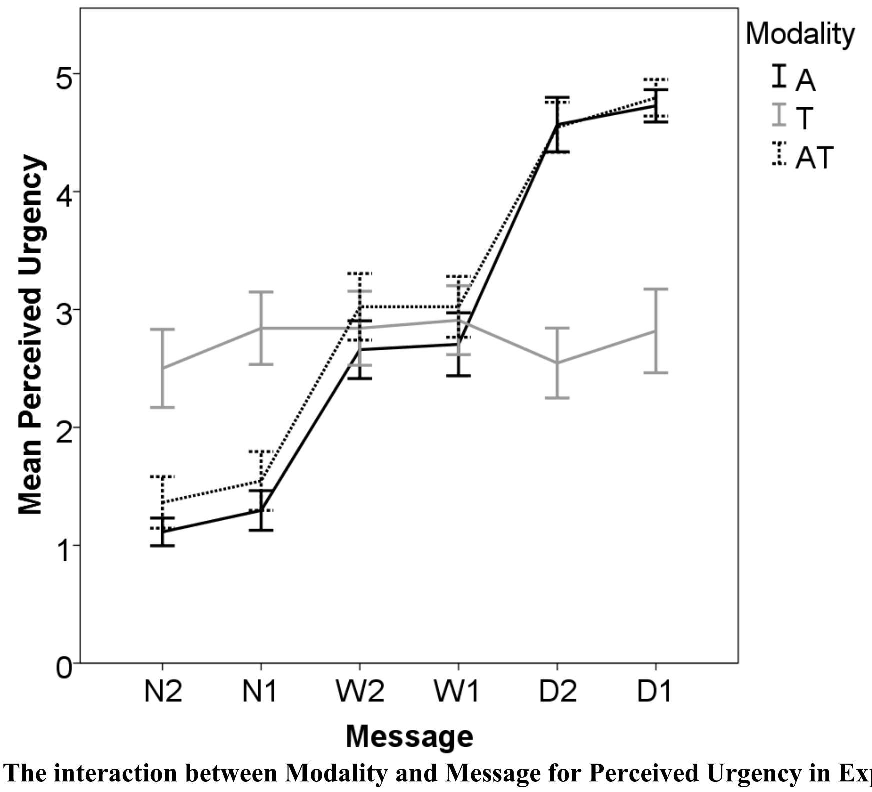


Figure 6-5: The interaction between Modality and Message for Perceived Urgency in Experiment 4 (H_{1a} – H_{1b}). Messages are sorted by their mean values of PU.

Data for Modalities T and AT, where there was a Design present, were analysed in terms of their PU using a three-way repeated measures ANOVA, with Modality, Message and Design as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Design, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above. ***Hypothesis H_{1c}:*** There was a significant main effect of Design ($F(2.38,102.36) = 17.98$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PU compared to P ($F(1,43) = 7.27$, $r = 0.38$, $p < 0.05$), P higher ratings compared to PRI ($F(1,43) = 4.28$, $r = 0.30$, $p < 0.05$) and PRI higher ratings compared to PR ($F(1,43) = 10.08$, $r = 0.44$, $p < 0.05$). Therefore, H_{1c} was accepted. See Figure 6-6 for mean ratings of PU across designs. ***Interactions between main effects (involving H_{1c}):*** There was a significant interaction between Modality and Design ($F(3,129) = 9.29$, $p < 0.001$) and Message and Design ($F(15,645) = 2.20$, $p < 0.05$) (see Figure 6-7). Contrasts revealed that the difference of ratings of design PR compared to PRI described above, was higher for T modality compared to AT ($F(1,43) = 4.75$, $r = 0.31$, $p < 0.05$). They also revealed that while design PI had higher ratings compared to P for message W_1, it had similar ones for D_2 ($F(1,43) = 7.80$, $r = 0.39$, $p < 0.05$).

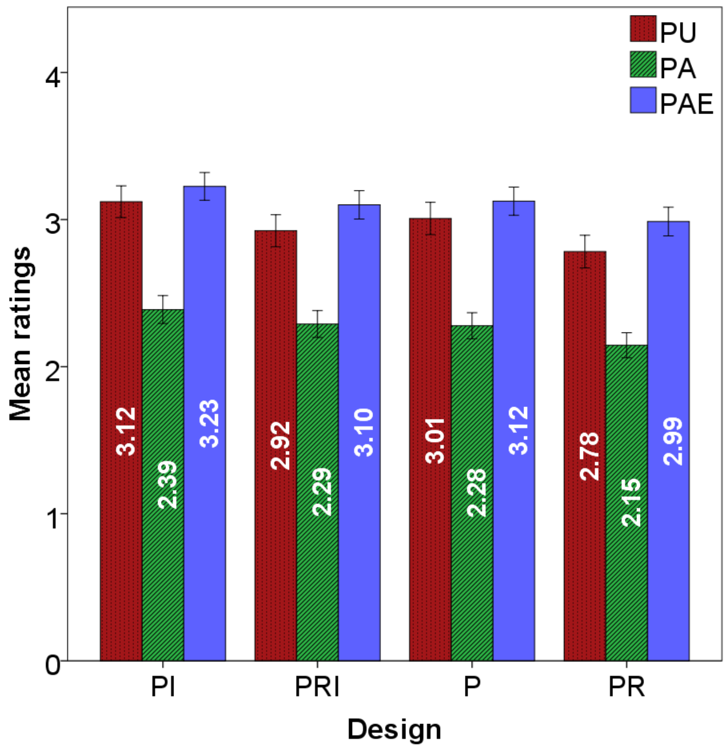


Figure 6-6: Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across designs (hypothesis H_{1c}, H_{2c}, H_{3c}).

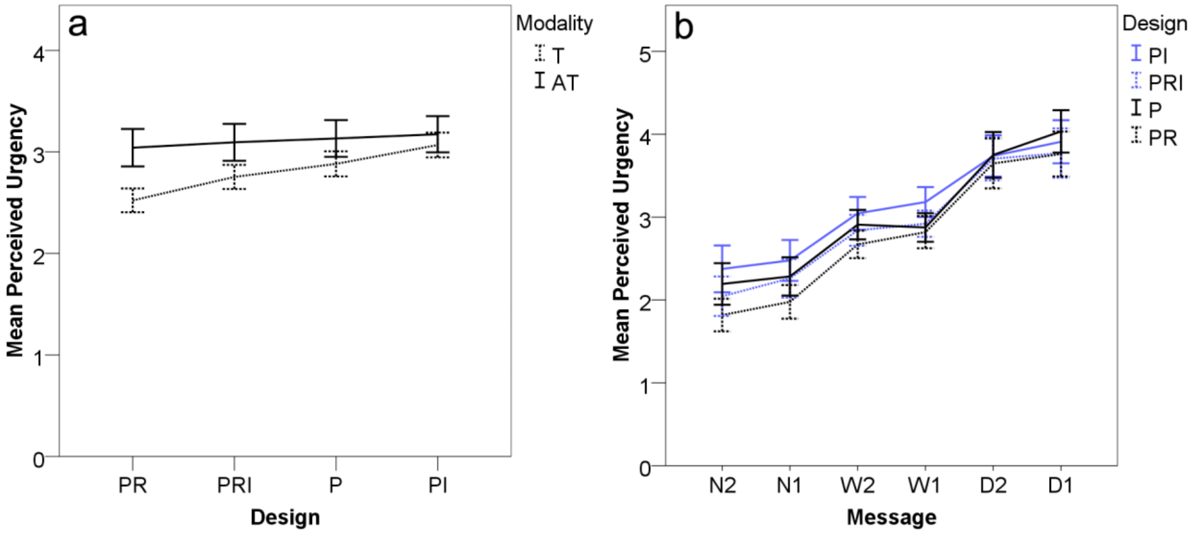


Figure 6-7: The interaction between Modality and Design (H_{1b} – H_{1c}, a) and Message and Design (H_{1a} – H_{1c}, b) for Perceived Urgency in Experiment 4. Messages and designs are sorted by their mean values of PU.

6.3.4.2. Perceived Annoyance

Data for PA were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Message and Modality × Message, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{2a}:*** There was a significant effect of Message ($F(2.46,105.94) = 15.03$, $p < 0.001$). Contrasts revealed that D_2 and N_2 had higher PA than D_1 ($F(1,43) = 17.52$, $r = 0.54$, $p < 0.001$), which in turn had higher PA than N_1, W_1 and W_2 ($F(1,43) = 5.11$, $r = 0.32$, $p < 0.05$). Therefore, H_{2a} was accepted. ***Hypothesis H_{2b}:*** There was a significant main effect of Modality ($F(2,86) = 8.42$, $p < 0.001$). Contrasts revealed that modality AT created higher ratings of PA compared to A and T ($F(1,43) = 13.53$, $r = 0.49$, $p < 0.001$). Therefore, H_{2b} was accepted. See Figure 6-3 for mean ratings of PA across modalities, Figure 6-4 for ratings across messages and Table 6-3 for pairwise comparisons of PA between messages.

	W_2	W_1	N_1	D_1	N_2	D_2
W_2		.337	.257	**.000**	**.000**	**.000**
W_1	.337		.553	**.002**	**.000**	**.000**
N_1	.257	.553		**.029**	**.000**	**.000**
D_1	**.000**	**.002**	**.029**		.105	**.000**
N_2	**.000**	**.000**	**.000**	.105		.762
D_2	**.000**	**.000**	**.000**	**.000**	.762	

Table 6-3: Pairwise comparisons between messages for Perceived Annoyance (H_{2a}). Messages are sorted by their mean values of PA. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects (H_{2a}, H_{2b}): There was a significant interaction between Modality and Message ($F(5.84,251.30) = 9.52$, $p < 0.001$). Contrasts revealed that D_1 had higher PA compared to N_1 in modality A ($F(1,43) = 13.77$, $r = 0.49$, $p = 0.001$) and AT ($F(1,43) = 4.67$, $r = 0.31$, $p < 0.05$) while there was no such difference in modality T. Further, A had higher PA than T for message D_2 while there was no difference of ratings between these two modalities for N_2 ($F(1,43) = 8.15$, $r = 0.40$, $p < 0.05$). See Figure 6-8 for the interaction between Modality and Message.

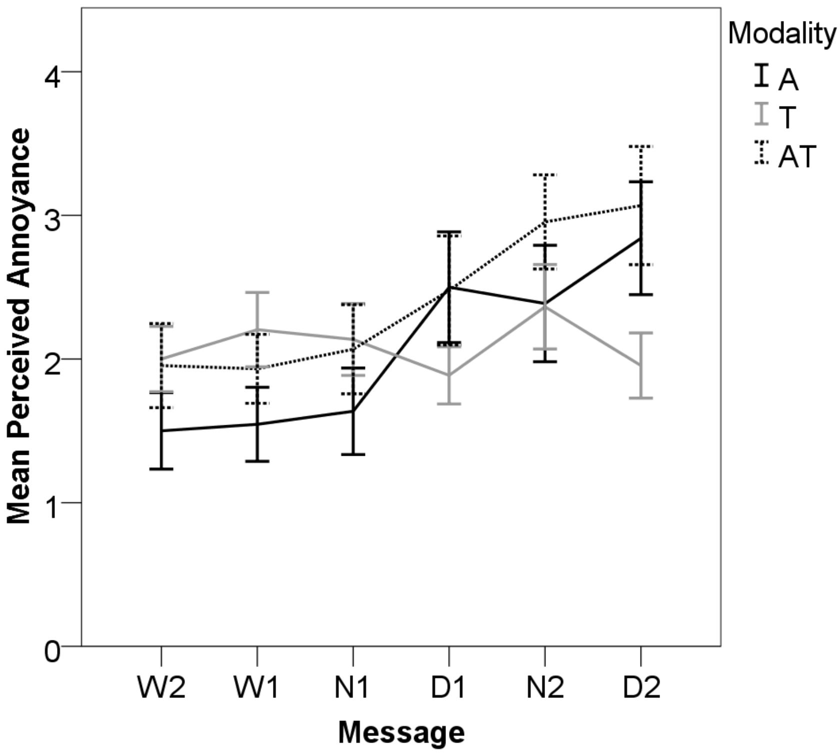


Figure 6-8: The interaction between Modality and Message for Perceived Annoyance in Experiment 4 (H_{2a} – H_{2b}). Messages are sorted by their mean values of PA.

Data for Modalities T and AT, where there was a Design present, were analysed for PA using a three-way repeated measures ANOVA, with Modality, Message and Design as factors.

Mauchly's test revealed that the assumption of sphericity had been violated for Design and Modality × Message × Design, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above. ***Hypothesis H_{2c}:*** There was a significant main effect of Design ($F(2.46,105.95) = 13.31$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PA compared to PRI and P ($F(1,43) = 9.49$, $r = 0.42$, $p < 0.05$) and the latter created higher ratings of PA compared to PR ($F(1,43) = 13.50$, $r = 0.49$, $p < 0.001$). Therefore, H_{2c} was accepted. See Figure 6-6 for mean ratings of PA across designs. ***Interactions between main effects (involving H_{2c}):*** There was a significant interaction between Message and Design ($F(15,645) = 1.78$, $p < 0.05$) and Modality, Message and Design ($F(9.24,397.50) = 2.00$, $p < 0.05$) (see Figure 6-9). Contrasts revealed that, while ratings of PA for message W_2 were higher when using design PI compared to PRI, they did not differ for message W_1 ($F(1,43) = 15.84$, $r = 0.52$, $p < 0.001$). Further, ratings for message W_2 were higher when using design P compared to PRI, while this was reversed for message W_1 ($F(1,43) = 4.50$, $r = 0.31$, $p < 0.05$). Similarly, while ratings for message N_1 were higher when using design PI compared to PRI, the effect was reversed for message D_1 ($F(1,43) = 4.20$, $r = 0.30$, $p < 0.05$). Also, ratings of message N_2 were lower when using design PR compared to P, while these two designs did not create different ratings for D_2 ($F(1,43) = 5.66$, $r = 0.34$, $p < 0.05$). Finally, in modality T, ratings for D_1 were higher when using design P compared to PRI, which was reversed for D_2 ($F(1,43) = 5.46$, $r = 0.33$, $p < 0.05$). Similarly, in modality T, ratings for D_1 were higher when using design P compared to PR, and they did not differ for these designs for D_2 ($F(1,43) = 4.61$, $r = 0.31$, $p < 0.05$). However, in modality AT the above effects were not observed.

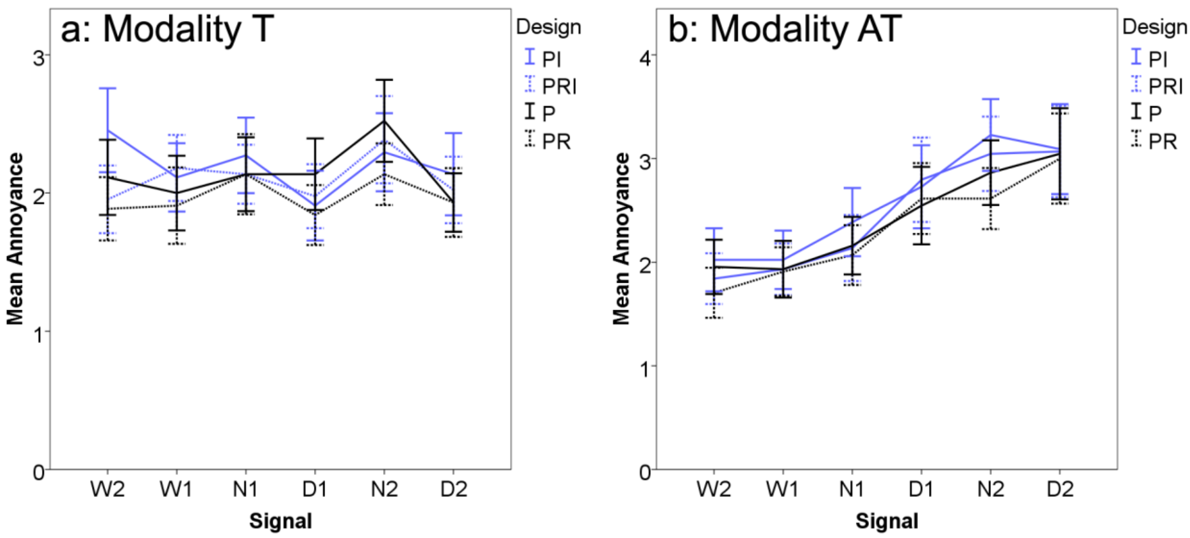


Figure 6-9: The interaction between Message and Design for Perceived Annoyance (H_{2a} – H_{2c}) in modalities T (a) and AT (b) for Experiment 4.

6.3.4.3. Perceived Alerting Effectiveness

Data for PAE were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Modality, Message and Modality × Message, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{3a}:*** There was a significant effect of Message ($F(3.48,149.83) = 55.23$, $p < 0.001$). Contrasts revealed that D_1 and D_2 were rated higher in PAE compared to W_2 ($F(1,43) = 6.23$, $r = 0.35$, $p < 0.05$), W_1 higher compared to N_1 ($F(1,43) = 32.66$, $r = 0.66$, $p < 0.001$) and N_1 higher compared to N_2 ($F(1,43) = 19.73$, $r = 0.56$, $p < 0.001$). Therefore, H_{3a} was accepted. ***Hypothesis H_{3b}:*** There was a significant main effect of Modality ($F(1.63,70.28) = 28.48$, $p < 0.001$). Contrasts revealed that modality AT created higher ratings of PAE compared to A ($F(1,43) = 13.99$, $r = 0.49$, $p = 0.001$) and A higher ratings compered to T ($F(1,43) = 18.44$, $r = 0.55$, $p < 0.001$). Therefore, H_{3b} was accepted. See Figure 6-3 for mean ratings of PAE across modalities, Figure 6-4 for mean ratings across messages and Table 6-4 for the pairwise comparisons of PAE between messages.

	N_2	N_1	W_1	W_2	D_2	D_1
N_2		**.000**	**.000**	**.000**	**.000**	**.000**
N_1	**.000**		**.000**	**.000**	**.000**	**.000**
W_1	**.000**	**.000**		.105	**.001**	**.000**
W_2	**.000**	**.000**	.105		**.017**	**.000**
D_2	**.000**	**.000**	**.001**	**.017**		.274
D_1	**.000**	**.000**	**.000**	**.000**	.274	

Table 6-4: Pairwise comparisons between messages for Perceived Alerting Effectiveness (H_{3a}). Messages are sorted by their mean values of PAE. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects (H_{3a}, H_{3b}): There was a significant interaction between Modality and Message ($F(6.68,287.05) = 25.20$, $p < 0.001$), indicating that the differences in ratings of PAE described above were not present in modality T when comparing D_2 with W_2 ($F(1,43) = 15.40$, $r = 0.51$, $p < 0.001$), W_1 with N_1 ($F(1,43) = 6.95$, $r = 0.37$, $p < 0.05$) and N_1 with N_2 ($F(1,43) = 9.38$, $r = 0.42$, $p < 0.05$). Further, the differences between D_2 and W_2 described in the main effect were more pronounced in modality A compared to AT ($F(1,43) = 5.20$, $r = 0.33$, $p < 0.001$). See Figure 6-10 for the interaction between Modality and Message.

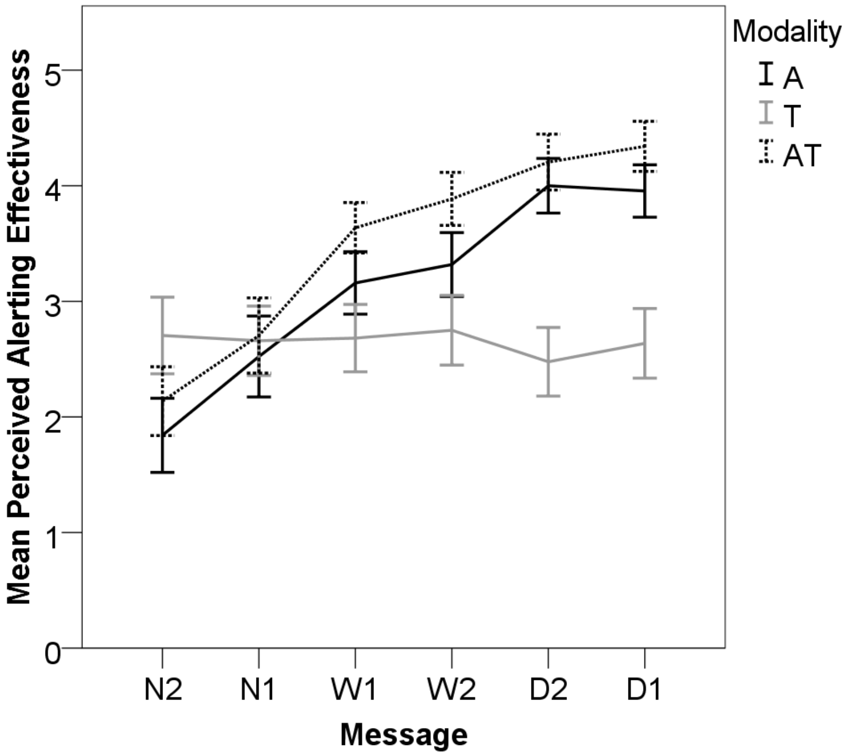


Figure 6-10: The interaction between Modality and Message for Perceived Alerting Effectiveness in Experiment 4 (H_{3a} – H_{3b}). Messages are sorted by their mean values of PAE.

Data for Modalities T and AT, where there was a Design present, were analysed for PAE using a three-way repeated measures ANOVA, with Modality, Message and Design as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Message × Design and Modality × Message × Design, therefore Degrees of Freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above. ***Hypothesis H_{3c}:*** There was also a significant main effect of Design ($F(3,129) = 12.90$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PAE compared to P and PRI ($F(1,43) = 6.98$, $r = 0.37$, $p < 0.05$) and the latter higher ratings compared to PR ($F(1,43) = 9.11$, $r = 0.42$, $p < 0.001$). Therefore, H_{3c} was accepted. See Figure 6-6 for mean ratings of PAE across designs.

From Experiment 4 it was clear that AT was rated higher in all measures compared to A and T, showing clear evidence of the usefulness of the designed cues when modalities were combined. Further, it was evident that the urgency designed in the warnings was reflected in their PU. PAE escalated according to PU, indicating that messages signifying situations of higher importance were regarded as more useful. PA was higher for messages of both L_H and L_L and lower for L_M, allowing for several interpretations which will be discussed later. Finally, ratings were not as responsive to messages in the T modality, a further indication of the higher utility of the cues when presented multimodally. In order to investigate the ability of participants to recognise the tactile cues without the audio present, Experiment 5 was performed immediately after Experiment 4, investigating the recognition accuracy of the messages.

6.4. Experiment 5: Recognition Accuracy of Language-based Cues

6.4.1. Motivation

Experiment 5 focused on objective measures related to the recognition of the tactile counterparts of the new alerts, the Speech Tactons. Since the new cues were evaluated using subjective measures in Experiment 4, an evaluation using objective measures would provide results on how effective these cues would be in a time-bound task. The task involved recognition of the tactile cues, with no audio cues present, so as to evaluate to what extent the Speech Tactons would be able to stand alone. This would be important to evaluate, since effective recognition of the unimodal cues would increase the complexity of messages to be conveyed through vibration. In this way, a complete set of results for the new alerts designed would be presented, looking into their effectiveness using both subjective (Experiment 4) and objective measures (Experiment 5).

6.4.2. Design

Experiment 5 investigated the recognition accuracy of participants when exposed to the T warnings. To investigate participants' performance in identifying individual cues, as well as recognizing their urgency, two measures were used. Recognition Accuracy (RecA) was 1 when participants recognized the exact message, *e.g.* responded "N_2" when the message was indeed N_2, and 0 in all other cases. Recognition Accuracy of Urgency (RecAU) was 1 when participants recognized the urgency of the message, *e.g.* responded "D_1" or "D_2" when the message was D_2, and 0 in all other cases. A 6×4 within subjects design was used with Message and Design as the independent variables and RecA and RecAU as the dependent ones. In line with (Salminen et al., 2012; Tuuri et al., 2010), it was hypothesized that the rhythmic and intensity variations of the tactile cues would affect recognition, leading to possible identification. Further, longer cues were expected to produce better recognition results, since participants would have more time to identify them. Finally, as in (L. M. Brown et al., 2006), it was expected that the added roughness and intensity to the tactile cues would affect their recognition. These features were specifically expected to increase recognition, due to the more distinct character of the resulting cues. As a result, there were the following hypotheses:

- RecA will be influenced by Message (H_{4a}) and Design (H_{4b}).
  - Specifically, RecA was expected to increase in longer cues (cues of lower designed urgency), and in designs involving higher cue intensity and roughness.
- RecAU will be influenced by Message (H_{5a}) and Design (H_{5b}).
  - Specifically, RecAU was expected to increase in longer cues (cues of lower designed urgency), and in designs involving higher cue intensity and roughness.

6.4.3. Procedure

Participants and equipment were identical to Experiment 4. After completing Experiment 4 participants were again exposed to the 54 cues (6 A, 24 T, 24 AT) in a random order, to further familiarize them with the signals and the mapping between T and AT. Then, they were presented with only the 24 T warnings, repeated three times each in a random order, resulting in 72 trials. Participants were asked to map each warning to one of the A messages (D_1, D_2, W_1, W_2, N_1 or N_2). They were able to do this by selecting one option out of six available, each displaying the full text of the speech warning. They were also free to feel the T cues as many times as needed before responding. The experiment lasted about 30 minutes and finally participants were debriefed and paid £6 for participating to both experiments.

6.4.4. Results

Data for RecA and RecAU were treated as dichotomous and analysed with Cochran's Q tests. ***Hypothesis H_{4a}:*** It was found that N_1 had higher RecA than D_1 ($Q(1) = 6.03$, $p < 0.05$), D_2 ($Q(1) = 9.56$, $p < 0.05$) and W_1 ($Q(1) = 5.65$, $p < 0.05$). Further, N_2 had higher RecA than D_1 ($Q(1) = 16.12$, $p < 0.001$), D_2 ($Q(1) = 23.25$, $p < 0.001$), W_1 ($Q(1) = 18.31$, $p < 0.001$), W_2 ($Q(1) = 10.38$, $p < 0.05$) and N_1 ($Q(1) = 5.23$, $p < 0.05$). Therefore, H_{4a} was accepted. ***Hypothesis H_{4b}:*** There were no significant differences in RecA found between designs. Therefore, H_{4b} was rejected. ***Hypothesis H_{5a}:*** In terms of RecAU, it was found that W_1 had lower values compared to D_1 ($Q(1) = 25.31$, $p < 0.001$), D_2 ($Q(1) = 24.31$, $p < 0.001$, W_2 ($Q(1) = 12.79$, $p < 0.001$), N_1 ($Q(1) = 11.11$, $p < 0.001$) and N_2 ($Q(1) = 24.25$, $p < 0.001$). Also, N_1 had lower RecAU compared to D_1 ($Q(1) = 5.08$, $p < 0.05$) and D_2 ($Q(1) = 4.07$, p

< 0.05). Therefore, H_{5a} was rejected. ***Hypothesis H_{5b}:*** Additionally, design P had higher RecAU compared to PR (78% vs. 71%, $Q(1) = 7.86$, $p < 0.05$). Designs PI (73%) and PRI (75%) did not show significant differences in RecAU compared to any other designs. Therefore, H_{5b} was rejected. See Table 6-5 for mean RecA and RecAU across messages.

	D_1	D_2	W_1	W_2	N_1	N_2
RecA	**50%**	**50%**	**52%**	**57%**	**61%**	**68%**
RecAU	**80%**	**80%**	**60%**	**74%**	**73%**	**79%**

Table 6-5: RecA and RecAU across messages for Experiment 5 (hypotheses H_{4a}, H_{5a}).

6.5. Discussion

The results of PU support the argument that the messages designed conveyed the desired urgency (H_{1a}). Participants rated the messages as expected, highlighting that available guidelines in (C. L. Baldwin & Moore, 2002; Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002) are also valid for multimodal messages. It is interesting how a set of simple guidelines to a voice actor succeeded in producing messages of distinct differences in average frequency and peak. This is arguably an indication of the applicability of guidelines presented in (Elizabeth Hellier et al., 2002). Even messages of the same level of designed urgency presented different rating results, which still were significantly different to messages of different levels. This can provide potential for selecting different cues for one level and enriching the interaction. In a further result, AT messages had higher ratings of PU (H_{1b}), which can be an improvement compared to A or T messages, when a situation of high criticality needs to be conveyed. Similar improvements of speech warnings when combined with visuals have been observed in studies like (Cao, Mahr, et al., 2010; Cao, Theune, et al., 2010). With the studies described in this chapter, evidence has been presented that tactile cues can also improve responses to speech warnings. This adds to the existing body of work, suggesting enhanced responses to multimodal signals versus unimodal ones, *e.g.* (Bridget A. Lewis et al., 2013), as well as to such observations in Experiments 1, 2 and 3. Further, T messages did not present highly different ratings of PU, adding to the argument that such cues work better when used multimodally.

In terms of tactile designs (H_{1c}), it was clear that intensity was the main factor that led to higher PU (since results showed that PI was rated higher than P and PRI higher than PR), while roughness led to lower PU (since results showed that PRI and PR were rated lower than PI and P). This strengthens the evidence that intensity of the tactile part is useful to create more urgent messages and can be compared with (C. L. Baldwin & Moore, 2002; Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002), where high intensity of audio affected PU ratings. Roughness seems to produce the opposite effect, unlike some prior studies such as (L. M. Brown et al., 2006). However, it needs to be noted that in (L. M. Brown et al., 2006) roughness was not used to design urgency per se, but to signify more or less important scheduling events, and the tactile cues were not speech based. The above results are arguably promising when designing AT cues based on Speech Tactons and provide a variety of ways to do this.

Data for PA, while presenting interesting variations, had low values overall, since all average values across all factors observed were below 3 (moderately annoying). This is an improvement compared to previous studies of this thesis using such cues, where annoyance was higher for T, *e.g.* in Experiment 1. Consistent with previous chapters, multimodal messages were rated higher in PA (H_{2b}). The T modality again created less variation in ratings of annoyance across messages, adding to the argument that T cues can be better used to enhance the responses rather than used standalone. Interestingly, both cues of high and low designed urgency were rated higher in PA (H_{2a}, cues D_2 and N_2). This could be partly supported by anecdotal evidence, since some participants commented firstly that they would not like warnings for non-important events, and secondly that although L_H cues were more annoying, this was desired by them since it would increase their alertness. Looking at the results of PAE, which will be discussed later, this interpretation could be further supported. It is therefore argued, that on the one hand more modalities create more annoyance, but this is not necessarily a flaw when the event signified is critical. Since intermediate designed urgency is less annoying to participants according to the results of Experiment 4, unimodal signals seem to be a good option to choose in this case. In terms of tactile design (H_{2c}), cues using intensity were again rated higher (PI was rated highest). This adds to the argument that intensity can be used for urgent events, where higher PA can be tolerated. Roughness did not lead to as high ratings (PR was rated lowest), arguably making this feature a better candidate for low urgency cues, also considering the similarly low PU ratings for roughness.

Results for PAE are arguably encouraging, since they are similar to PU, with often higher average values. As generally observed in this study, more modalities increase ratings in all measures (H_{3b}), and PAE is no exception. As in (Salminen et al., 2012), participants rated the audiotactile messages as more arousing and dominant compared to the audio ones. The results of Experiment 4 show that Speech Tactons are also rated as more effective when combined with audio. With the technique suggested for designing these Tactons, a simple way to derive such messages from speech is suggested. As mentioned earlier, various techniques have been suggested in the past to map speech to vibration, but with mixed results. This chapter's contribution is a comprehensive examination of how speech and tactile cues are perceived by drivers and shows positive results not previously observed, namely that alerts signifying more critical events are rated as more effective (H_{3a}). These results are also partly supported by anecdotal evidence of participants stating that they valued more being warned about important events. They can also arguably explain why higher ratings of annoyance for L_H messages were acceptable according to some participants' comments during unplanned discussions after the end of Experiments 4 and 5, while for L_L they were not. This also relates to (Judy Edworthy et al., 2003), where highly urgent messages were perceived as more appropriate. Therefore, warnings of high criticality can arguably be more alerting even at the cost of more annoyance. In terms of tactile design (H_{3c}), cues with high intensity were rated as more effective compared to cues with added roughness (PI rated highest and PR lowest), addressing open questions of (E. E. Hoggan & Brewster, 2006) on how intensity would perform compared to roughness, and suggesting that intensity can be suitable as a design parameter.

Finally, recognition accuracy produced acceptable values overall, but especially good values of RecAU. For values of RecA (H_{4a}, H_{4b}), a random response, indicating that participants were just guessing the messages, would provide RecA percentages of 100% ÷ 6 = 16.7%. In Experiment 5, the lowest RecA observed was 50%, well above that value. Looking at RecAU (H_{5a}, H_{5b}), the results are even more encouraging. Interestingly, high urgency messages performing poorer in RecA performed best in RecAU. This is also partly supported by some participants' comments, during unplanned discussions after the experiments, mentioning that it was easy for them to recognise which level a message belonged to, but not as easy to tell which message it was, especially when messages were short. For longer messages, RecA performance was also high (H_{4a}), since participants had more time to distinguish the different properties of the T cues. This is arguably also an indication that fewer individual T cues could be better recognised compared to more, since the cues would be more different. It is also in line with (L. M. Brown et al., 2006), where the number of Tactons needed to be reduced to achieve better recognition results. Experiment 5 was not intended to suggest Speech Tactons to be presented on their own, but only along with audio. This is especially true for L_H cues, where ambiguity of message meanings cannot be tolerated. However, results can even support individual presentation, if cues are limited in number and not urgent.

Finally, P showed better RecAU values compared to PR, indicating that roughness may also hinder recognition and should be avoided as a single design parameter.

In terms of the interactions between main effects, the interaction between Message and Modality ($H_{1a} – H_{1b}$, $H_{2a} – H_{2b}$, $H_{3a} – H_{3b}$), indicated that Speech Tactons produce a lower variation in ratings of PU, PA and PAE when presented unimodally. As described above, this can lead to the suggestion that they are better distinguishable, and therefore more suggested, in combination with speech, in order to provide higher cue saliency. Further, interactions between Modality and Design showed an elevated variation of ratings of PU for unimodal tactile cues in different tactile designs ($H_{1b} – H_{1c}$). This could be attributed to the lower difficulty of recognising different features of tactile cues when they are presented alone. Finally, interactions between Message and Design for PA ($H_{2a} – H_{2c}$) presented a low effect size in all cases and highlight how annoyance can present variations in subjective ratings, electing messages of all levels of urgency as most annoying, counter to the results of PU and PAE, where messages of higher designed urgency were rated higher.

6.6. Conclusions & Statement of Findings

In Experiments 4 and 5, a set of language-based cues in the audio and tactile modalities were designed and evaluated. This was the first step in order to compare truly multimodal abstract and language-based cues and answer the research question of *"How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?"* Results showed that the addition of these new cues improved subjective responses of drivers to speech warnings. The warnings were clearly distinguished in terms of urgency, their annoyance was low and their alerting effectiveness changed similarly to urgency, increasing for more urgent messages and for multimodal cues. Recognition accuracy of the tactile cues' urgency was high overall and recognition accuracy of individual messages was higher for longer cues. This provides potential for using the tactile cues even alone for non-critical events, if their number is limited. Speech Tactons are suggested to accompany speech warnings, so as to make use of the observed advantages of multimodal cues, but not for low urgency situations, to avoid annoyance. With the technique provided, these tactile cues can be easily designed and added to warnings that will improve drivers' responses. Further, the full set of audio, visual and tactile cues are now available both in their abstract and their language-based designs, to be compared in Chapter 7 and answer RQ-3. As a result, the following guidelines can be derived from this chapter:

- Speech Tactons improve warnings in all measures used, so they are suggested as an addition to speech warnings. The use of Speech Tactons along with audio, results to cues with Perceived Urgency and Perceived Alerting Effectiveness, that increase for cues of higher designed urgency, while overall Perceived Annoyance remains low;
- Annoyance is higher but more acceptable for high urgency warnings compared to lower urgency ones. This is because warnings of high designed urgency are perceived as more effective and more annoying, while warnings of low designed urgency are perceived as less effective and more annoying. In all cases though, annoyance is kept at low levels with the cues used;
- Speech Tactons can be recognised acceptably in terms of their urgency even if they are presented alone. It is suggested that the number of cues used to signify urgency with Speech Tactons is no more than one per level of designed urgency, to avoid confusion between cues.

The next set of experiments (Experiments 6 and 7) combined the work done in Experiments 1 – 5, by performing a comparison between Abstract and Language-based multimodal warnings. Having effectively designed and evaluated a set of Language-based warnings in Experiments 4 and 5, a next logical step was to directly compare them with Abstract cues, designed and evaluated in Experiments 1, 2 and 3. This would provide direct results on their relative effectiveness, allowing for guidelines on advantages and disadvantages of each cue design. Experiments 4 and 5 did not use a simulated driving task, nor did they utilise Language-based warnings in the visual modality. This would be beneficial, since a more dynamic character of the experimental task would be more closely related to a real driving task, while the use of visuals would need to be considered in order to design truly multimodal cues. Therefore, the above limitations would be addressed in Experiments 6 and 7, providing a set of truly multimodal cues along the audio, visual and tactile modalities, with varying message contents. Further, the driving tasks used to assess the cue effectiveness would combine recognition (as used in Experiment 2) and reaction (as used in Experiment 3), so as to present complete results of Abstract versus Language-based multimodal cues, varying in urgency and message content, evaluated in simulated driving task with different levels of criticality (*i.e.* a non-critical recognition task, and a critical reaction task).

7. Comparing Abstract and Language-based Multimodal Driver Displays

7.1. Introduction

This chapter completes the answer to the research question: *How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?* Having designed and evaluated a set of language-based multimodal displays varying in urgency in Chapter 6, a comparison of these displays with abstract displays is performed. Multimodal abstract displays have already been designed and evaluated in Chapters 4 and 5, but never compared exhaustively to language-based ones. As also described in Section 2.4, previous experiments have evaluated the performance of abstract versus more informative audio (J. Edworthy, Walters, et al., 2000; Cristy Ho & Spence, 2005; Denis McKeown & Isherwood, 2007) or tactile cues (Rob Gray et al., 2014). However, no research has studied all multimodal combinations of these warnings and how their simultaneous presentation affects responses. This is important in order to provide guidelines on the effectiveness of these two types of messages and the best modalities to utilize for message display.

Speech Tactons, designed in Chapter 6, presented promising results when combined with speech warnings. However, they were not tested in the driving context, which may affect performance. In order to provide complete guidelines for Speech Tactons, their evaluation using a driving task is essential. This is why Experiment 6 provided an evaluation comparing abstract and language-based displays in a non-critical recognition task (as performed in Experiment 2 for abstract displays only), while Experiment 7 provided an evaluation comparing these display in a critical reaction ask (as performed in Experiment 3 for abstract displays only). In this way, the two types of cues were exhaustively evaluated using objective measures, which can provide insights on their utility as driver alerts.

This chapter presents a first comparison between abstract and language-based warnings, across all combinations of audio, visual and tactile modalities. Speech, text and Speech Tactons were compared to abstract pulses in Experiments 6 and 7. All warnings were evaluated in terms of recognition time of cue urgency without any critical event present, in line with Experiment 2 and (Cao, van der Sluis, et al., 2010), and response time to high urgency cues during a critical event to assess the resulting behaviour, in line with Experiment 3 and studies like (Cristy Ho, Tan, et al., 2005; Scott & Gray, 2008). In this way, the cues would be evaluated in the context of both critical and less critical situations, allowing generalisation of the results in a wider set of contexts.

Section 7.2 describes the cue design used in Experiments 6 and 7. Section 7.3 presents Experiment 6 evaluating the recognition time of abstract and language-based multimodal warnings in the absence of a critical event. Section 7.4 presents Experiment 7, evaluating reaction times of these warnings with a critical event present. Section 7.5 discusses the findings, and Section 7.6 presents the derived conclusions and guidelines from this chapter.

7.2. Warning Design

To compare responses to abstract versus language-based warnings, cues from Experiments 1, 2 and 4, 5 were used, utilizing respectively repeated pulses and language-based messages. As in these experiments, cues were presented in all combinations of the audio, visual and tactile modalities: Audio (A), Visual (V), Tactile (T), Audio + Visual (AV), Audio + Tactile (AT), Tactile + Visual (TV), Audio + Tactile + Visual (ATV).

7.2.1. Abstract Warnings

The abstract warnings consisted of repeated tones and were similar to Experiments 1 and 2. As in these experiments, three Levels of Designed Urgency (LDU) were created, indicating conditions varying in importance. L_H (Level High) signified situations of high urgency, such as an impending collision, L_M (Level Medium) situations of medium urgency, such as a broken headlamp and L_L (Level Low) situations of low urgency, such as an advertisement. There were 21 signals: 7 signals with the above modalities (A, T, V, AT, AV, TV, ATV) × 3 Levels of Designed Urgency. The warnings consisted of pure tones, colours or vibrations delivered as repeated pulses. Pulse rate increased as signals became more urgent, as in (B. A. Lewis & Baldwin, 2012). Warnings of the same urgency level had the same pulse rate, independent of modality. 8 pulses having 0.1 *sec* single pulse duration and interpulse interval were used for L_H, 5 pulses having 0.17 *sec* single pulse duration and interpulse interval for L_M and 2 pulses having 0.5 *sec* single pulse duration and 0.5 *sec* interpulse interval for L_L. All warnings had 1.5 *sec* duration. Auditory warnings were varied additionally in base frequency, as in (Judy Edworthy et al., 1991b; B. A. Lewis & Baldwin, 2012; D. C. Marshall et al., 2007) (1000 Hz for L_H, 700 Hz for L_M and 400 Hz for L_L).Visual warnings were also varied in colour, in line with (B. A. Lewis & Baldwin, 2012) (Red for L_H, Orange for L_M and Yellow for L_L [13]). A C2 Tactor from Engineering Acoustics[14] was used for the tactile warnings. Tactile warnings had a frequency of 250 Hz, the nominal centre frequency of the C2. The above warnings showed significantly different ratings of perceived urgency in Experiments 1 and 2 and were selected as good candidate abstract signals to convey differently urgent events multimodally.

Contrary to the fixed intensity of the cues in Experiments 1 and 2, the intensity of audio and tactile cues was decreased as their designed urgency decreased for three reasons. Firstly, annoyance levels in Experiment 1 were higher in the tactile modality, an effect that was ameliorated by varying intensity as urgency decreased in Experiment 4. Secondly, the good recognition results achieved for language-based cues in Experiment 5 provided a good potential for a similar result for the abstract tones. Finally, it was desirable to have a fair comparison between abstract and language-based warnings and avoid any observed effects to be accounted for by different intensities. Therefore, the intensity of speech cues of the same urgency level, which will be described below, was also used in the abstract cues. Thus, in both audio and tactile cues, L_H messages had a peak of -1.9 *dBFS*, L_M had a peak of -11.1 *dBFS* and L_L had a peak of -16.5 *dBFS*. Simultaneous delivery of unimodal signals was used in the multimodal ones, to create a synchronous effect of sound, vibration, visuals and all their combinations.

7.2.2. Language-based Warnings

The language-based warnings used were the best performing cues in terms of recognition accuracy from Experiment 5. Three speech messages designed to convey three different urgency levels, L_H, L_M and L_L were used: *"Danger! Collision Imminent"* for L_H, *"Warning! Left side headlamp out"* for L_M and *"Notice! Call and win free tickets"* for L_L. As described in Chapter 6, all messages were recorded by a female voice actor, who in line with (Elizabeth Hellier et al., 2002) was instructed to narrate the message of L_H in an urgent manner, as if a loved one was in imminent danger. The L_M message was spoken non-urgently, as if in a friendly conversation with nothing interesting about the situation and the L_L message was

[13] Red was *RGB*(255,0,0), Orange was *RGB*(255,127,0) and Yellow was *RGB*(255,255,0).
[14] http://www.atactech.com/PR_tactors.html

spoken in a monotone, deadpan manner. The L_H message was 1.7 *sec* long and had a peak of -1.9 *dBFS* and an average frequency of 377 *Hz*. The L_M message was 2.7 *sec* long, had a peak of -11.1 *dBFS* and an average frequency of 285 *Hz*. Finally, the L_L message was 3.7 *sec* long, had a peak of -16.5 *dBFS* and an average frequency of 202 *Hz*.

For the Speech Tactons, all stimuli designed were auditory, to be used with a C2 tactor. To construct the auditory cues, the fundamental frequency F_0 (pitch) of each sample of the speech recordings was obtained, which resulted in alternating pure tones for each utterance. Then, the changes in intensity of the original sound files were used in the tones. All tactile cues retained the rhythm and intensity variations of the original recordings. The resulting values of average frequency of all tactile cues never differed to the average frequency of the audio more than ±10Hz.

Finally, for the visual cues, the text of the warnings was displayed in the same colour as the abstract cues of the respective LDU (Red for L_H, Orange for L_M and Yellow for L_L). A possible limitation of this approach is that the effects of text meaning and text colour were not measured separately. However, a consistent colouring between abstract and language based visual cues was maintained for simplicity, in line with (B. A. Lewis & Baldwin, 2012). Thus, 21 different language-based cues were created, 7 cues with all modalities (A, T, V, AT, AV, TV, ATV) × 3 Levels of Designed Urgency. For all modifications, Praat[15] and Audacity[16] software were used. In all, there were 42 different warnings, 21 abstract and 21 language-based ones[17]. These warnings were evaluated in two experiments, looking into how quickly and accurately participants would respond when exposed to them.

7.3. Experiment 6: Comparing Abstract and Language-based Cues in terms of Recognition Time

7.3.1. Motivation

As found in Experiment 2, the use of abstract cues to design multimodal warnings of varying urgency can be effective, since the level of urgency can be recognized in a time-based task,

[15] http://www.fon.hum.uva.nl/praat/
[16] http://audacity.sourceforge.net/
[17] All warnings are available at http://goo.gl/XHViGY

resulting to participant responses that vary according to the urgency designed in the cues. Experiment 2 investigated a recognition task for the cues, looking into how quickly and accurately participants would be able to recognize the warnings' level of designed urgency. However, all the warnings used in Experiment 2 were abstract, not addressing the possibility of using warnings that are more semantically associated to the signified event. This investigation could be beneficial, since it would reveal the relative advantages of warnings that vary in design (abstract and language-based) providing guidelines on how they can be used multimodally as driving alerts, depending on the demands of the situation. Communicating the content of the signified event through language may be beneficial for example when the event is not obvious, and conversely when the event is obvious and imminent, an abstract alert may be more suitable. Previous studies such as (J. Edworthy, Walters, et al., 2000) did not discover significant differences between speech and non-speech cues in terms of subjective measures. However, the cues investigated were not multimodal, and the task used was not time-based, requiring no immediate recognition in the cues. This was addressed with Experiments 6 and 7, starting with a recognition task in the absence of a critical event (Experiment 6), and continuing with a reaction task where a critical event increased the urgency of the simulated situations (Experiment 7).

7.3.2. Design

Experiment 6 evaluated how quickly and accurately participants were able to recognize the level of urgency of the presented multimodal warnings. A 7×3×2 within subjects design was used with Modality, Level of Designed Urgency (LDU) and Information as the independent variables and Recognition Time (RecT) and Recognition Accuracy (RecA) as the dependent variables. Modality had 7 levels (A, T, V, AT, AV, TV, ATV), LDU had 3 levels (L_H, L_M, L_L) and Information had 2 levels (Abstract, Language-based). In line with (Carryl L Baldwin & May, 2014; Cristy Ho & Spence, 2005), the reaction time performance of speech cues was expected to be close to non-speech cues. Further, lack of difference in perceived urgency observed in (J. Edworthy, Walters, et al., 2000) would be investigated further using a simulated driving task. It was however expected, that abstract cues and cues of higher LDU would have an advantage in all measures, due to their simplicity and shorter length. Finally, the advantage of using multimodal warnings observed in all previous experiments would be evaluated when both abstract and language-based cues would be utilized. As a result, there were the following hypotheses:

- RecT will be influenced by Modality (H_{1a}), LDU (H_{1b}) and Information (H_{1c});
  - Specifically, RecT was expected to decrease in multimodal as opposed to unimodal cues and in higher levels of LDU, while abstract warnings were expected to create quicker responses compared to language-based warnings.
- RecA will be influenced by Modality (H_{2a}), LDU (H_{2b}) and Information (H_{2c}).
  - Specifically, RecA was expected to increase in multimodal as opposed to unimodal cues and in higher levels of LDU, while abstract warnings were expected to create more accurate responses compared to language-based warnings.

7.3.3. Procedure

Twenty participants (10 female) aged between 20 and 38 years (M = 25.05, SD = 5.11) took part in this experiment. They had not participated in previous experiments. They all held a valid driving license and had between 1 and 20 years of driving experience (M = 6.05, SD = 5.23). There were two left handed participants and all reported normal hearing and normal or corrected to normal vision. They were either University students or employees.

The experiment took place in a University room, where participants sat in front of 27-inch Dell 2709W monitor and a PC running the simulator software (see Figure 7-1). In the software, a three lane road in a rural area was depicted, with a lead car maintaining a steady speed in the central lane, as in Experiments 1, 2 and 3. Participants used a Logitech G27 gaming wheel and pedals to steer the simulated vehicle and to brake. Inputs were logged with a frequency of 50 *Hz*. Participants wore a set of Sennheiser HD 25-1 headphones and a wristband on their left wrist with a C2 Tactor attached on the inside of the band, in line with Experiment 4. This simulated tactile feedback being presented by a smart watch. To cover any noise from the Tactor, car sound was played throughout the experiment. For two participants, sound and vibration were slightly adjusted to maintain comfortable intensities. Visual abstract cues used delivered through coloured circles that flashed in the top central area of the screen, and were sized 400×400 pixels (about 12×12 *cm*). Visual language-based cues were coloured text displaying each warning, which appeared once and for as long as the warning was uttered in the top central area of the screen, and were sized 200×800 pixels (about 24×6 *cm*). The visual cues did not obstruct the lead car and were designed to simulate a Head-Up Display. Abstract and language-based visual cues were also designed so as to occupy roughly the same area on the screen (about 144 cm^2). Figure 7-1 shows the experimental setup and visual cues.

Figure 7-1: The experimental setup. The visual signals are Abstract L_H (a), L_M (b) and L_L (c) and Language-based L_H (d), L_M (e) and L_L (f).

Participants were welcomed and provided with a brief introduction to the experiment. Afterwards, they were exposed to all the warnings as follows: first, a label with the text *"Level High (H) Warnings of HIGH urgency will follow"* appeared on the screen, then the 7 abstract warnings of L_H were played once to half of the participants, in the following order: A → T → V → AT → AV → TV → ATV and then the 7 language-based warnings of L_H were played in the same order of modality. Afterwards, a label with the text *"Level Medium (M) Warnings of MEDIUM urgency will follow"* appeared and then the 7 abstract warnings of L_M were displayed followed by the 7 language-based L_M ones, keeping the same order for modalities. Finally, a label with the text *"Level Low (L) Warnings of LOW urgency will follow"* appeared, followed by the 7 L_L abstract and then the 7 L_L language-based warnings as above. To the other half of the participants, first the language-based cues were played in each LDU and then the abstract ones, in the same manner as above. This procedure was chosen to minimize any order effects when presenting abstract and language-based cues, while still presenting them in a memorable way. The training lasted about 6 *min* for each participant. They were then asked to drive for 90 *sec*, to get accustomed to the simulator.

In the main part of the study, participants were presented with a driving scene, where they drove a simulated vehicle along a straight rural road following a car in front. Participants were able to steer the vehicle but did not use the accelerator pedal. The vehicle controlled by the participants maintained a constant speed of just above 70 *mph*. This speed was chosen in order to exceed the UK motorway speed limit (70 *mph*) creating a hazardous driving situation and requiring the drivers' attention. While steering the vehicle, the warnings were displayed to the participants in a random order and with a random interval of any integral value between (and including) 11–19 *sec*. These values were chosen to be similar to previous driving studies with repeated exposure to stimuli, *e.g.* (Cristy Ho & Spence, 2005), as well as previous experiments in this thesis. Each stimulus was played twice. This resulted in a total of 82 stimuli (42 warnings × 2 presentations). Participants were asked to identify the urgency level of each stimulus by pressing one of three buttons on the steering wheel as quickly as possible. Buttons were labelled with letters (H, M or L) according to the urgency levels – topmost for L_H, middle for L_M, bottom for L_L. Participants were asked to maintain a central lane position. The whole experiment lasted about 30 minutes and participants were then prepared for the next experiment, which followed immediately.

7.3.4. Results

7.3.4.1. Recognition Time

Data of two participants were discarded. In the first case, this was because the participant mentioned after the experiment that they were giving their opinion on how urgent were the cues by pressing the buttons, rather than identifying the given urgency of the cues. In the second case, it was because the participant was visibly inattentive to the experiment. All remaining data for recognition time were analysed using a three-way repeated measures ANOVA, with Modality, Level and Information as factors. Mauchly's test showed that the assumption of sphericity had been violated for Modality and the interaction between Modality and LDU and Modality and Information. Therefore, Degrees of Freedom were corrected with Greenhouse–Geisser sphericity estimates.

Hypothesis H_{1a}: There was a significant main effect of Modality ($F(3.01,102.33) = 103.13$, $p < 0.001$). Contrasts revealed that AV, ATV, V and TV warnings elicited significantly quicker responses compared to A and AT ($F(1,34) = 22.77$, $r = 0.59$, $p < 0.001$), which in turn had quicker responses compared to T ($F(1,34) = 106.78$, $r = 0.87$, $p < 0.001$). ***Hypothesis H_{1b}:*** There was a significant main effect of LDU ($F(2,68) = 74.13$, $p < 0.001$). Contrasts revealed that L_H warnings were recognised quicker compared to the L_M and L_L ones ($F(1,34) = 89.05$, $r = 0.85$, $p < 0.001$). ***Hypothesis H_{1c}:*** There was a significant main effect of Information ($F(1,34) = 37.55$, $p < 0.001$). Contrasts revealed that Abstract warnings were recognised quicker than Language-based ones (1.41 *sec* on average for Abstract vs. 1.57 *sec* for Language-based warnings, $F(1,34) = 37.55$, $r = 0.72$, $p < 0.001$). As a result, hypotheses H_{1a}, H_{1b} and H_{1c} were accepted. See Figure 7-2 for Recognition Times across Modalities, Figure 7-3 for RecT across LDU and Table 7-1 for pairwise comparisons of RecT across Modalities.

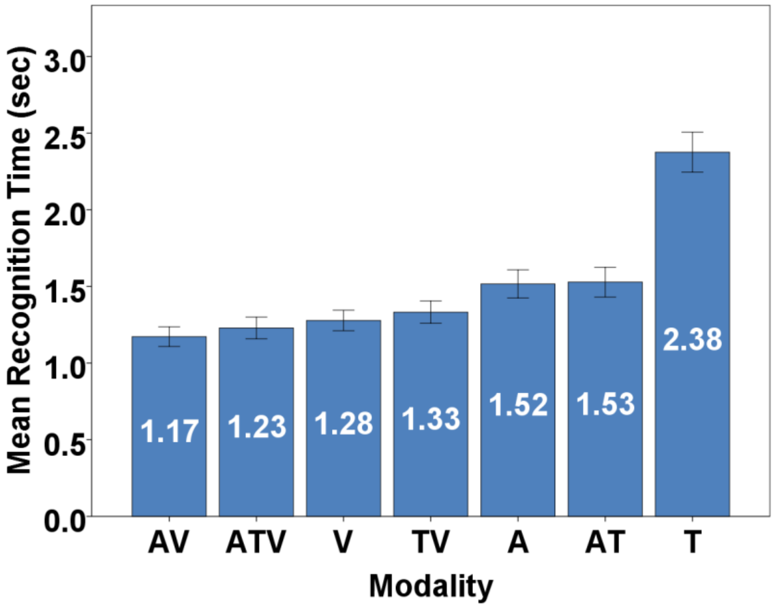


Figure 7-2: Recognition times for Experiment 6 across Modalities (hypothesis H_{1a}). Modalities in graphs are sorted by mean values. Error bars represent 95% Confidence Intervals.

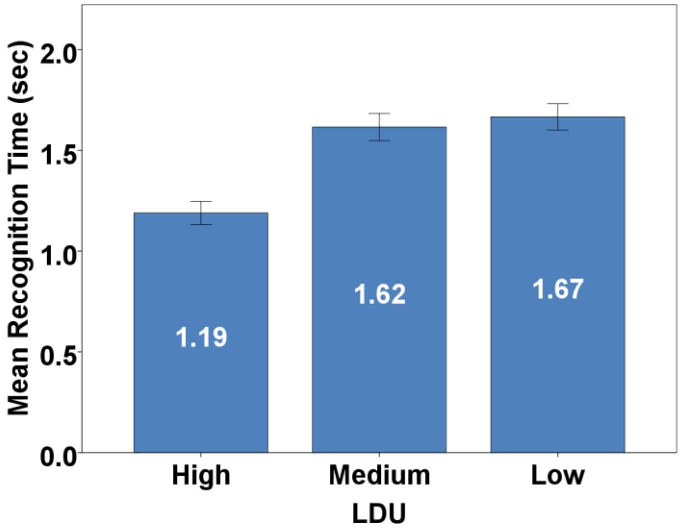


Figure 7-3: Recognition times for Experiment 6 across Levels of Designed Urgency (hypothesis H_{1b}).

	AV	ATV	V	TV	A	AT	T
AV		.185	**.000**	**.001**	**.000**	**.000**	**.000**
ATV	.185		.297	.067	**.000**	**.000**	**.000**
V	**.000**	.297		.166	**.000**	**.000**	**.000**
TV	**.001**	.067	.166		**.000**	**.000**	**.000**
A	**.000**	**.000**	**.000**	**.000**		.860	**.000**
AT	**.000**	**.000**	**.000**	**.000**	.860		**.000**
T	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 7-1: Pairwise comparisons between modalities for Recognition Time (H_{1a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Modality and LDU ($F(7.28,247.60) = 2.63$, $p < 0.05$, see Figure 7.4). Contrasts revealed that A warnings had slower recognition times than TV for L_M compared to L_H ($F(1,34) = 14.55$, $r = 0.55$, $p < 0.05$). Although TV warnings with L_M cues were quicker than L_L in recognition, this effect was reversed for A ($F(1,34) = 4.73$, $r = 0.35$, $p < 0.05$). Finally, that AT warnings had quicker recognition times than T for L_H compared to L_M ($F(1,34) = 6.04$, $r = 0.39$, $p < 0.05$). There was a significant interaction between Modality and Information ($F(3.68,125.07) = 26.82$, $p < 0.001$, see Figure 7.5). Contrasts revealed that for TV warnings Abstract signals were recognised quicker than Language-based ones, but this was reversed for A warnings ($F(1,34) = 11.87$, $r = 0.51$, $p < 0.05$). Further, that while for T warnings Abstract signals were recognised quicker than Language-based ones, this was reversed for AT ($F(1,34) = 52.37$, $r = 0.78$, $p < 0.05$).

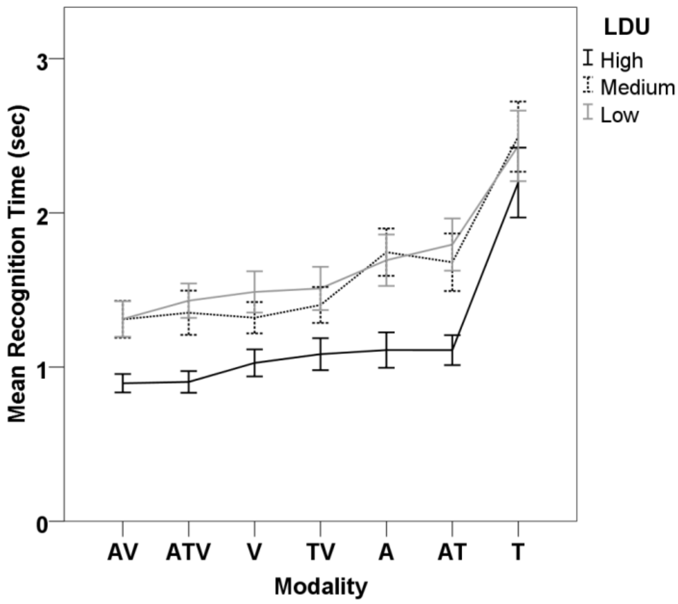


Figure 7-4: The interaction between Modality and LDU for Recognition Time (H_{1a} – H_{1b}).

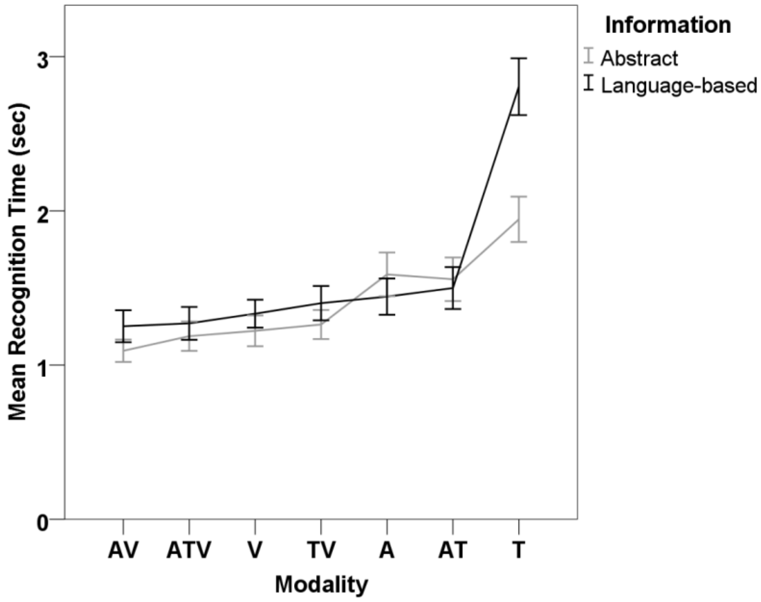


Figure 7-5: The interaction between Modality and Information for Recognition Time (H_{1a} – H_{1c}).

7.3.4.2. Recognition Accuracy

In all, there were 1512 participant responses and only 1 trial where a participant failed to respond. For the rest, 1366 responses were correct (90.4%) and 145 incorrect (9.6%). Data for recognition accuracy were treated as dichotomous (with values "correct" or "incorrect") and analysed with Cochran's Q tests. ***Hypothesis H_{2a}:*** Cochran's Q tests revealed that participants made significantly more mistakes in modality T compared to all the rest of the modalities. Specifically, out of 228 trials for each modality, there were 80 mistakes for T versus 16 for A ($Q(1) = 47.63$, $p < 0.001$), 17 for V ($Q(1) = 46.69$, $p < 0.001$), 18 for AT ($Q(1) = 46.89$, $p < 0.001$), 9 for AV ($Q(1) = 63.81$, $p < 0.001$), 12 for TV ($Q(1) = 57.80$, $p < 0.001$) and 13 for ATV ($Q(1) = 54.08$, $p < 0.001$). ***Hypothesis H_{2b}:*** Cochran's Q tests also revealed that there were significantly more mistakes in L_L compared to L_H and L_M. Specifically, out of 532 trials for each level, there were 80 mistakes for L_L versus 38 for L_H ($Q(1) = 16.04$, $p < 0.001$) and 47 for L_M ($Q(1) = 10.78$, $p < 0.001$). ***Hypothesis H_{2c}:*** There was no significant difference in number of mistakes between Abstract and Language-based cues. Specifically, out of 798 trials for each type of Information, there were 88 mistakes for Abstract versus 77 for Language-based cues, $Q(1) = 0.83$, $p = 0.36$. As a result, hypotheses H_{2a} and H_{2b} were accepted and H_{2c} was rejected.

7.4. Experiment 7: Comparing Abstract and Language-based Cues in terms of Reaction Time

7.4.1. Motivation

Experiment 7 evaluated the speed and accuracy of responses to abstract versus language-based multimodal displays, simulating a reaction task to a critical event, *i.e.* a car braking in front. This task was also used in Experiment 3, and showed quicker responses to the critical event in the presence of critical alerts. In Experiment 7, this finding was used and warnings of high designed urgency were only presented when the critical event occurred. Warnings of medium and low urgency were therefore ignored by the participants, simulating more convincingly the reaction during a real driving task. In this way, results of Experiment 6 looking into a recognition task, and subjective results like the ones presented in (J. Edworthy, Walters, et al., 2000) would be elaborated for the context of a critical task, completing the comparison between abstract and language-based warnings of varying urgency. In this way, the guidelines for the utility of both designs would be informed by performance observed in both a non-critical (Experiment 6) and a critical task (Experiment 7).

7.4.2. Design

Experiment 7 evaluated how quickly participants were able to respond to presented multimodal warnings of high urgency (L_H). Other than their response time to this task, two driving metrics suggested in studies such as (D. P. Brumby et al., 2011) and also used in Experiment 3 were used. These were the Root Mean Square Error (RMSE) of the vehicle's lateral deviation and steering angle. As mentioned in Section 3.2, lower lateral deviation and steering angle can indicate lower driver distraction, see also (Lindgren et al., 2009; Y. C. Liu, 2001). Further, the variable Time was used to measure the effect of the warning presentation on participants' driving behaviour by comparing the metrics described before and after the warning presentation.

A 7×3×2×2 within subjects design was used with Modality, LDU, Information and Time as the independent variables. Response Time (ResT), Lateral Deviation (LatDev) and Steering Angle (SteAng) were the dependent variables. As in the previous experiment, Modality had 7 levels (A, T, V, AT, AV, TV, ATV), LDU had three levels (L_H, L_M, L_L) and Information had 2 levels (Abstract, Language-based). Finally, Time had 2 levels: Before cue was presented and After cue was presented. As in Experiment 3 and (D. P. Brumby et al., 2011; John D. Lee et al., 2004), response times and lane keeping behaviour were expected to be affected by the warnings, since they would pose an additional load on the driving task. As in Experiment 6, multimodal cues, cues of higher LDU and abstract cues were expected to perform best in the time-based task used (ResT). However, they were also expected to disrupt the driving metrics the most, due to their increased saliency. As a result, there were the following hypotheses:

- ResT when reacting to L_H warnings and a car braking event will be influenced by Modality (H_{3a}) and Information (H_{3b});
    - Specifically, ResT was expected to decrease when multimodal as opposed to unimodal cues were used as warnings, while abstract warnings were expected to create quicker responses compared to language-based warnings.
- LatDev when reacting to L_H warnings and a car braking event will be influenced by Modality (H_{4a}), Information (H_{4b}) and Time (H_{4c});
    - Specifically, LatDev was expected to increase when multimodal as opposed to unimodal cues were used as warnings, while abstract warnings were expected to create higher LatDev compared to language-based warnings.
- SteAng when reacting to L_H warnings and a car braking event will be influenced by Modality (H_{5a}), Information (H_{5b}) and Time (H_{5c});
    - Specifically, SteAng was expected to increase when multimodal as opposed to unimodal cues were used as warnings, while abstract warnings were expected to create higher SteAng compared to language-based warnings.
- LatDev when exposed to L_M and L_L warnings without a car braking event will be influenced by LDU (H_{6a}), Modality (H_{6b}), Information (H_{6c}) and Time (H_{6d});
    - Specifically, LatDev was expected to increase with L_M warnings and when multimodal as opposed to unimodal cues were used as warnings, while abstract warnings were expected to create higher LatDev compared to language-based warnings.
- SteAng when exposed to L_M and L_L warnings without a car braking event will be influenced by LDU (H_{7a}), Modality (H_{7b}), Information (H_{7c}) and Time (H_{7d}).
    - Specifically, SteAng was expected to increase with L_M warnings and when multimodal as opposed to unimodal cues were used as warnings, while abstract

warnings were expected to create higher SteAng compared to language-based warnings.

7.4.3. Procedure

Participants and equipment were identical to the previous experiment. It took place after participants completed the previous experiment and had a short break. Participants were presented with the same driving scene showing a vehicle maintaining a constant speed of just above 70 *mph*. Other than steering the vehicle, participants were able to respond by pressing the brake pedal. While steering the vehicle, the warnings were again displayed to the participants in a random order and with a random interval of any integral value between (and including) 11–19 *sec*. Each stimulus was again played twice, resulting in a total of 82 stimuli (42 warnings × 2 presentations). When there was a L_H warning, the vehicle in front started braking towards the participant vehicle along with the presentation of the warning. In case of an L_M or L_L warning, the vehicle in front continued driving and did not brake.

Participants were asked to maintain a central lane position throughout the experiment. They were instructed to respond by pressing the brake pedal as quickly as possible when there was a L_H warning presented along with the car in front braking. Finally, they were instructed to ignore the L_M and L_L warnings and not to respond to them. This process was chosen because responding to some warnings and ignoring others would create an increased workload for participants, requiring higher attention. As shown in Experiment 3, the presentation of warnings along with a critical event resulted in quicker responses, which is desired in this situation. Finally, testing responses to L_H cues was considered as more ecologically valid, since participants would not have to respond promptly to L_M or L_L warnings in a real setting.

Participants' ResT was calculated from the onset of the L_H stimulus and start of the braking event of the lead car, until the participant first pressed the brake pedal. Their LatDev and SteAng were logged for 4.7 *sec* (from 5.7 *sec* to 1 *sec* before any stimulus was displayed), forming their baseline value for driving performance. They were logged again for 4.7 *sec* immediately after the stimulus to assess the warning effects on driving. The value of 4.7 *sec* was chosen, since it was the duration of the longest of all cues (3.7 *sec*), increased by 1 *sec*. Thus, any effects occurring throughout the longest possible duration of a cue plus a small period of time afterwards would be recorded.

For both LatDev and SteAng, the RMSE values were then computed from the logged values. As a result, out of the 82 overall trials, there were 28 values of ResT ([7 Abstract L_H cues + 7 Language-based L_H cues] × 2 presentations). Also, since LatDev and SteAng were logged in all cases (L_H, L_M and L_L cues), for each of the 82 trials there were two values for their LatDev (baseline value and value after the cue was displayed) and two values for their SteAng (baseline value and value after the cue was displayed). The whole experiment lasted about 30 *min* and participants were then debriefed about the purpose of both experiments and paid £6 for their participation.

7.4.4. Results

7.4.4.1. Response Time

The data for one participant were discarded, since they were resting their foot on the brake pedal contrary to the instructions. The remaining data for response times to L_H cues were analysed using a two-way repeated measures ANOVA, with Modality and Information as factors. Mauchly's test showed that the assumption of sphericity had been violated for Modality, therefore Degrees of Freedom were corrected with Greenhouse–Geisser sphericity estimates. ***Hypothesis H_{3a}:*** There was a significant main effect of Modality ($F(4.17,154.37) = 18.83$, $p < 0.001$). Contrasts revealed that ATV, AV, AT and A warnings caused quicker responses compared to TV and T ones ($F(1,37) = 7.46$, $r = 0.41$, $p < 0.05$), which in turn had quicker responses compared to V warnings ($F(1,37) = 6.92$, $r = 0.40$, $p < 0.05$). ***Hypothesis H_{3b}:*** There was no significant effect of Information ($F(1,37) = 1.37$, $p = 0.25$). As a result, H_{3a} was accepted and H_{3b} was rejected. See Figure 7-6 for Response Times across Modalities and Table 7-2 for pairwise comparisons of ResT across Modalities.

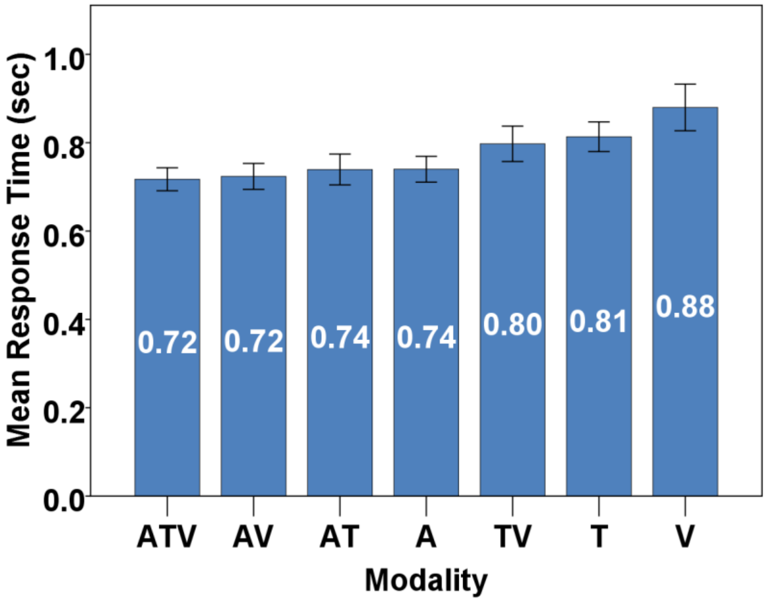


Figure 7-6: Response times for Experiment 7 across Modalities (hypothesis H_{3a}).

	ATV	AV	AT	A	TV	T	V
ATV		.753	.194	.098	**.000**	**.000**	**.000**
AV	.753		.216	.172	**.000**	**.000**	**.000**
AT	.194	.216		.989	**.024**	**.001**	**.000**
A	.098	.172	.989		**.010**	**.000**	**.000**
TV	**.000**	**.000**	**.024**	**.010**		.369	**.001**
T	**.000**	**.000**	**.001**	**.000**	.369		**.012**
V	**.000**	**.000**	**.000**	**.000**	**.001**	**.012**	

Table 7-2: Pairwise comparisons between modalities for Response Time (H_{3a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

7.4.4.2. Lateral Deviation and Steering Angle

Data for LatDev when reacting to L_H cues by braking were analysed using a three-way repeated measures ANOVA, with Modality, Information and Time as factors. Time had two levels: Before cue was displayed (baseline data) and After cue was displayed. ***Hypothesis H_{4a}:*** No significant effects of Modality were observed. ***Hypothesis H_{4b}:*** There was an effect of Information which approached significance ($F(1,39) = 4.00$, $p = 0.053$), suggesting that the average LatDev both before and after the exposure to Language-based cues may be lower compared to Abstract cues, but not significantly so. ***Hypothesis H_{4c}:*** There was a significant main effect of Time ($F(1,39) = 5.65$, $p < 0.05$). Contrasts revealed that LatDev was higher after the cues were displayed (0.47 *m* on average before the L_H cue and the car braking event vs. 0.51 *m* after the cue and the event, $F(1,39) = 5.65$, $r = 0.36$, $p < 0.05$).

Data for SteAng when reacting to L_H cues by braking were also analysed using a three-way repeated measures ANOVA, with Modality, Information and Time as factors. As above, Time had two levels: Before cue was displayed (baseline data) and After cue was displayed. ***Hypotheses H_{5a}, H_{5b}:*** There were no other significant effects observed for Modality and Information. ***Hypothesis H_{5c}:*** There was a significant main effect of Time ($F(1,39) = 26.63$, $p < 0.001$). Contrasts revealed that SteAng was higher after the cues were displayed (0.07 *rad* on average before the L_H cue and the car braking event vs. 0.08 *rad* after the cue and the event, $F(1,39) = 26.63$, $r = 0.64$, $p < 0.001$). As a result of the above, H_{4c} and H_{5c} were accepted and H_{4a}, H_{4b}, H_{5a} and H_{5b} were rejected.

Data for LatDev and SteAng when exposed to L_M and L_L cues without reacting were analysed using a four-way repeated measures ANOVA, with Modality, Level, Information and Time as factors. ***Hypotheses H_{6a} - H_{6d} and H_{7a} - H_{7d}:*** There were no significant effects observed for LDU, Modality, Information or Time for neither LatDev nor SteAng. As a result, H_{6a} - H_{6d} and H_{7a} - H_{7d} were all rejected.

7.5. Discussion

7.5.1. Recognition Time and Accuracy

Looking at the results of Experiment 6, results for recognition time showed an advantage of abstract cues when identifying the level of designed urgency (H_{1c} was accepted), although they did not present any differences in recognition accuracy (H_{2c} was rejected). This can be partly explained by the fact that these cues were shorter in length overall. In studies like (Cao, Mahr, et al., 2010), speech cues caused longer response times compared to abstract ones. It is noted, however, that the task of Experiment 6 was an identification one, requiring recall of the cues' LDU. As will be discussed later, this is arguably different to a simple response task, in which abstract and language-based cues performed similarly in this study. As a guideline, abstract multimodal cues can be used for quicker identification compared to language-based ones in a non-critical task. Although the speech cues were designed so that the distinctive word describing their LDU (*Danger*, *Warning* or *Notice*) came first in the message, it seemed that the cue length still required more time to interpret compared to the short abstract pulses. Interestingly, language-based warnings performed better in terms of RecT compared to abstract ones in modalities A and AT (interaction between Modality and Information, H_{1a} – H_{1c}), indicating that when there is sound conveying the information, speech can also be a good means to do so. However, the above guideline still holds, since the main effect showed better performance of abstract cues overall. In future work, even simpler language-based cues could be evaluated to compare their recognition time with abstract ones.

In terms of modalities (H_{1a}), results of RecT are very similar to Experiment 2, where only abstract cues were used. As in Experiment 2, the visual modality seems to have played an important role in participants recognizing the cues' LDU, since the modalities with the shortest recognition times all included visuals. Other than V, all other better performing cues were multimodal, similar to the previous experiments of this thesis, where the presence of more modalities enhanced responses in all measures. The presence of V in the group of best performing modalities confirms the role of visuals when interpreting such messages. However, V alone cannot arguably be recommended for critical situations, as it has been shown that it suffers in terms of performance when a visual critical event occurs (see Experiment 3). Also, as will be discussed later, V cues performed worse when users were reacting to an imminent collision event. As a guideline, abstract cues including visuals can be used to effectively inform about non-critical driving events. Combining results of perceived annoyance in Experiment 1, which increases as modalities used increase, bimodal rather than trimodal cues for this case can be recommended.

A disadvantage of unimodal tactile cues in terms of recognition time was found. As also anecdotally mentioned by several participants in unplanned and unstructured discussions after the end of both experiments, cues were harder to identify by some, when not accompanied by sound and / or visuals clarifying their meaning. This disadvantage of unimodal tactile cues was also found in Experiment 2, where only abstract cues were evaluated, indicating that this difficulty holds also when language-based cues are used. The results of recognition accuracy, where the T cues were the worst performing compared to all other modalities (H_{2a} was rejected), also can add to this observed disadvantage. Thus, it is recommended to avoid the delivery of messages through vibration alone when recognition time is important, since this may slow down their interpretation.

In terms of LDU, recognition of L_H cues was quickest, confirming that they were conveying an increased level of urgency (H_{1b} was accepted). This was again in line with Experiment 2 and showed that the design used was effective in conveying high urgency in both abstract and language-based warnings. Combined with the recognition accuracy results, where L_L

cues performed the poorest (H_{2b} was accepted), it can be concluded that the cues designed can afford quick recognition in more urgent situations. In terms of low urgency situations, it is concluded that cues should be used cautiously, since a driver response is arguably not essential in such cases (*e.g.* presenting an advertisement) and combined with the low performance observed, it may also be disruptive. Finally, the difference in RecT between L_M and L_L was less pronounced for unimodal cues (A and T) compared to multimodal (AT and VT), arguably indicating how the added salience of multimodal cues can accelerate the cue recognition (interaction between Modality and LDU, H_{1a} – H_{1b}).

7.5.2. Response Time, Lateral Deviation and Steering Angle

Results for response time in Experiment 7 were similar to Experiment 3, with warnings including audio creating quickest responses compared to the rest. With the exception of A, all warnings in the quickest performing group were multimodal (H_{3a} was accepted). This is in line with Experiment 3, although in that study all best performing warnings were multimodal. This enhanced performance of warnings including audio could be arguably attributed to the reliance on cues different than visual for reacting to a visual critical event. As shown in Experiment 3, visual cues can suffer in terms of response times when users were exposed to a critical event in the simulator. This was confirmed in Experiment 7, since unimodal visual cues performed the poorest and only when accompanied by sound or by sound & vibration did they create quicker responses. As a guideline, sound is a viable means of creating quick responses in highly urgent situations. It is noted that the task used in this study was a response one, where participants did not have to evaluate the cues' content, but rather automatically react. In this way, it was possible to assess performance of cues in the presence of an event requiring an imminent response, where identifying the cues' content may be less critical.

Interestingly, there was no difference in response time performance between abstract and language-based cues (H_{3b} was rejected). This is an indication that the designed language-based warnings perform as well as the abstract ones in this task. This is in line with (J. Edworthy, Walters, et al., 2000), where no difference was found in terms of how urgent abstract and language-based warnings were perceived. Although not a perception task, Experiment 7 showed similar response results for these two types of warnings. It could be argued that as long the cues' content is clear, the response to such high urgency warnings is more affected by the modality they are delivered in (multimodally and including audio) than by their content (language-based or abstract sounds). Considering the results of Experiment 4, where language-based cues received low annoyance ratings overall, these cues seem to present an advantage over abstract pulses in a critical situation. As will be described below, a trend towards better lane keeping performance when exposed to the language-based warnings is an additional indication of this advantage.

In terms of lateral deviation and steering angle, the results showed that the exposure to L_H cues led to poorer lane-keeping compared to prior to exposure to the cues (H_{4c} and H_{5c} were accepted). This is in line with Experiment 3, where the presence of cues did not improve or slightly worsened these metrics. It is therefore argued that the presence of cues along with critical events can create a distraction to the driving task. This is expected, since it is an additional factor for the driver to address. It has also been confirmed by studies like (Biondi, Rossi, Gastaldi, & Mulatti, 2014), where a startling effect of beeping cues was observed, leading to degradation of driving metrics. Additionally, since there is a physical reaction to the cues with braking, some increase in the driving metrics values is arguably justified. As long as this increase is not dramatic, and as long as the set of cues improves response performance compared to the absence of them, as has been shown to do in Experiment 3, this seems to be a necessary drawback when exposed to critical warnings. This also suggests that the use of warnings should be scarce, unless they signify critical events.

As described earlier, there was marginally better overall driving performance with the language-based cues, however the results did not reach significance (H_{4b} and H_{5b} were rejected). Therefore, a definite guideline on their advantage in this case cannot be provided, but they seem to create a trend towards better lane keeping behaviour. This could be addressing the problem of beeping cues in (Biondi et al., 2014), since speech may avert startling effects created by abstract sounds. This new finding could be further examined in future work, by investigating the use of less prominent speech cues in critical situations, or using abstract looming warnings found in studies like (Cristy Ho et al., 2013), where intensity in the cues changes with time. Finally, there was no influence of modality in lane keeping performance (H_{4a} and H_{5a} were rejected), indicating a uniform effect of the presented cues across modalities in the critical simulated task.

The presence of L_M and L_L cues, which had to be ignored by participants, did not disturb the driving metrics, irrespective of LDU, Modality, Information or Time (H_{6a} - H_{6d} and H_{7a} - H_{7d} were all rejected). This is arguably an important finding, since non-critical warnings should not add additional burden to the main task of driving, irrespective of their design, criticality or modality. Participants were very accurate in discriminating the L_H cues from the L_M and L_L ones and in reacting to L_H. In only one case out of 1120 trials did a participant mistakenly react to a L_M cue and in no case did anyone react to a L_L one. These are encouraging results for all the cues designed, showing their suitability for use in contexts of intermediate or low criticality, which may occur more frequently when driving.

7.6. Conclusions & Statement of Findings

This chapter provides the answer to the research question of *"How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?"* After the design of truly multimodal language-based cues in Chapter 6, the comparison of these cues with abstract ones in this chapter provided a set of implications related to their design and the modalities they utilised. The cues in this chapter were compared in terms of objective measures, deriving results on their recognition along different levels of designed urgency, as well as the reaction they are able to reduce when presented along with a critical event. This extends the objective comparisons derived in Chapter 6, presenting a complete set of results relating to both abstract and language-based warnings for drivers. In this chapter, although the differences between the two types of cues were only observed in an identification task with low criticality (Experiment 6), the multimodal character of the cues improved observed reactions in both Experiment 6 and Experiment 7. However, a degradation of steering performance (although arguably not a dramatic one) was observed in this case. Finally, the interchangeable character of cue designs when used in a critical event in Experiment 7 was identified, highlighting a wider design space for critical vehicle warnings. In summary, the following guidelines can be derived from this chapter:

- Abstract cues can result to quicker recognition in a low criticality task, *i.e.* recognizing warning urgency with no critical event on the road. They can be used as alerts for tasks that do not require immediate responses by the driver;
- Multimodal cues including visuals are suitable for the same task, since participants rely on a visual interpretation of the cues. When multimodal cues are necessary for salience, but still outside a critical context, visual cues can help interpretation of the message to be conveyed;

- Multimodal cues including audio create quicker responses in a high criticality task, *i.e.* responding to a car in front braking sharply. Audio is an effective means to create quick and accurate reactions to critical events when used multimodally;
- Abstract and language-based cues have similar response times when a critical event is presented and can be used interchangeably. However, the use of language-based cues marginally improves driving performance, leading to a possible advantage when signifying a critical event without disrupting driving performance;
- In high urgency situations, the use of warnings can lead to a slight degradation of steering performance, but they improve response times. However, when alerting the driver is essential, warnings are suggested, since their absence can lead to an absence of response to a critical event.

Experiments 6 and 7 concluded a series of experiments in this thesis, which focused on manual driving. An extensive set of guidelines on driver warnings were derived from these experiments, related to cue design, *i.e.* modalities to be used and designed urgency in the cues (Experiments 1 – 7), and semantic content of the cues (Experiments 6 and 7). There was also extensive investigation on the situations under which the cues were delivered, *i.e.* non-critical situations with an absence of critical event (Experiments 1, 2, 4, 5, 6), and critical situations, where an event requiring imminent response was presented along with the warnings (Experiments 3 and 7). All these results, although useful, do not address the more and more relevant use case of autonomous driving, which differs in many aspects to manual driving, and can motivate the design and evaluation of new warnings. This is why Experiments 8, 9 and 10 conclude the work in this thesis, by designing a set of warnings for autonomous driving scenarios, aiming to present a set of guidelines on the use of modality, urgency and message content in autonomous driving situations. The use of both subjective (Experiment 8) and objective measures (Experiments 9 and 10) follows the experimental design of the previous experiments of this thesis, and presents an evaluation of warnings signifying a very critical part of the autonomous vehicle interaction: the handover of control between the car and the driver. After a set of multimodal warnings varying in urgency signifying a handover of control are designed, they are evaluated in a series of experiments, initially seeking participants' subjective views on the warnings (Experiment 8), and finally simulating this handover situation across contexts of varying criticality, in order to assess the warnings' performance and their suitability as driving alerts.

8. Investigating Multimodal Driver Displays for Autonomous Vehicles

8.1. Introduction

As described in Section 2.6, although the interest in autonomous driving and its implications for transport is increasing (Kyriakidis et al., 2014), there is a distinct lack of research in partially autonomous scenarios, where a transition between autonomous and manual modes occurs (see NHTSA Level 3 Limited Self-Driving Automation (National Highway Traffic Safety Administration, 2013)). These scenarios can occur due to various reasons, like a vehicle malfunction or the driver's intention to take control from or give control to the vehicle. These scenarios have not yet been fully explored, while effective ways to communicate such situations to the driver have not yet been designed. Therefore, designing warnings that signify this point of transition of control (referred to as a handover) is essential. In parallel, as vehicle automation increases, drivers are more likely to engage in tasks other than driving. Gaming is a popular activity that drivers are expected to engage in, see for example (Krome et al., 2015; Neubauer et al., 2014). A critical handover often examined is an automation failure, since it happens unexpectedly, leaving little time to react (C. Gold et al., 2013; B. K. Mok et al., 2015; Pfromm et al., 2015). Signifying handovers with multimodal warnings (Naujoks et al., 2014), using varying message contents (Koo et al., 2014) and evaluating transition times (C. Gold et al., 2013; Christian Gold & Bengler, 2014) are important aspects of this critical case. However, there has been no study on designing effective multimodal warnings for diverse handover situations when the driver of an autonomous car is busy with a demanding side task. Further, there is no work on how handovers can be facilitated by multimodal warnings, whether from the car or from the area of the distracting interaction taking place though the side task. This motivated the final research question of this thesis: *How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?* This chapter answers this question with Experiments 8, 9 and 10, evaluating a set of multimodal warnings designed for vehicle handovers.

Experiments 1 – 7 provided an elaborate investigation of abstract and language-based warnings in manual driving scenarios. Utilising both subjective and objective measures, they resulted to a set of guidelines of using modality, urgency and message content to alert drivers in varying manual driving situations. Experiments 8, 9 and 10 follow the experimental framework set in the previous experiments, this time in autonomous driving scenarios. The experimental designs used in Experiments 8, 9 and 10 are similar to the previous studies of this thesis, so as to create comparable guidelines for the use of multimodal warnings varying in urgency and message content for different autonomous driving situations. Firstly, a set of scenarios for handovers that vary in urgency is envisioned and the appropriate multimodal warnings are designed. These warnings are then evaluated both in terms of subjective (Experiment 8) and objective measures (Experiment 9), in line with Experiment 4. They are also assessed in a critical scenario, during an automation failure, in Experiment 10. There, both abstract and language-based designs are used, in line with Experiments 6 and 7. In this way a complete set of guidelines on warning design for autonomous handovers is provided.

Section 8.2 describes Experiment 8, evaluating perceived urgency, annoyance and alerting effectiveness of the cues. Section 8.3 describes Experiment 9, evaluating response times and observed lateral deviation when returning to manual driving in both critical and non-critical scenarios. Section 8.4 initially discusses the findings of Experiments 8 and 9. Section 8.5 describes Experiment 10 evaluating the same driving metrics when returning to driving after a critical automation failure, and Section 8.6 discusses the findings of Experiment 10. Finally, Section 8.7 provides a general discussion for this chapter and Section 8.8 concludes with a set of guidelines for multimodal warning design for autonomous handovers.

8.2. Experiment 8: Investigating Subjective Responses to Language-based Warnings Designed for Autonomous Handovers

8.2.1. Warning Design

In order to investigate autonomous handover scenarios of varying urgency, a set of six speech messages covering a range of possible handovers of control between the car and the driver were designed. Three different Levels of Designed Urgency (LDU) were used for the envisioned situations (Level High - L_H, Level Medium - L_M and Level Low - L_L), as in all previous studies of this thesis. Situations where either the car would hand over control to the driver (CD) or the driver to the car (DC) were used. This resulted to six speech messages, presented in Table 8-1 (L_HCD, L_MCD, L_LCD, L_HDC, L_MDC and L_LDC). The messages used were adjusted from (J.D. Lee et al., 2008), containing a set of in-vehicle messages prioritized according to SAE J2395 (SAE, 2002). Adjustments were in order to avoid resemblance between messages in terms of rhythm. High priority messages in (J.D. Lee et al., 2008) were mapped to L_H, intermediate priority ones to L_M and low priority ones to L_L. The word *"Danger!"* was added before each L_H message, *"Warning!"* before L_M and *"Notice!"* before L_L, since this has shown to provide distinctively different urgency ratings in previous studies (C. L. Baldwin & Moore, 2002; Elizabeth Hellier et al., 2002). Finally, in L_H the handover was enforced, since imminent actions would be needed in such critical situations, while in L_M and L_L, the handover was only requested. This was reflected in the text of the messages, concluding with whether the messages were an enforced handover or handover request. In line with (Koo et al., 2014), the messages explained why a handover was necessary, rather than how the handover would happen.

Urgency	*Handover of Control*	
	CD	**DC**
L_H	Danger! Collision Imminent You have control! **D:** 2.7 *sec* **P:** -0.0 *dBFS* **AF:** 371 *Hz*	Danger! Object in roadway I have control! **D:** 3.0 *sec* **P:** -0.2 *dBFS* **AF:** 346 *Hz*
L_M	Warning! GPS signal weak Want to take over? **D:** 3.8 *sec* **P:** -11.0 *dBFS* **AF:** 309 *Hz*	Warning! Dense fog ahead May I take over? **D:** 3.4 *sec* **P:** -9.4 *dBFS* **AF:** 296 *Hz*
L_L	Notice! Toll ahead, 5 pounds Want to take over? **D:** 4.7 *sec* **P:** -18.5 *dBFS* **AF:** 212 *Hz*	Notice! New email from John May I take over? **D:** 4.2 *sec* **P:** -20.4 *dBFS* **AF:** 211 *Hz*

Table 8-1: The messages designed, using situations of High (L_H), Medium (L_M) and Low urgency (L_L). The handovers of control were from Car to Driver (CD) or from Driver to Car (DC). For each message the duration (D), peak (P) and average frequency (AF) of the audio are reported.

The messages were recorded by a female voice actor using a Rode NT2-A[18] condenser microphone. In line with (Elizabeth Hellier et al., 2002) and Experiment 4, the actor was instructed to speak messages of L_H urgently, as if a loved one was in imminent danger. L_M messages were spoken non-urgently, as if in a friendly conversation with nothing interesting about the situation and L_L messages were spoken in a monotone, deadpan manner. L_H messages were slightly modified to remove pauses between sentences so as to decrease duration. As tactile equivalents of the audio warnings, Speech Tactons with a C2 Tactor[19] were used, which were constructed following the procedure described in Section 6.2. Finally, for the visual warnings, the text of the warnings was displayed for the duration of the utterance and varied in colour, in line with Experiments 6 and 7 (Red for L_H, Orange for L_M and Yellow for L_L[20]). For all modifications, Praat[21] and Audacity[22] software was used.

The designed warnings were presented in all combinations of the audio, visual and tactile modalities: Audio (A), Visual (V), Tactile (T), Audio + Visual (AV), Audio + Tactile (AT), Tactile + Visual (TV), Audio + Tactile + Visual (ATV). As a result, 42 different cues were created, 7 cues with all modalities (A, T, V, AT, AV, TV, ATV) × 3 Levels of Designed Urgency (L_H, L_M, L_L) × 2 Situations (CD, DC). These warnings were evaluated in Experiments 8 and 9, looking into subjective (Experiment 8) and objective responses (Experiment 9) of participants when exposed to the cues. This has never been studied before and is essential in order to provide insights on how such warnings would be perceived.

8.2.2. Motivation

Experiment 8 focused on collecting subjective measures for a set of multimodal language-based warnings varying in urgency, signifying a handover of control in autonomous cars. In line with Experiments 1 and 4, the aim was to investigate whether the designed urgency in the warnings would be recognised in participants' responses. This is important, since, as discussed in Experiments 1 and 4, a match of designed a perceived urgency is key in designing effective alerts. Similarly, annoyance needed to be kept in manageable levels, so as to avoid creating alerts that would be disruptive. This is why perceived annoyance was

[18] http://www.rode.com/microphones/nt2-a
[19] http://www.atactech.com/PR_tactors.html
[20] Red was *RGB*(255,0,0), Orange was *RGB*(255,127,0) and Yellow was *RGB*(255,255,0).
[21] http://www.fon.hum.uva.nl/praat/
[22] http://audacity.sourceforge.net/

further assessed in the warnings. Finally, alerting effectiveness, as discussed in Experiment 4, can provide an effective means of assessing the suitability of an alert according to participants' opinions, and can be an indicator of suitability of the alert. To summarise, values of perceived urgency matching the designed urgency, low values of perceived annoyance, and high values of perceived alerting effectiveness were desirable subjective results, so as to create alerts, that convey the appropriate criticality and are considered effective, without annoying drivers. Experiment 8 was designed to assess these warnings' properties, in line with Experiments 1 and 4 of this thesis.

8.2.3. Design

For Experiment 8, a 7×3×2 within subjects design was used with Modality, LDH and Situation as the independent variables and Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) as the dependent ones. In line with Experiment 4 and (Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002), it was hypothesized that the designed urgency of the cues would be reflected in their perceived urgency. Perceived annoyance and alerting effectiveness were expected to also increase with cue urgency, as also observed in (Carryl L Baldwin, 2011). Finally, the work presented by Naujoks, Mai & Neukum (Naujoks et al., 2014), investigating the quality of signified handovers during an automation failure (situation CD), was extended by investigating multimodal rather than audio and visual warnings, and also covering the situation DC. It was expected that this situation would increase subjective measures, since the driver would be responsible for the outcome of the manoeuvre required as a result of the automation failure. As a result, there were the following hypotheses:

- The ratings of PU will be influenced by Modality (H_{1a}), LDU (H_{1b}) and Situation (H_{1c});
  - Specifically, PU was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in the situation CD, where the driver would be intended to take action in order to take control of the vehicle.
- The ratings of PA will be influenced by Modality (H_{2a}), LDU (H_{2b}) and Situation (H_{2c});
  - Specifically, PA was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in the situation CD, where the driver would be intended to take action in order to take control of the vehicle.
- The ratings of PAE will be influenced by Modality (H_{3a}), LDU (H_{3b}) and Situation (H_{3c}).

- Specifically, PAE was expected to increase in multimodal as opposed to unimodal cues, in cues of higher designed urgency, and in the situation CD, where the driver would be intended to take action in order to take control of the vehicle.

8.2.4. Procedure

Twenty-one participants (3 female) aged between 18 and 29 years (*M* = 21.00, *SD* = 2.84) took part in this experiment. They had not participated in previous experiments, except one, who had participated in Experiments 1 and 2. They all held a valid driving license and had between 1 and 8 years of driving experience (*M* = 3.36, *SD* = 2.01). All were right handed University students and reported normal vision and hearing.

The experiment took place in a University room where participants sat in front of 27-inch Dell 2709W monitor and a PC running the experimental driving simulator. Auditory cues were displayed through a set of Sennheiser HD 25-1 headphones. Tactile cues through a wristband on participants' left hand with a C2 Tactor attached to it, in line with Experiment 4. Visual cues were coloured text appearing for the duration of the utterance of the audio in the top centre of the screen, simulating a Head up Display (HuD), in line with Experiments 6 and 7. They were sized 228×700 pixels (about 7×21 *cm*). Participants provided all responses using a mouse. To cover any Tactor noise, car sound was played during the experiment.

After being welcomed and explained the experimental procedure, the 42 cues were displayed in a random order for participants to familiarize themselves with the signals. For each cue, they could either repeat it or proceed to the next when they felt familiar with it. In the main experiment, they were presented with the cues when sitting in front of a driving simulator depicting a rural road scene with a straight road and a car in front, in line with previous experiments in this thesis. The participants' car was self-driving. They were asked to imagine they were sitting in the driver's seat of an autonomous vehicle, wearing a wrist mounted device like a smart watch for vibration. Participants rated all cues in terms of PU, PA and PAE, by completing a 5-point Likert scale, in line with (Carryl L Baldwin, 2011) and Experiment 4. In all ratings, the scale was labelled: Not at all (1), Slightly (2), Moderately (3), Very (4) and Extremely (5). Each cue was presented twice, resulting to 84 trials.

8.2.5. Results

8.2.5.1. Perceived Urgency

Data for PU were analysed using a three-way repeated measures ANOVA, with Modality, LDU and Situation as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{1a}:*** There was a significant main effect of Modality ($F(3.20,131.25) = 29.88$, $p < 0.001$). Contrasts revealed that modalities were rated for PU in the following order: T and V lower than A ($F(1,41) = 20.11$, $r = 0.57$, $p < 0.001$), A lower than TV and AV ($F(1,41) = 7.80$, $r = 0.40$, $p < 0.05$), TV and AV lower than AT ($F(1,41) = 5.51$, $r = 0.34$, $p < 0.05$) and AT lower than ATV ($F(1,41) = 7.62$, $r = 0.40$, $p < 0.05$). See Figure 8-1 for mean values of PU across modalities and Table 8-2 for pairwise comparisons of PU across Modalities. ***Hypothesis H_{1b}:*** There was a significant main effect of LDU ($F(1.58,64.56) = 306.02$, $p < 0.001$). Contrasts revealed that levels were rated in the following order: L_L lower than L_M ($F(1,41) = 151.02$, $r = 0.89$, $p < 0.001$) and L_M lower than L_H ($F(1,41) = 282.06$, $r = 0.93$, $p < 0.001$) (see Figure 8-2). ***Hypothesis H_{1c}:*** There was a significant main effect of Situation ($F(1,41) = 24.52$, $p < 0.001$). Contrasts revealed that DC was rated lower than CD (Mean values of DC: 2.79, of CD: 2.99, $F(1,41) = 24.52$, $r = 0.61$, $p < 0.001$). As a result, hypotheses H_{1a}, H_{1b} and H_{1c} were accepted.

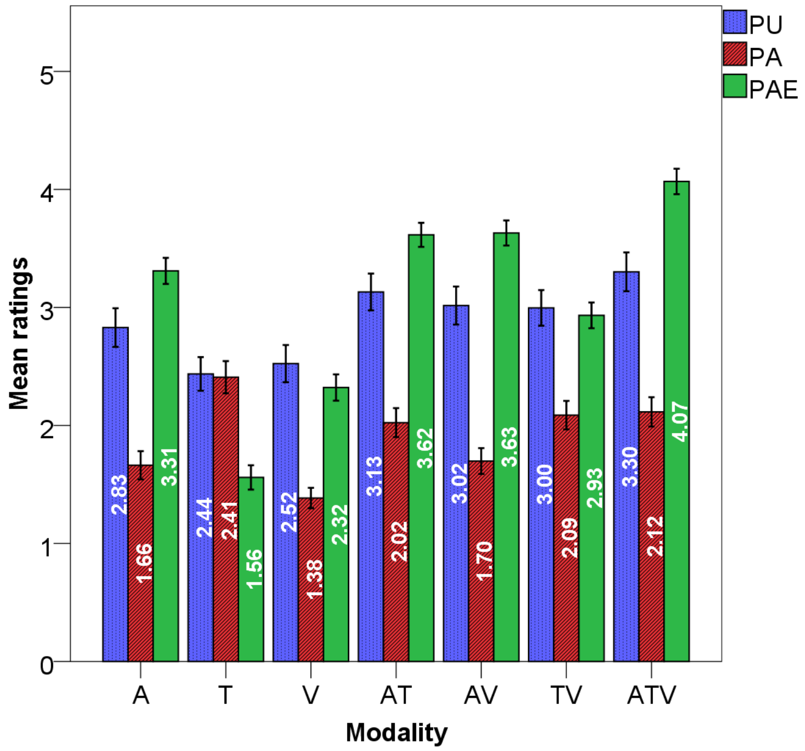


Figure 8-1: Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across modalities (hypotheses H_{1a}, H_{2a}, H_{3a}). Error bars indicate 95% confidence intervals.

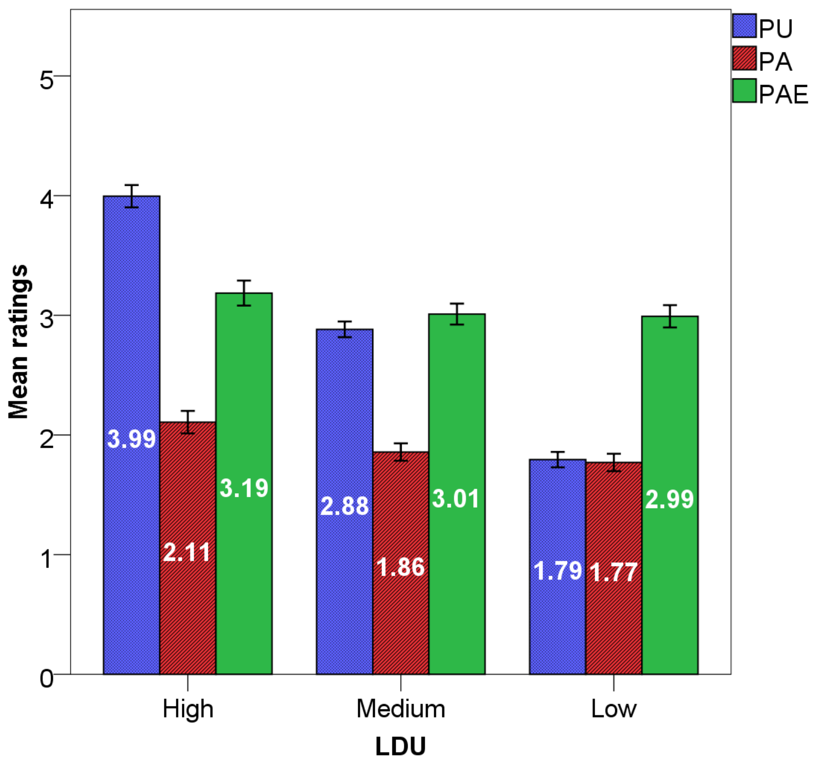


Figure 8-2: Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across LDU (hypotheses H_{1b}, H_{2b}, H_{3b}).

	T	V	A	TV	AV	AT	ATV
T		.435	**.001**	**.000**	**.000**	**.000**	**.000**
V	.435		**.000**	**.000**	**.000**	**.000**	**.000**
A	**.001**	**.000**		**.008**	**.000**	**.000**	**.000**
TV	**.000**	**.000**	**.008**		.740	**.026**	**.000**
AV	**.000**	**.000**	**.000**	.740		**.024**	**.000**
AT	**.000**	**.000**	**.000**	**.026**	**.024**		**.009**
ATV	**.000**	**.000**	**.000**	**.000**	**.000**	**.009**	

Table 8-2: Pairwise comparisons between modalities for Perceived Urgency (H_{1a}). Modalities are sorted by their mean values of PU. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Modality and LDU ($F(6.30,258.14) = 22.61$, $p < 0.001$), indicating that in modality T the ratings of PU were not differencing between L_H and L_M ($F(1,41) = 10.68$, $r = 0.45$, $p < 0.01$), while ratings of PU were not different between modalities V and A for L_M ($F(1,41) = 23.80$, $r = 0.61$, $p < 0.001$). Conversely, ratings of TV were lower compared to A in L_H, contrary to the main effect ($F(1,41) = 21.93$, $r = 0.59$, $p < 0.001$), while ratings of TV were higher compared to AV in L_M and L_L, again contrary to the main effect ($F(1,41) = 17.69$, $r = 0.55$, $p < 0.001$). Finally, in L_L ratings of ATV were lower compared to AT, contradicting the main effect ($F(1,41) = 5.34$, $r = 0.34$, $p < 0.05$). A significant interaction between LDU and Situation ($F(1.64,29.17) = 17.82$, $p < 0.05$), indicated that the described differences in ratings of Situations were stronger in L_H ($F(1,41) = 9.29$, $r = 0.43$, $p < 0.01$) and L_L ($F(1,41) = 29.64$, $r = 0.65$, $p < 0.001$). Finally, a significant interaction between all three factors ($F(7.21,295.53) = 2.31$, $p < 0.01$), indicated that in CD and L_H, the differences in ratings between ATV and AT were more pronounced ($F(1,41) = 5.69$, $r = 0.35$, $p < 0.05$). See Figure 8-3 for the interaction between Modality and LDU and Figure 8-4 for the interaction between LDU and Situation.

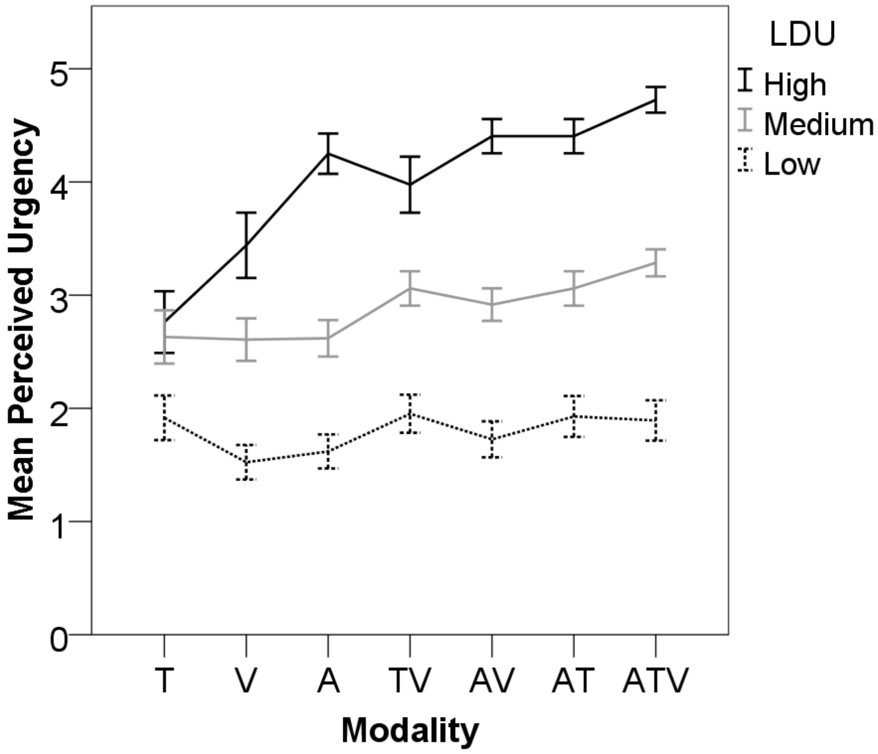


Figure 8-3: The interaction between Modality and LDU for Perceived Urgency (H_{1a} – H_{1b}). Modalities are sorted by their mean values of PU.

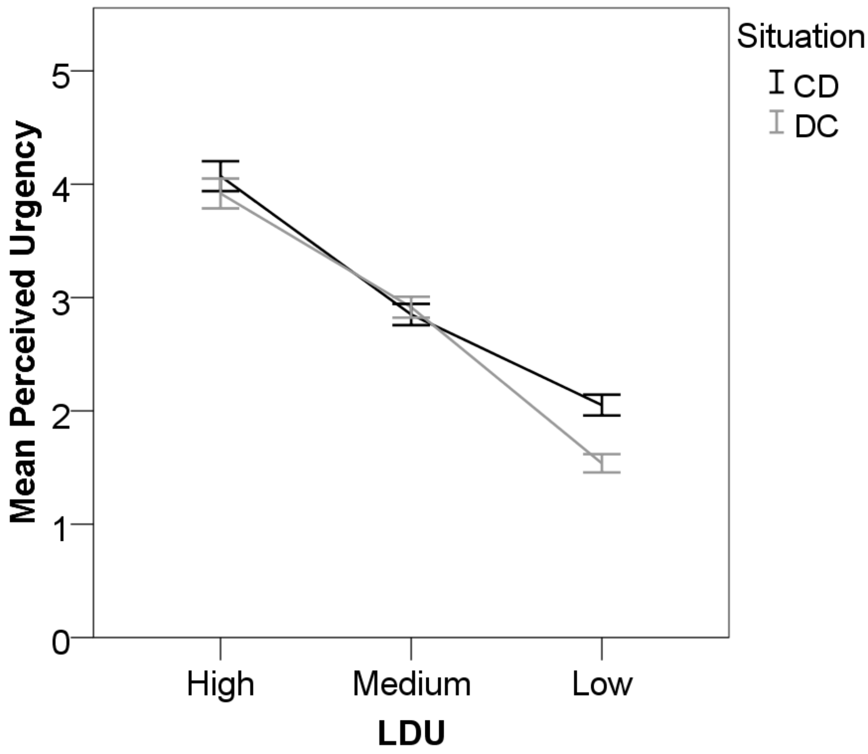


Figure 8-4: The interaction between LDU and Situation for Perceived Urgency (H_{1b} – H_{1c}).

8.2.5.2. Perceived Annoyance

Data for PA were analysed using a three-way repeated measures ANOVA, with Modality, LDU and Situation as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{2a}:*** There was a significant main effect of Modality ($F(2.82,115.63) = 20.52$, $p < 0.001$). Contrasts revealed that modalities were rated for PA in the following order: V lower than A and AV ($F(1,41) = 9.70$, $r = 0.44$, $p < 0.01$), A and AV lower than AT, TV and ATV ($F(1,41) = 14.21$, $r = 0.51$, $p < 0.01$) and AT, TV and ATV lower than T ($F(1,41) = 4.63$, $r = 0.32$, $p < 0.05$) (see Figure 8-1). See Table 8-3 for the pairwise comparisons of PA between modalities. ***Hypothesis H_{2b}:*** There was a significant main effect of LDU ($F(1.17,48.11) = 7.10$, $p < 0.05$). Contrasts revealed that L_L and L_M were rated lower than L_H ($F(1,41) = 8.09$, $r = 0.41$, $p < 0.001$, see Figure 8-2). As a result, hypotheses H_{2a} and H_{2b} were accepted, while H_{2c} was rejected. ***Interactions between main effects:*** There was a significant interaction between Modality and LDU ($F(5.45,223.47) = 3.60$, $p < 0.01$), indicating that ratings for L_M were as low as L_L for ATV but as high as L_H for T ($F(1,41) = 11.18$, $r = 0.46$, $p < 0.01$, see Figure 8-5).

	V	A	AV	AT	TV	ATV	T
V		**.003**	**.002**	**.000**	**.000**	**.000**	**.000**
A	**.003**		.517	**.000**	**.001**	**.000**	**.000**
AV	**.002**	.517		**.001**	**.001**	**.000**	**.000**
AT	**.000**	**.000**	**.001**		.421	.102	**.004**
TV	**.000**	**.001**	**.001**	.421		.754	**.000**
ATV	**.000**	**.000**	**.000**	.102	.754		**.037**
T	**.000**	**.000**	**.000**	**.004**	**.000**	**.037**	

Table 8-3: Pairwise comparisons between modalities for Perceived Annoyance (H_{2a}). Modalities are sorted by their mean values of PA. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

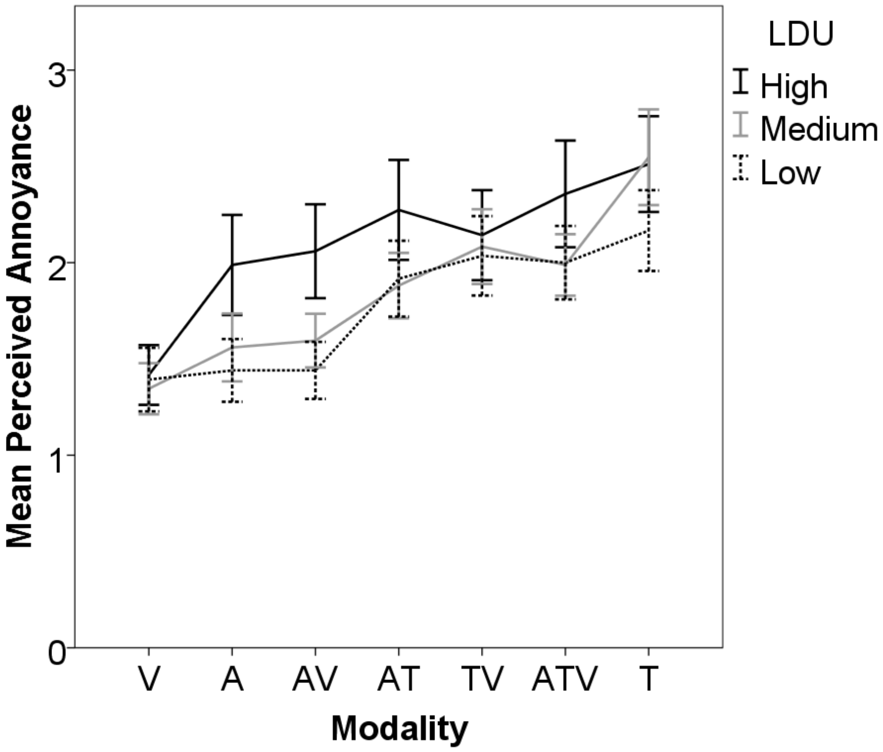


Figure 8-5: The interaction between Modality and LDU for Perceived Annoyance (H_{2a} – H_{2b}). Modalities are sorted by their mean values of PA.

8.2.5.3. Perceived Alerting Effectiveness

Data for PAE were analysed using a three-way repeated measures ANOVA, with Modality, LDU and Situation as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{3a}:*** There was a significant main effect of Modality ($F(3.38,138.75) = 146.70$, $p < 0.001$). Contrasts revealed that modalities were rated for PAE in the following order: T lower than V ($F(1,41) = 43.34$, $r = 0.72$, $p < 0.001$), V lower than TV ($F(1,41) = 76.83$, $r = 0.81$, $p < 0.001$), TV lower than A ($F(1,41) = 12.23$, $r = 0.48$, $p < 0.01$), A lower than AT and AV ($F(1,41) = 17.23$, $r = 0.54$, $p < 0.001$), AT and AV lower than ATV ($F(1,41) = 32.64$, $r = 0.67$, $p < 0.001$) (see Figure 8-1). See Table 8-4 for the pairwise comparisons of PAE between modalities. ***Hypothesis H_{3b}:*** There was a significant main effect of LDU ($F(1.36,55.63) = 4.78$, $p < 0.05$). Contrasts revealed that L_L and L_M were rated lower than L_H ($F(1,41) = 8.34$, $r = 0.41$, $p < 0.01$) (see Figure 8-2). ***Hypothesis H_{3c}:*** There was a significant main effect of Situation ($F(1,41) = 5.45$, $p < 0.05$). Contrasts revealed that CD was rated lower than DC (Mean values of CD: 3.02, DC: 3.10, $F(1,41) = 5.45$, $r = 0.34$, $p < 0.05$). As a result, hypotheses H_{3a}, H_{3b} were accepted, and H_{3c} was rejected.

	T	V	TV	A	AT	AV	ATV
T		**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
V	**.000**		**.000**	**.000**	**.000**	**.000**	**.000**
TV	**.000**	**.000**		**.001**	**.000**	**.000**	**.000**
A	**.000**	**.000**	**.001**		**.000**	**.000**	**.000**
AT	**.000**	**.000**	**.000**	**.000**		.856	**.000**
AV	**.000**	**.000**	**.000**	**.000**	.856		**.000**
ATV	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 8-4: Pairwise comparisons between modalities for Perceived Alerting Effectiveness (H_{3a}). Modalities are sorted by their mean values of PAE. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Modality and Level ($F(7.53,308.69) = 12.13$, $p < 0.001$), indicating that for modality V ratings were lower in L_H compared to L_M and L_L ($F(1,41) = 5.87$, $r = 0.35$, $p < 0.05$), while they were similar across levels for modality T ($F(1,41) = 7.36$, $r = 0.39$, $p < 0.05$) and TV ($F(1,41) = 4.81$, $r = 0.32$, $p < 0.05$). They also revealed that for modality A ratings were returning to accordance with the main effect and were higher in L_H than in L_M and L_L, where they did not differ to each other ($F(1,41) = 10.12$, $r = 0.44$, $p < 0.01$). For ATV, this higher rating of L_H was even more pronounced ($F(1,41) = 14.55$, $r = 0.52$, $p < 0.001$). However in modality AT, other than the difference between L_H and the other two levels, L_M also rated higher than L_L ($F(1,41) = 7.27$, $r = 0.39$, $p < 0.05$). Finally, there was a significant interaction between Modality and Situation $F(6,246) = 2.43$, $p < 0.05$), indicating that the described differences in ratings for situations CD and DC were more pronounced in modalities V ($F(1,41) = 5.27$, $r = 0.34$, $p < 0.05$) and A ($F(1,41) = 8.13$, $r = 0.41$, $p < 0.01$). See Figure 8-6 for the interaction between Modality and LDU and Figure 8-7 for the interaction between LDU and Situation.

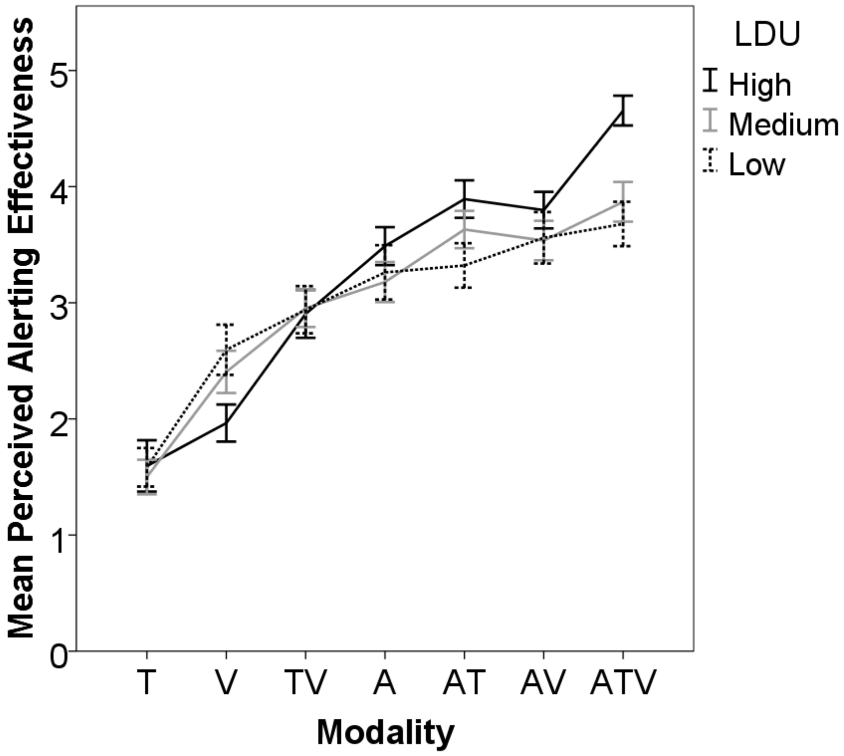


Figure 8-6: The interaction between Modality and LDU for Perceived Alerting Effectiveness (H_{3a} – H_{3b}). Modalities are sorted by their mean values of PAE.

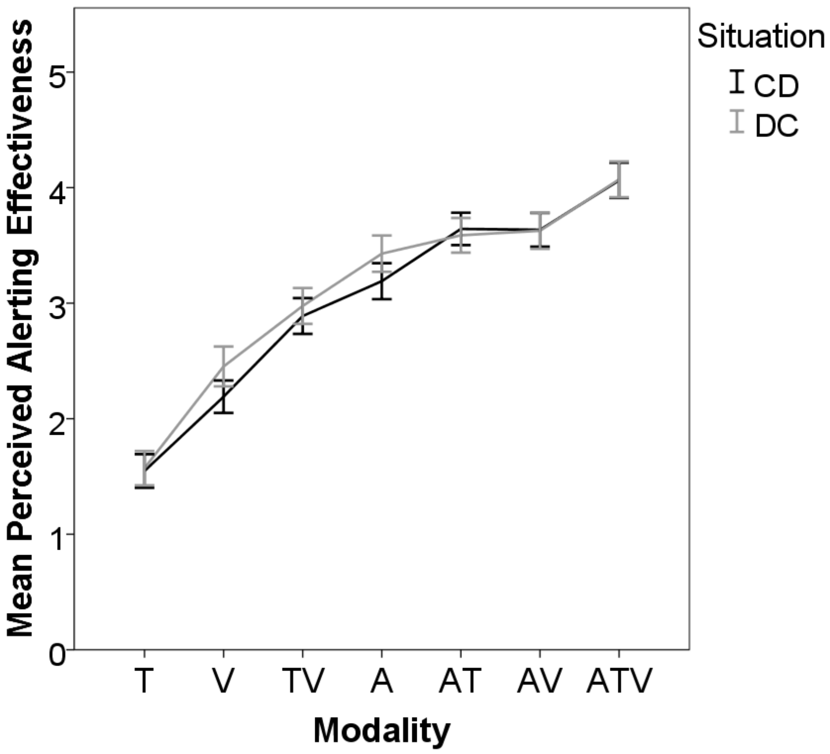


Figure 8-7: The interaction between LDU and Situation for Perceived Alerting Effectiveness (H_{3b} – H_{3c}). Modalities are sorted by their mean values of PAE.

The results of Experiment 8 showed clearly that participants identified the designed urgency of all the signals and rated highly urgent warnings as more effective and more annoying, but with low values of annoyance overall, similar to Experiment 4. Multimodal warnings were rated as more urgent and more effective, while unimodal tactile warnings as the most annoying and least effective. To investigate how the cues supported handover situations, the messages relating to cases where the driver needed to take control (situation CD) were assessed in Experiment 9.

8.3. Experiment 9: Investigating Handover Times for Language-based Multimodal Warnings

8.3.1. Motivation

Experiment 9 focused on collecting objective measures for the designed language-based multimodal warnings varying in urgency, signifying handovers of control in autonomous cars. As discussed in Experiments 2, 3, 6 and 7, collecting objective measures for multimodal alerts approximates a real driving scenario, where reactions need to happen in a specific amount of time. This approach was also followed for Experiment 9, where a set of handovers from the car to the driver and *vice versa* was accompanied by multimodal language-based alerts. The performance to this task in terms of speed and accuracy would reveal the effectiveness of the designed warnings and provide a baseline set of results in the context of autonomous handovers of control. An experimental design similar to the one used in Experiments 2, 3, 6 and 7 was used, so as provide comparable guidelines in an autonomous driving scenario aside to guidelines in a manual one, and present an extensive set of results for multimodal warnings varying in urgency in both driving modes. As discussed in previous chapters, rapid and accurate reactions to driver alerts are essential, especially when the events signified are critical. Therefore, Experiment 9 collected data on reactions to the warnings in the designed situations, which would assess their effectiveness in a simulated driving task, and provide insights on their utility in a road scenario.

8.3.2. Design

Experiment 9 investigated how quickly and effectively distracted participants would be able to resume control in an autonomous vehicle. The focus was on situation Car to Driver (CD), since this is the situation where the driver would need to act quickly and take over control of the vehicle in a real driving scenario. The handover was either requested (L_M, L_L) or enforced (L_H). A task was designed, where the driver would be distracted by a tablet game while not driving but would need to return to driving periodically. The experiment investigated how quickly and accurately this transition would happen and how it would affect driving after it happened. A 7×3 within subjects design was used, with Modality and LDU as the independent variables and Response Time (ResT), Response Accuracy (ResA) and Lateral Deviation after Handover (LDaH) as the dependent ones. In line with Naujoks, Mai & Neukum (Naujoks et al., 2014), the urgency of the situation, also reflected in the warnings, was expected to affect the driving metrics. Further, a richer set of scenarios would be evaluated, by varying the warning urgency, the modalities used and the nature of the required response, and investigating the influence of this intervention. In line with Experiments 3, 6 and 7, it was expected that multimodal cues and cues of higher designed urgency would positively influence all metrics. As a result, there were the following hypotheses:

- The observed ResT will be influenced by Modality (H_{4a}) and LDU of the warnings (H_{4b});
  - Specifically, ResT was expected to decrease in multimodal as opposed to unimodal cues and higher levels of LDU.
- The observed ResA will be influenced by Modality (H_{5a}) and LDU of the warnings (H_{5b});
  - Specifically, ResA was expected to increase in multimodal as opposed to unimodal cues and higher levels of LDU.
- The observed LDaH will be influenced by Modality (H_{6a}) and LDU of the warnings (H_{6b}).
  - Specifically, LDaH was expected to decrease in multimodal as opposed to unimodal cues and higher levels of LDU.

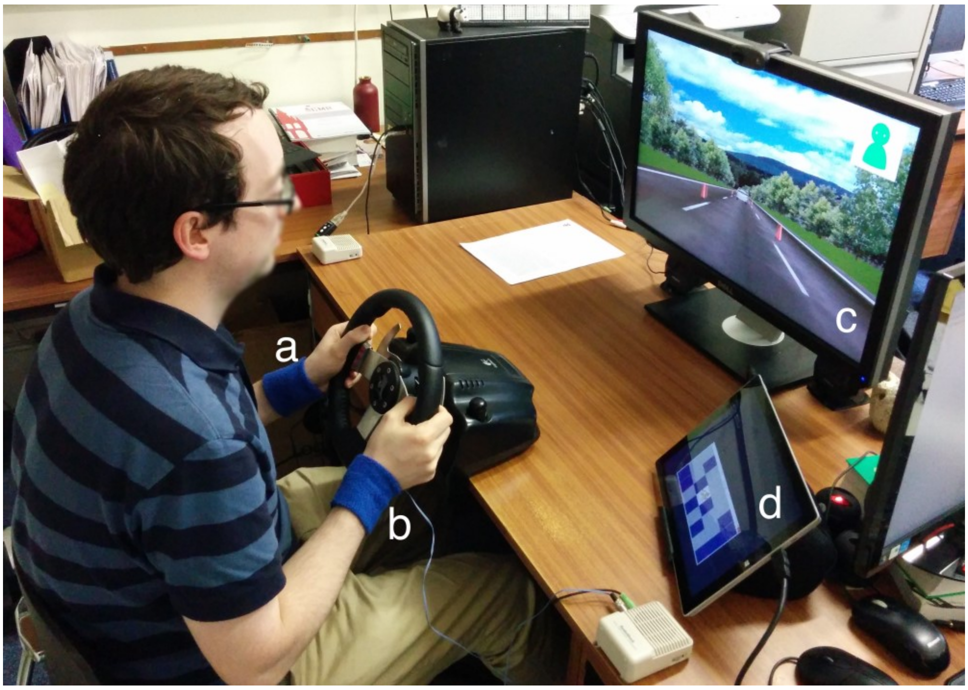


Figure 8-8: The setup of Experiments 9 and 10, with the tactile wristbands (a, b), the driving simulator (c) and the tablet (d). The tablet and actuators used in Experiment 9 were slightly different to the picture, but performed the same functions. Further, in Experiment 9 a set of headphones was worn by the participants. Finally, wristband (b) was only used in Experiment 10.

Figure 8-9: The setup of Experiment 9. In a critical event which the car could not address, a handover to the driver was enforced (a). In non-critical events, a handover to the driver was requested (b),(c). When in autonomous mode (d), drivers were playing a tablet game (e). Icons on the top right corner of the simulator screen denote whether the car is controlled by the driver (person icon) or is autonomous (car icon). See also http://youtu.be/ni048BpTDG8 for a video of the experimental setup.

8.3.3. Procedure

Experiment 9 followed immediately after Experiment 8. Participants and equipment were identical, with the addition of a Logitech G27 gaming wheel and pedals, and a ViewSonic View Pad 10pro 10-inch tablet computer. For this experiment, a side-task on the tablet, in line with (C. Gold et al., 2013), was introduced to participants. This would decrease their engagement with the driving task and provide a more challenging transition back to driving (de Winter et al., 2014). It is also likely that drivers may play games on tablets when not driving. The side task was the Concentration memory game previously used in the context of multimodal home reminders (Warnock, McGee-Lennon, & Brewster, 2011) (see Figure 8-9.e). It was chosen since it has a well-defined set of performance metrics and requires high levels of concentration. The driving scene used was similar to Experiment 8, depicting a curvy rural road with a car in front (Figure 8-9.d). After getting familiarized with the game and the driving simulator, participants were asked to focus on playing the game, unless interrupted by a warning. They were instructed to rest their left hand on the table and their feet on the floor. They were able to use their right hand to play the game by using the tablet, which was placed on a table top stand on the right of the screen, to better accommodate their dominant hand. If they matched all cards in the 3×8 grid, the game would reload automatically with a different random set of cards. While playing the game, the participants' car was self-driving in autonomous mode in the centre of the lane at a speed of around 70 *mph*. This mode simulated Level 3 Automation by NHTSA (National Highway Traffic Safety Administration, 2013) and did not require any driver intervention, but did require availability for occasional control. See Figure 8-8 for the setup of Experiment 9.

At random intervals of any integral value between (and including) 27–32 *sec* (in line with previous experiments and (Merat et al., 2014), where automation was regularly disengaged) a warning was presented. If it was an L_L or L_M warning (messages L_LCD and L_MCD), control from the driver was requested. This simulated events, where taking control of the vehicle was not critical. In this case, participants were instructed to press a labelled button on the steering wheel to return to driving, in line with (Naujoks et al., 2014) (see Figure 8-9.b, and 7-9.c for the visual warnings for these situations). If the warning presented was a L_H one (message L_HCD), control from the driver was required. This simulated an automation failure the vehicle could not correct and therefore a switch to manual mode would be needed. To create a more critical situation, the car in front in the L_H case started braking along with the warning presentation, as in Experiments 3, 6 and 7 (see Figure 8-9.a for the visual warning for this situation). Participants were then handed control and were instructed to brake immediately with their right foot and then return to normal driving. Once the participants stepped on the brake, the car in front would return to its original position away from the participants' car. It should be noted that the interval of 27–32 *sec* had limited ecological validity, since events requiring attention are expected to occur less frequently. However, it was necessary to be able to evaluate all the different cues designed. Other studies have used similar intervals, *e.g.* (Cristy Ho, Tan, et al., 2005).

Each warning was presented twice in the above setting, resulting in a total of 42 presentations (7 Modalities × 3 Levels of Designed Urgency × 2 presentations). When back to driving, participants were able to steer using the wheel for 10 *sec* (they did not need to use the accelerator pedal). During this period, they were asked to maintain a central lane position. After 10 *sec*, the car automatically took control and the next trial started. The mode was always indicated on the screen with a car icon for autonomous mode or a person icon for manual (see Figures 8-9.a- and 8-9.d, top right). Even though in a real driving scenario, drivers might not wish to regain control in non-critical situations, participants were still asked to react as quickly as possible in all cases, to be able to measure response time.

Participants' ResT was calculated from the onset of a stimulus until the participants pressed the button on the steering wheel (for L_L, L_M) or pressed the brake pedal (for L_H). If participants did not respond to a cue, their ResA was 0. Otherwise, their ResA was 1 if they performed the right action in the first instance (pressing the pedal or the button) otherwise it was 0 (if they performed the wrong action initially, even if they later corrected it). Their LDaH was the RMSE of their lane position values, logged for 10 *sec* after they pressed the button on the steering wheel (for L_L, L_H) or 10 *sec* after the onset of a L_H stimulus and start of the braking event of the lead car (for L_H). The value of 10 *sec* was chosen since it has shown to be an adequate time to come back to driving in handover situations (Merat et al., 2014). Experiments 8 and 9 together lasted about 60 minutes and participants were then debriefed and paid £6.

8.3.4. Results

8.3.4.1. Response Time

In all there were 882 trials. If participants did not respond to a cue (which was the case in 83 trials – 9.4%), their ResT was adjusted to the maximum available time to respond, 10 *sec*, to allow for a two factor ANOVA analysis. Data for ResT were analysed using a two-way repeated measures ANOVA, with Modality and LDU as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. **Hypothesis H_{4a}:** There was a significant main effect of Modality ($F(2.35,96.19) = 99.22$, $p < 0.001$). Contrasts revealed that AV, ATV, A, AT and TV caused quicker responses than T ($F(1,41) = 4.98$, $r = 0.33$, $p < 0.05$) and T created quicker responses than V ($F(1,41) = 127.67$, $r = 0.87$, $p < 0.001$). See Figure 8-10 for mean values of ResT across modalities and Table 8-5 for pairwise comparisons of ResT across modalities. **Hypothesis H_{4b}:** There was a significant main effect of LDU (Mean values of L_H: 2.15 *sec*, L_M: 3.41 *sec*, L_L: 3.45 *sec*, $F(1.55,63.47) = 37.27$, $p < 0.001$). Contrasts revealed that L_H cues caused quicker responses than L_M and L_L ones ($F(1,41) = 47.19$, $r = 0.73$, $p < 0.001$). As a result, H_{4a} and H_{4b} were accepted. ***Interactions between main effects:*** There was a significant interaction between Modality and LDU L_M ($F(5.31,217.68) = 1.96$, $p < 0.05$), indicating that contrary to the main effect, TV created quicker responses than AT in L_M ($F(1,41) = 16.96$, $r = 0.54$, $p < 0.001$).

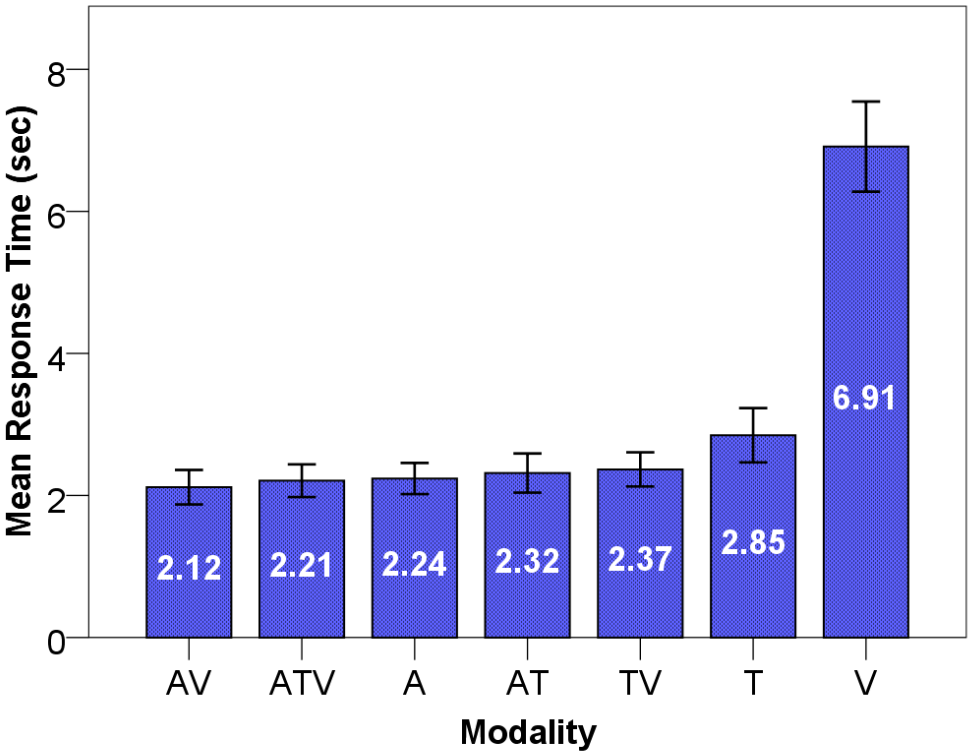


Figure 8-10: Response Times (ResT) across modalities for Experiment 9 (hypothesis H_{4a}). Modalities are sorted by their mean values of ResT.

	AV	ATV	A	AT	TV	T	V
AV		.541	.432	.242	.097	**.003**	**.000**
ATV	.541		.781	.506	.340	**.004**	**.000**
A	.432	.781		.463	.353	**.004**	**.000**
AT	.242	.506	.463		.693	**.016**	**.000**
TV	.097	.340	.353	.693		**.031**	**.000**
T	**.003**	**.004**	**.004**	**.016**	**.031**		**.000**
V	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 8-5: Pairwise comparisons between modalities for Response Time (H_{4a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

8.3.4.2. Response Accuracy

Hypothesis H_{5a}: The values of ResA for Modalities were as follows: V: 40%, T: 88%, TV: 92%, ATV: 93%, A: 94%, AT: 95% and AV: 98%. Data for ResA were treated as dichotomous and analysed with Cochran's Q tests. It was found that modality V was the least accurate compared to all other modalities (all comparisons were significant with $p < 0.001$ and $52.94 \leq Q(1) \leq 71.05$). It was also found that T was less accurate than AT ($Q(1) = 5.40$, $p < 0.05$) and AV ($Q(1) = 9.94$, $p < 0.01$). Finally, AV was more accurate than TV

($Q(1) = 5.33$, $p < 0.05$) and ATV($Q(1) = 5.44$, $p < 0.05$). ***Hypothesis H_{5b}:*** The resulting values of ResA for LDU were as follows: L_H: 83%, L_M: 86%, L_L: 89%. Cochran's Q tests revealed that L_H was less accurate than L_L ($Q(1) = 4.63$, $p < 0.05$). As a result, H_{5a} was accepted and H_{5b} was rejected.

8.3.4.3. Lateral Deviation after Handover

Data for LDaH were analysed using a two-way repeated measures ANOVA, with Modality and LDU as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{6a}:*** There was a significant main effect of Modality ($F(1.28,52.54) = 12.03$, $p < 0.001$). Contrasts revealed that V warnings resulted in higher LDaH compared to all other modalities (AV, A, AT, TV, T and ATV, $F(1,41) = 11.62$, $r = 0.47$, $p < 0.01$). See Figure 8-11 for mean values of LDaH across modalities and Table 8-6 for pairwise comparisons of LDaH across modalities. ***Hypothesis H_{6b}:*** There was a significant main effect of LDU ($F(1.02,41.78) = 36.06$, $p < 0.001$). Contrasts revealed that L_H warnings led to higher LDaH compared to L_M and L_L ($F(1,41) = 35.10$, $r = 0.68$, $p < 0.001$). As a result, H_{6a} was accepted and H_{6b} was rejected. ***Interactions between main effects:*** There was a significant interaction between LDU and Modality ($F(1.25,51.20) = 18.71$, $p < 0.001$). Contrasts revealed that for L_H, LDaH values increased in modality V, while for L_M and L_L they decreased ($F(1.25,51.20) = 18.86$, $r = 0.56$, $p < 0.001$).

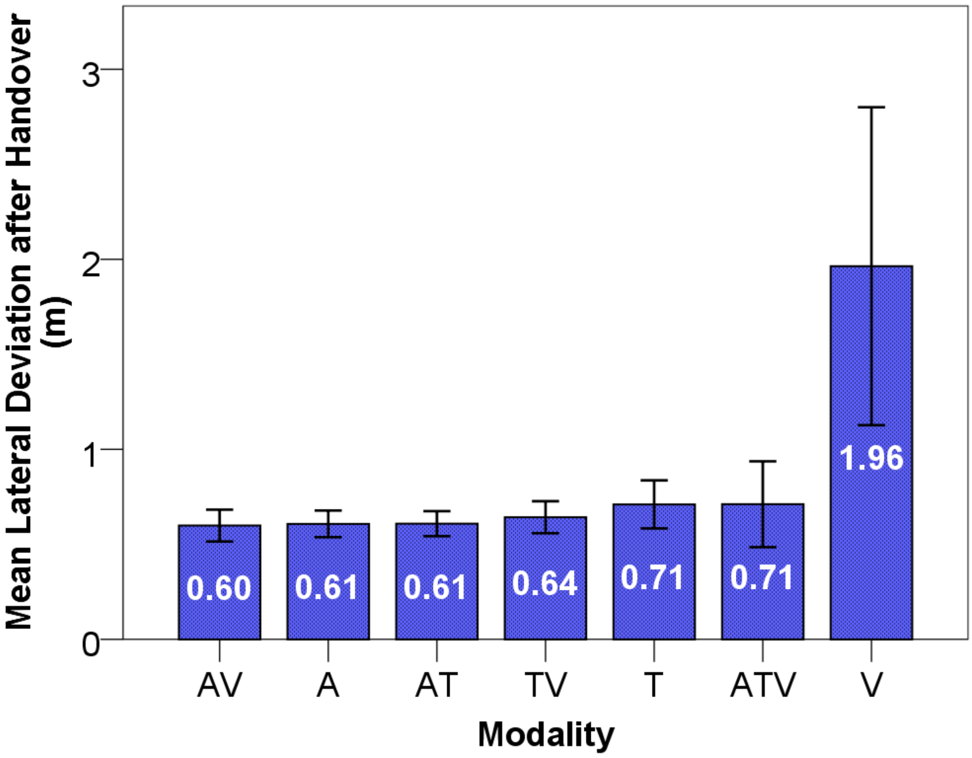


Figure 8-11: Lateral Deviation after Handover (LDaH) across modalities for Experiment 9 (hypothesis H_{6a}). Modalities are sorted by their mean values of LDaH.

	AV	A	AT	TV	T	ATV	V
AV		.863	.785	.427	.058	.257	**.001**
A	.863		.982	.538	.095	.398	**.001**
AT	.785	.982		.543	.122	.353	**.001**
TV	.427	.538	.543		.385	.546	**.001**
T	.058	.095	.122	.385		.994	**.001**
ATV	.257	.398	.353	.546	.994		**.001**
V	**.001**	**.001**	**.001**	**.001**	**.001**	**.001**	

Table 8-6: Pairwise comparisons between modalities for Lateral Deviation after Handover (H_{6a}). The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Finally, the results of the tablet game performance were as follows: 134.70 *sec* mean time to complete one game, 0.54 *Clicks per Second* and 0.33 *Superfluous Views per Click*[23]. These are similar to (Warnock et al., 2011), showing that participants were attentive to the game and it as effectively engaging them.

[23] Superfluous views show how many decisions (pictures tapped) were not successful. When a picture was viewed it was marked as 'seen'. Every subsequent viewing of that picture failing to match it to another picture was a superfluous view.

8.4. Discussion for Experiments 8 and 9

The results of Experiment 8 showed clearly that participants were able to identify the designed urgency of the cues, since their perceived urgency matched their designed urgency. This confirmed results of previous studies like (Carryl L Baldwin, 2011; Elizabeth Hellier et al., 2002) and arguably showed the cues would be suitable for autonomous handovers. Ratings were higher as the number of modalities increased (H_{1a}, H_{2a}, H_{3a}), in line with previous experiments. Finally, situation CD, in which the driver needed to act during a handover, was rated higher in terms of urgency compared to DC (H_{1c}). This could reflect an increased alertness on behalf of the driver when their intervention is needed, as opposed to when the handover is suggested or performed by the car. It is an interesting finding, since it highlights how drivers' perception of urgency could have been affected by their involvement in the situation.

Regarding the interactions between main effects for PU, in line with Experiment 4, it was found that unimodal tactile cues presented lower variability of ratings across levels compared to the rest of the cues (interaction between Modality and LDU, H_{1a} – H_{1b}). This could be due to the reduced salience of these cues and has also been discussed in previous chapters, where the unimodal use of tactile was not suggested when salience is required in the cues. The same interaction revealed similar ratings of PU for modalities V and A in L_M, arguably further indicating the limited salience that unimodal cues create. In L_H, an elevated perceived urgency achieved by unimodal audio cues could be attributed to the highly urgent character of the audio message for this level, which was evidently not perceived urgent as its tactile and visual counterpart. The interaction between LDU and Situation (H_{1b} – H_{1c}), and between all three factors (H_{1a} – H_{1b} – H_{1c}) revealed that the differences in PU between situations were more pronounced in L_H and L_L, and in modalities ATV and AT, possibly indicating that the more distinguishable cue design in the above settings was able to convey the desired urgency more effectively.

In terms of perceived annoyance, the observed values were low overall (between "Slightly" and "Moderately"). Cues in L_H were rated as more annoying, confirming previous results of this thesis (H_{2b}). The highest rated modality for annoyance was T, indicating that Speech Tactons are more acceptable when used in conjunction with other modalities, as in Experiment 4 (H_{2a}). This was especially true in L_H and L_M, where the message would be more important to act upon in a real situation compared to L_L (interaction between Modality and LDU, $H_{2a} – H_{2b}$). It is interesting that annoyance ratings of T surpassed even trimodal ratings, while many participants during unplanned discussions after the end of the experiment anecdotally reported that Speech Tactons were less understood when used alone. This confirms that the use of Speech Tactons unimodally is not suggested.

Ratings of alerting effectiveness confirmed the importance of informing drivers about critical events, as in Experiment 4. Regarding hypothesis H_{3b}, L_H was rated higher than L_M and L_L, showing that alerts about most critical situations are considered more effective. There was a particular preference to cues containing audio (H_{3a}), which can indicate the saliency of the cues. Unimodal T and V cues, but also TV ones were rated as less effective compared to A. Visual cues were rated low in L_H (interaction between Modality and LDU, $H_{3a} – H_{3c}$), possibly indicating that participants appreciated more salient ways to be informed during critical situations. The same interaction revealed a similar rating of T and TV cues across levels, which could indicate a uniform perceived effectiveness for these cues independently of how urgent they were designed to be. Finally, contrary to hypothesis H_{3c}, situation DC was rated as more effective compared to CD, especially in modalities V and A (interaction between Modality and Situation, $H_{3a} – H_{3c}$). This points to an elevated desire to maintain situation awareness when the driver is not involved in the manoeuvre about to take place. This can also indicate that the preference for simpler cues for situation DC was due to the fact that this situation does not require driver action, thus it can be communicated with less modalities.

Regarding handover times in Experiment 9, it was found that in L_H participants returned to driving significantly quicker than in L_M and L_L (H_{4b}). This is a novel finding, arguably confirming that reaction to a critical situation can be reflected in participant responses in this context. Although there were different motor requirements to perform each task (using the foot to brake vs the hand to press a button), this indicates that participants had increased alertness in L_H situation. This could also be supported by some participants' unplanned comments after the experiment, mentioning that they felt inclined to respond more quickly in an urgent event. Conversely, contrary to hypothesis H_{5b}, L_H cues created less accurate responses, which possibly indicates an elevated demand of the manual task performed in L_H.

A significant increase in handover time was observed for the V condition which had a much higher ResT (H_{4a}, also observed in (Naujoks et al., 2014)). This possibly reflects the visual attention required by the game, which led to a high number of missed responses, and low response accuracy values (H_{5a}). Drivers took an average of 6.9 *sec* to notice visual warnings about handovers, which would be catastrophic in real driving. Thus, great care must be taken using unimodal visual displays when the driver's visual attention is focused towards another device. Experiment 10 investigates whether this problem can be ameliorated by presenting the warnings in the interaction area, by placing warnings on the tablet. However, with the current arrangement, the use of unimodal visuals for autonomous vehicles handovers is not recommended. The second worse performing modality both in terms of ResT as well as ResA was T, again highlighting similar limitations to Experiment 5, when interpreting vibration in unimodal presentation of Speech Tactons. One limitation of this study was that the game was purely visual. It is common for games to use audio and tactile feedback too. Experiment 10 used a game with more feedback types to study if these interfere with the different modalities of feedback from the car.

In terms of LDaH, it had highest values for the visual modality (H_{6a}), in line with (Naujoks et al., 2014). This is due to the high number of failures to respond to handovers signified visually in L_H (15 out of 42 trials). These failures are critical, since they would leave the car uncontrolled. Of course, this particular transition would be much more difficult to test in a real driving setting. Again, the high distraction caused by the side task is arguably reflected in this finding. Presenting handover warnings visually on a HuD would not help overcome this distraction if the HUD was not inside the driver's visual field, and thus such visual feedback is not recommended.

The fact that LDaH was decreased for L_M and L_L, and especially in modality V (H_{6a}, and interaction between Modality and LDU, H_{6a} – H_{6b}) is actually a negative result, indicating a high number of failures to respond to a unimodal visual warning. If participants did not respond, the car would remain under autonomous mode and values of LDaH were zero. These decreased LDaH values would not necessarily lead to problems, since, in the real world, driver intervention is not essential. However, they can still indicate that unimodal visuals are not salient enough to attract attention during a handover.

8.5. Experiment 10

8.5.1. Motivation

As discussed earlier in this chapter, and in Section 2.6, interaction with games while driving is a little explored topic, with available studies mainly in cases where the car is fully autonomous and no intervention is expected (see for example (Krome et al., 2015; Terken et al., 2013)). This motivated Experiments 8 and 9, where autonomous handovers were signified by the simulator. However, resuming control with the help of warnings originating from the area of the gaming interaction as opposed to coming from the car itself has not been studied. This is an interesting investigation, since the expectation is that the saliency of the warnings could be improved when drivers receive them from the area of the distracting interaction of the game. Evidence of the effectiveness of presenting cues from an external device in the car have been discussed in prior studies (see Section 2.6), however with a limited set of cues. Experiment 10 addressed this gap by investigating a set of urgent multimodal warnings designed for an automation failure, requiring immediate driver attention. Warnings were delivered either from a simulator (as in Experiments 8 and 9) which is the most common approach in the literature, or from a tablet where the user was playing a game as a secondary task. As a further intervention, different warning designs were compared, utilising abstract and language-based cues never before used in this setting. In this way, an extensive comparison of multimodal warning designs, as presented in Experiments 6 and 7, would be performed in an autonomous driving scenario and for critical warnings in Experiment 10. As discussed in Experiments 6 and 7, varying cue design can increase the flexibility of conveying a message depending on the situation at hand. Finally, the use of objective measures for handovers of control would conclude along with Experiment 9 the collection of performance data for an autonomous driving situation, and add the resulting insights to this thesis' guidelines on multimodal warnings varying in urgency and message content, in both manual and autonomous driving settings.

8.5.2. Warning Design

The warnings designed addressed a highly urgent situation, where a car would hand over control to the driver during a critical event, due to an automation failure. The abstract warnings consisted of pure tones, colours or vibrations delivered as repeated pulses, as in Experiments 1 and 2. In line with these, the warnings had an increased pulse rate to convey high urgency. They consisted of 8 pulses having 0.1 *sec* single pulse duration and interpulse interval and had 1.5 *sec* duration. The auditory warning varied additionally in base frequency (1000 *Hz*) in line with (Judy Edworthy et al., 1991b). The visual warning also varied in colour and was Red (*RGB*(255,0,0)). The tactile warning had a frequency of 150 *Hz*, the nominal centre frequency of the ELV-1411A Tactor[24], used to deliver vibrational messages. The Tactor chosen was different to Experiments 8 and 9, and was more silent during operation. This was useful, since no headphones were used to mask the Tactor noise in Experiment 10. As in Experiments 6 and 7, the abstract audio and tactile cues had the same intensity as the speech cues. Simultaneous delivery of unimodal signals was used in the multimodal cues, creating a synchronous effect of sound, vibration and visuals.

For the language-based warnings, the speech message used was taken from Experiments 8 and 9. It was a high priority message according to (J.D. Lee et al., 2008; SAE, 2002), with the word "Danger!" added in the beginning to increase perceived urgency, in line with (Carryl L Baldwin, 2011; Judy Edworthy et al., 2003). At the end of the message an explanation that the driver had vehicle control was added, as in Experiments 8 and 9. The resulting message was "*Danger! Collision Imminent. You have control!*". The message was spoken urgently by a female actor, as if a loved one was in danger, in line with (Judy Edworthy et al., 2003) and previous experiments of this thesis. It was modified to remove pauses and decrease duration. The resulting duration of the message was 2.7 *sec*, with a peak of -0.0 *dBFS* and an average frequency of 371 *Hz*. The tactile equivalent of the audio warning was a Speech Tacton delivered with the ELV-1411A Tactor, which was constructed as described in Section 6.2. The duration of the tactile warning was also 2.7 *sec*, the peak -0.0 *dBFS* and the average frequency 370 *Hz.* The visual warning was the text of the warning displayed for the duration of the utterance in Red (*RGB*(255,0,0)), as in Experiments 8, 9.

All warnings were delivered either from the driving simulator in front of the participant or from a Windows tablet to the right of the driver, as will be described below. In this way, the location of the cues was varied. The abstract and language-based warnings were presented in all combinations of the audio, visual and tactile modalities: Audio (A), Visual (V), Tactile (T), Audio + Visual (AV), Audio + Tactile (AT), Tactile + Visual (TV), Audio + Tactile + Visual (ATV). As a result 28 different cues were created, 7 cues with all modalities (A, T,

---

[24] http://www.aactechnologies.com/category/45

V, AT, AV, TV, ATV) × 2 types of Information (Abstract, Language-based) × 2 Locations (Simulator, Tablet). These warnings were evaluated in an experiment looking at reaction times and driving metrics of participants when exposed to the cues.

8.5.3. Design

Experiment 10 investigated how quickly and effectively participants would be able to resume control of an autonomous car, while distracted by a game on a tablet. A similar task to Experiment 9 was used, where a periodical transition back to driving would be enforced due to an unexpected critical event. In line with Experiment 9, the study investigated how quickly and accurately such a transition would happen and how it would affect driving metrics. However, only critical warnings were used, varying in design, and delivered from different locations. The reason was that the focus of this study was critical handovers as a result of an automation failure, on which Experiment 9 did not primarily focus. Further, there is very little research on how to design automation failure warnings, which motivated Experiments 8, 9 and 10. Investigating how the delivery of cues from a tablet versus a simulator would affect results was not addressed in Experiment 9 or in any other study on the topic. This would be essential to investigate, since delivering warnings from the area of a distracting interaction may have the potential to attract attention when a vehicle is autonomous and focus is not on the road. As a result, a 7×2×2 within subjects design was used, with Modality, Information and Location as the independent variables and Response Time (ResT), Response Accuracy (ResA) and Lateral Deviation after Handover (LDaH) as the dependent ones. As in Experiment 9, ResT would be a measure of alertness when resuming driving, RA would indicate any missed responses and LDaH would show the level of distraction when resuming driving (lower LDaH would indicate lower distraction, see (Lindgren et al., 2009; Y. C. Liu, 2001)).

The expectations forming the hypotheses of Experiment 10 were firstly that the modalities used in the warnings would affect responses. As in Experiments 8 and 9, multimodal warnings were expected to be more effective that unimodal ones, while the visual displays on the simulator were expected to be problematic. In terms of Information, in line with Experiments 6 and 7, it was expected that abstract cues would create quick responses, while language-based ones would affect driving less. Finally, varying the location of the cues by delivering cues also from the tablet was expected to affect responses positively, since they would be closer to the participants' field of view, in line with (Miller et al., 2015). As a result, there were the following hypotheses:

- ResT will be influenced by Modality (H_{7a}), Information (H_{7b}) and Location (H_{7c});
    - Specifically, ResT was expected to decrease in multimodal as opposed to unimodal cues, abstract cues, and cues originating from the tablet.
- ResA will be influenced by Modality (H_{8a}), Information (H_{8b}) and Location (H_{8c});
    - Specifically, ResA was expected to increase in multimodal as opposed to unimodal cues, abstract cues, and cues originating from the tablet.
- LDaH will be influenced by Modality (H_{9a}), Information (H_{9b}) and Location (H_{9c}).
    - Specifically, LDaH was expected to decrease in multimodal as opposed to unimodal cues, language-based cues, and cues originating from the tablet.

8.5.4. Participants and Equipment

Twenty participants (7 female) aged between 20 and 45 years (M = 25.25, SD = 5.67) took part in the experiment. One participant had participated in Experiments 1 and 2 and one in Experiments 6 and 7. The rest had not participated in the previous experiments. There were 17 University students and 3 private employees. They had a valid driving license and between 1 and 24 years of driving experience (M = 6.18, SD = 5.50). All were right handed and reported normal vision and hearing.

The experiment took place in a University room, where participants sat in front of 27-inch Dell 2709W monitor, a PC running the driving simulator, a Microsoft Surface Pro 3[25] tablet PC running a game (placed to the right of the driver) and a Logitech G27 gaming wheel and pedals. The driving simulator software depicted a rural road scene with a curvy road and a car in front, which has been used in many studies, *e.g.* (Zhao et al., 2013). See Figure 8-8 for the setup of the experiment.

The tablet was running the Concentration memory game, used also in Experiment 9 and based on (Warnock et al., 2011) (see Figure 8-12.b). As in Experiment 9, it was a simple card matching game on a 3×8 grid. As it is likely that drivers will occupy themselves with

---

[25] http://www.microsoft.com/surface/en-us/products/surface-pro-3

other activities while an autonomous vehicle is driving itself, this task was chosen so as to decrease their engagement with driving and create a more challenging handover.

Three sounds were added to the game, so as to increase auditory distraction, aside to the visual and cognitive distraction that was added by the game. The first sound was a 100 *msec* long 440 *Hz* tone (musical note A_4) that sounded every time the participant touched the tablet screen. The second sound was a 100 *msec* 330 *Hz* tone (musical note E_4) that sounded every time a pair of pictures revealed was not a match. The third one was an Earcon with three tones (100 *msec* of 262 *Hz* followed by 100 *msec* of 330 *Hz* followed by 100 *msec* of 392 *Hz* – musical notes C_4, E_4, G_4). This sounded every time a pair of cards was matched. In this way, an additional sensory load was created in the audio modality, which was not present in Experiment 9. Ecological validity was also increased, since sound effects are frequently found in games.

Auditory cues and game sounds were displayed through three Betron portable speakers[26], one located behind the simulator monitor (one for the Simulator warning location) two behind the tablet (for the Tablet warning location and one for the game sounds). Tactile cues were displayed through a wristband on both of the participants' hands. The right hand was used for the Tablet location, since it was the hand interacting with the tablet and the left hand for the simulator location, being the hand remaining on the steering wheel. Pilot studies showed that this mapping was clear to participants and they were also familiarized with it during training with the cues (see below). Visual abstract cues were displayed through Red circles that flashed in the top central area of the monitor (for the Simulator location, see Figure 8-12.c) or the tablet (for the Tablet location, see Figure 8-12.b), and were sized 400×400 pixels (about 12×12 *cm* for the monitor and 5×5 cm for the tablet). Visual language-based cues used Red text displaying the words from the speech warning, which appeared once and for as long as the warning was uttered in the top central area of the screen, and was sized 228×700 pixels (about 17×7 *cm* for the monitor and 7×3 *cm* for the tablet, see Figure 8-12.d, 8-12.b). The visual cues did not obstruct the lead car on the monitor or the game on the tablet.

---

[26] https://www.betronstore.co.uk/portable-wireless-bluetooth-speakers/betron-pop-up-portable-travel-speaker-black.html

8.5.5. Procedure

After being welcomed and explained the experimental procedure, the 28 cues were displayed in a random order to participants for familiarization. For each cue, they could either repeat it or go to the next one when they were familiar with it. Afterwards, they were presented with the driving simulator software and the game to familiarize themselves. In the main experiment, as in Experiment 9, participants were asked to focus on the game, unless interrupted by a warning. They were able to use their right hand to play the game on the tablet, which was placed on a stand to the right of the simulator. This would be a standard setup for left-hand drive car. If all cards in a grid were matched, the game would reload with a new set of cards chosen randomly. While playing the game, the car was in autonomous mode in the centre of the lane at a speed of around 60 *mph*. The car simulated Level 3 Automation (see NHTSA (National Highway Traffic Safety Administration, 2013)) not requiring continuous intervention, but expecting availability for occasional control (see Figure 8-12.a for a screenshot of the simulator in autonomous mode).

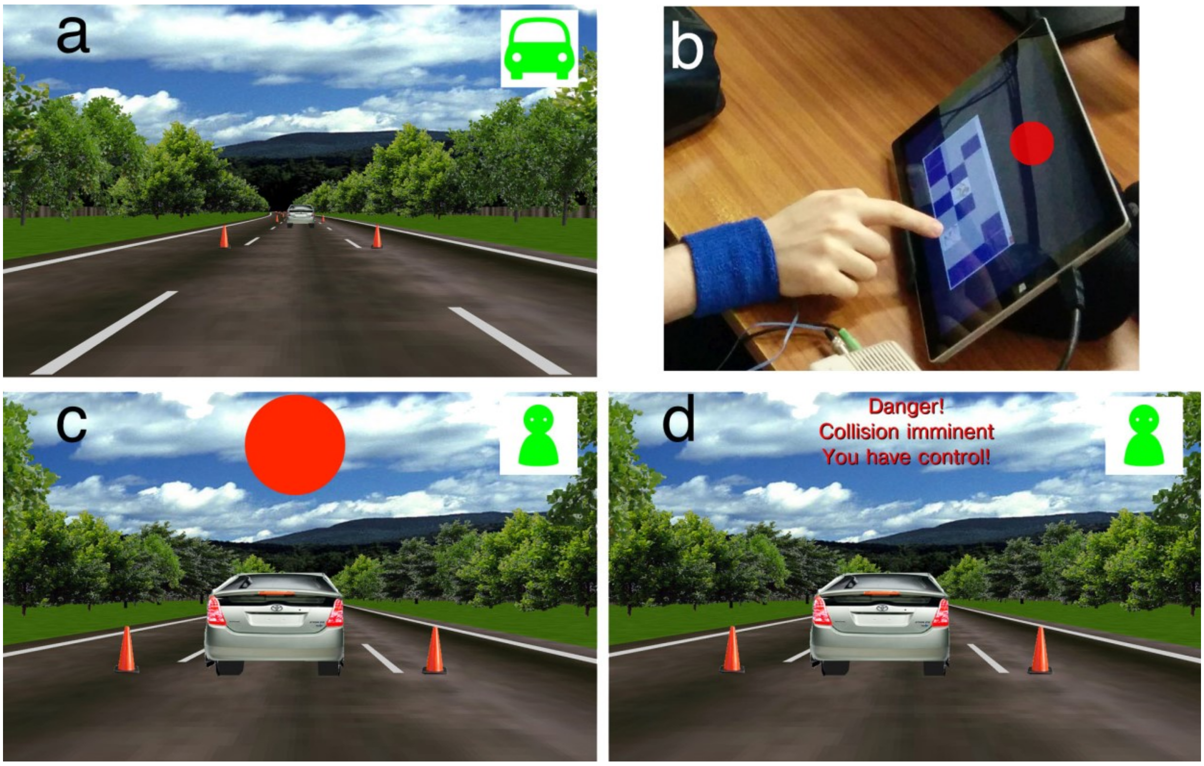


Figure 8-12: The driving simulator with the participant's car in autonomous mode, as indicated visually on the top right of the screen, and the car in front driving at a safe distance (a). The tablet game with some pairs already matched, indicated in grey (b). The handover situation, where the car in front brakes suddenly and the automation fails on the same time. In this case control is handed to the driver, as indicated visually on the top right of the screen (c,d) This handover is signified through an abstract warning (the visual warning is depicted in c) or a language-based warning (the visual warning is depicted in d).

At random intervals of any integral value between (and including) 27–32 *sec* (in line with Experiment 9) a warning was presented. In this case, control was passed to the driver (see Figures 8-12.c,8-12.d). This simulated an automation failure the vehicle could not correct and therefore a switch to manual mode was needed. To create a more critical situation, the car in front started braking at the same time as the presentation of the warning, as in situations signified with L_H warnings in Experiment 9. Participants were then handed control and were instructed to brake immediately with their right foot and return to safe driving. Once the participant braked, the car in front would advance away from the participant's car.

To manage experimental length, all abstract warnings were presented in one block of the experiment and all language-based ones in another, with the order of blocks counterbalanced across participants and with a small break between them. Each warning was presented twice in each block, resulting in a total of 56 presentations for both parts (7 Modalities × 2 types of Information × 2 Locations × 2 presentations). When back to driving, participants were able to steer using the wheel for 10 *sec* (there was no need to use the accelerator pedal). During this period, they were asked to stay in the centre of the lane. After 10 *sec*, the car automatically took over from the participant, initiating the next trial. On the top right of the screen, a car icon would be displayed when the car was in autonomous mode or a person icon for manual mode (see Figure 8-12.a, 8-12.c, 8-12.d).

Response time (ResT) was calculated from the onset of a stimulus until the participant pressed the brake pedal. If participants did not respond to a cue, their response accuracy (ResA) was 0, otherwise it was 1. Their Lateral Deviation after Handover (LDaH) was the RMSE of their lane position values, logged for 10 *sec* after the onset of a stimulus and start of the braking event of the lead car. As in Experiment 9, the value of 10 *sec* was chosen since it has shown to be an adequate time to come back to driving in handover situations (Merat et al., 2014). The experiment lasted about 45 minutes and participants were then debriefed and paid £6.

8.5.6. Results

8.5.6.1. Response Time

The data of one participant were excluded due to software issues. For the rest of the participants there were 1064 trials in total. If participants did not respond to a cue (which was the case in 14 trials – 1.3%), their ResT was adjusted to the maximum available time to respond, 10 *sec*, to allow a three factor ANOVA analysis.

Data for ResT were analysed using a three-way repeated measures ANOVA, with Modality, Information and Location as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis*** $\boldsymbol{H_{7a}}$***:*** There was a significant main effect of Modality ($F(2.11,78.19) = 34.95$, $p < 0.001$). Contrasts revealed that V caused slower responses compared to all other modalities, see Figure 8-13.a ($F(1,37) = 27.42$, $r = 0.65$, $p < 0.001$). Further, AV, AT, ATV and A created quicker responses compared to T and V ($F(1,37) = 17.45$, $r = 0.57$, $p < 0.001$), but not compared to TV. As a result, H_{7a} was accepted. ***Hypothesis*** $\boldsymbol{H_{7b}}$***:*** There was a significant main effect of Information, indicating that abstract cues caused faster responses than language based-ones (Mean value of ResT for Abstract cues: 1.11 *sec*, Mean value for Language-based cues: 1.42 *sec*, $F(1,37) = 20.50$, $r = 0.60$, $p < 0.001$). As a result, H_{7b} was accepted. ***Hypothesis*** $\boldsymbol{H_{7c}}$***:*** Finally, there was a significant main effect of Location, indicating that warnings from the tablet caused faster reaction times than simulator (Mean value of ResT for Simulator location: 1.35 *sec*, Mean value for Tablet location: 1.18 *sec*, $F(1,37) = 12.62$, $r = 0.50$, $p < 0.01$). As a result, H_{7c} was accepted. See Figure 8-13.a for values of ResT across modalities and Table 8-7 for the pairwise comparisons of ResT across Modalities.

	AV	AT	ATV	A	TV	T	V
AV		.845	.576	.689	**.004**	**.001**	**.000**
AT	.845		.721	.700	**.002**	**.001**	**.000**
ATV	.576	.721		.822	**.009**	**.002**	**.000**
A	.689	.700	.822		.064	**.000**	**.000**
TV	**.004**	**.002**	**.009**	.064		.134	**.000**
T	**.001**	**.001**	**.002**	**.000**	.134		**.000**
V	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 8-7: Pairwise comparisons between modalities for Response Time (H_{7a}). Modalities are sorted by their mean values of ResT. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Modality and Information ($F(1.87,69.13) = 21.04$, $p < 0.001$), indicating that the disadvantage of the V modality was stronger in language-based warnings ($F(1,37) = 22.11$, $r = 0.61$, $p < 0.001$). There was a significant interaction between Modality and Location ($F(2.19,81.06) = 23.14$,

$p < 0.001$), indicating that T warnings created quicker responses when coming from the simulator ($F(1,37) = 7.59$, $r = 0.41$, $p < 0.01$), while the observed disadvantage of V warnings was stronger when coming from the simulator ($F(1,37) = 33.25$, $r = 0.69$, $p < 0.001$). There was a significant interaction between Information and Location, indicating that the observed disadvantage of language-based cues was stronger when coming from the simulator compared to the tablet ($F(1,37) = 28.30$, $r = 0.66$, $p < 0.001$). Finally, there was an interaction between Modality, Information and Location ($F(2.35, 87.03) = 19.99$, $p < 0.001$), indicating that when coming from the simulator, the language-based V cues showed a disadvantage compared to TV cues, while when coming from the tablet the abstract V cues showed an advantage compared to TV ones ($F(1,37) = 30.54$, $r = 0.67$, $p < 0.001$). See Figure 8-13.c for the interaction between Modality and Information and 7-13.e for the interaction between Modality and Location for ResT.

8.5.6.2. Response Accuracy

Data for ResA were treated as dichotomous and analysed with Cochran's Q tests. ***Hypothesis H_{8a}:*** It was found that modality V was less accurate compared to AT, AV, TV and ATV ($Q(1) = 11.00$, $p < 0.01$) and also compared to T ($Q(1) = 9.00$, $p < 0.01$) and A ($Q(1) = 10.00$, $p < 0.01$). The observed values of RA for Modality were as follows: A: 99.3%, T: 98.7%, V: 92.8%, AT: 100%, AV: 100%, TV: 100%, ATV: 100%. ***Hypothesis H_{8b}:*** Abstract cues were more accurate than language-based ones (RA for Abstract: 99.4%, Language-based: 97.9%, $Q(1) = 8.00$, $p < 0.01$). ***Hypothesis H_{8c}:*** Finally, cues delivered through the tablet were more accurate than the simulator (RA for Simulator: 97.9%, Tablet: 99.4%, $Q(1) = 8.00$, $p < 0.01$). As a result, H_{8a}, H_{8b} and H_{8c} were accepted.

8.5.6.3. Lateral Deviation after Handover

There were 1120 trials for LDaH, since no data were excluded for this metric. Data for LDaH were analysed using a three-way repeated measures ANOVA, with Modality, Information and Location as factors. Due to sphericity violations, Degrees of Freedom were corrected using Greenhouse–Geisser estimates. ***Hypothesis H_{9a}:*** There was a significant main effect of Modality ($F(1.78,69.37) = 13.83$, $p < 0.001$). Contrasts revealed that V warnings created higher LDaH values compared to all other modalities ($F(1,39) = 16.76$, $r = 0.55$, $p < 0.001$). As a result, H_{9a} was accepted. ***Hypothesis H_{9b}:*** There was a significant main effect of Information, revealing that language-based warnings created higher LDaH than abstract

(Mean value of LDaH for Abstract cues: 0.94 *m*, Mean value for Language-based cues: 1.23 *m*, $F(1,39) = 7.03$, $r = 0.39$, $p < 0.05$). As a result, H_{9b} was accepted. ***Hypothesis H_{9c}:*** There was a significant main effect of Location revealing that, when warnings were coming from the tablet, LDaH was lower compared to when coming from the simulator ((Mean value of LDaH for Simulator location: 1.20 *m*, Mean value for Tablet location: 0.97 *m*, $F(1,39) = 10.18$, $r = 0.45$, $p < 0.01$). As a result, H_{9c} was accepted. See Figure 8-13.b for values of LDaH across modalities and Table 8-8 for the pairwise comparisons of ResT across Modalities.

	AV	ATV	AT	TV	A	T	V
AV		.512	.387	**.029**	.151	.059	**.000**
ATV	.512		.641	.075	.233	.058	**.000**
AT	.387	.641		.416	.481	.180	**.000**
TV	**.029**	.075	.416		.805	.236	**.000**
A	.151	.233	.481	.805		.265	**.000**
T	.059	.058	.180	.236	.265		**.000**
V	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	

Table 8-8: Pairwise comparisons between modalities for Lateral Deviation after Handover (H_{9a}). Modalities are sorted by their mean values of LDaH. The significance (*p*) values are reported after Bonferroni corrections and are in bold when they denote statistical significance.

Interactions between main effects: There was a significant interaction between Modality and Information ($F(1.63,63.73) = 12.01$, $p < 0.001$), revealing that the observed disadvantage of V warnings was mainly present in language-based warnings ($F(1,39) = 12.11$, $r = 0.49$, $p < 0.01$). There was a significant interaction between Modality and Location ($F(1.53,59.60) = 11.89$, $p < 0.001$), revealing that the observed disadvantage of V warnings was mainly present when they were coming from the simulator ($F(1,39) = 13.02$, $r = 0.50$, $p < 0.01$). There was a significant interaction between Information and Location, revealing that the observed disadvantage of language-based warnings was mainly present when they were coming from the simulator ($F(1,39) = 16.25$, $r = 0.54$, $p < 0.001$). Finally, there was a significant interaction between Modality, Information and Location ($F(1.76,68.80) = 11.57$, $p < 0.001$), revealing that while for warnings coming from the simulator, modality V created higher LDaH for language-based warnings, when warnings were coming from the tablet the disadvantage was mainly present in language-based T cues ($F(1,39) = 14.92$, $r = 0.53$, $p < 0.001$). See Figure 8-13.d for the interaction between Modality and Information and 7-13.f for the interaction between Modality and Location for LDaH.

Finally, in terms of the game performance, the results of the tablet game were as follows: 142.76 *sec* mean time to complete one game, 0.45 Clicks per Second and 0.27 Superfluous Views per Click. These are similar to (Warnock et al., 2011) and to Experiment 9, showing that participants were attentive to the game and confirming the demanding nature of this task, making it a good choice for use in driving experiments.

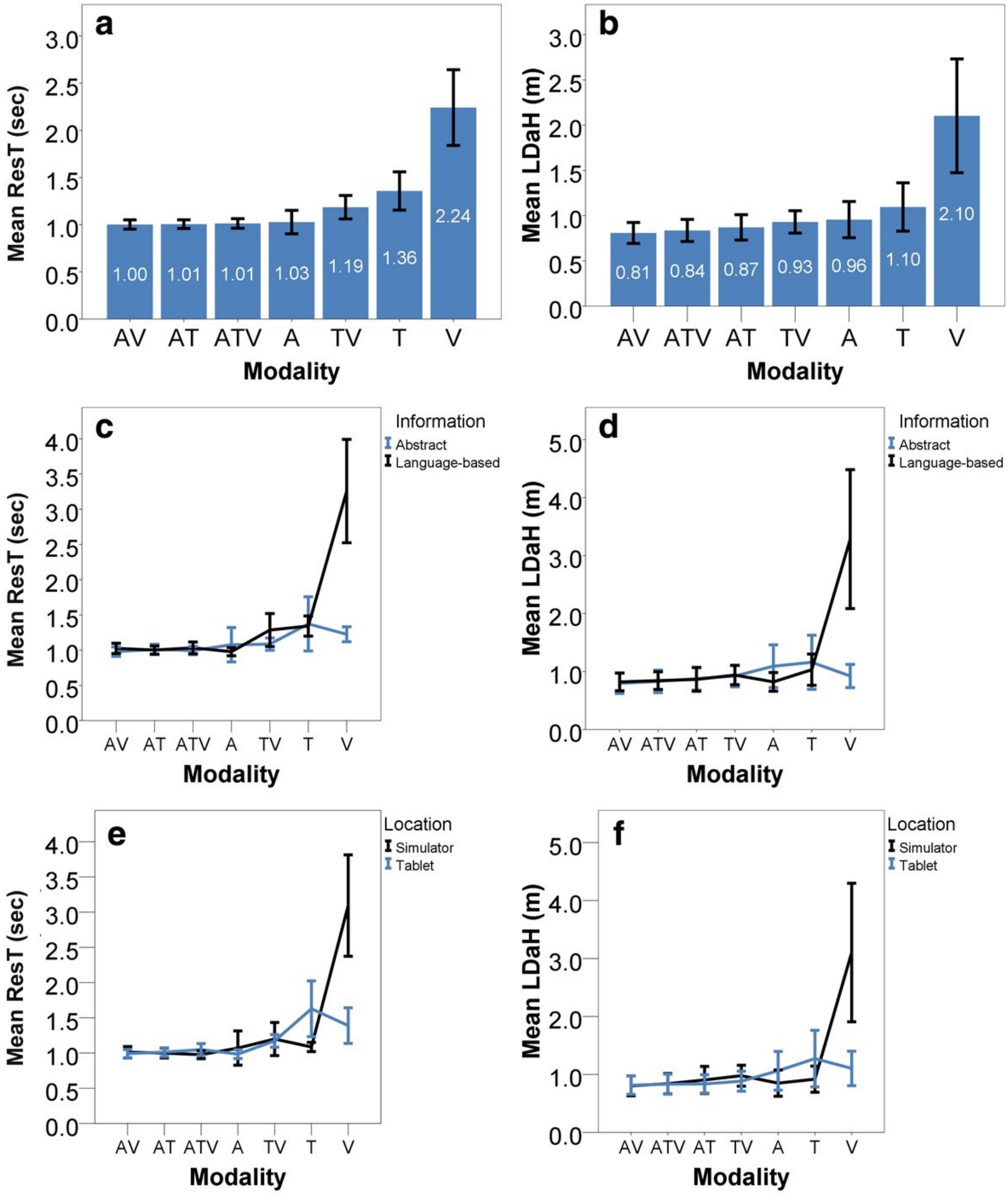


Figure 8-13: (a) Mean Response Time (ResT) and (b) mean Lateral Deviation after Handover (LDaH) across Modalities (H_{7a}, H_{9a}). (c) The interaction between Modality and Information for ResT (H_{7a} – H_{7b}) and (d) for LDaH (H_{9a} – H_{9b}). (e) The interaction between Modality and Location for ResT (H_{7a} – H_{7c}) and (f) for LDaH (H_{9a} – H_{9c}). Error bars indicate 95% confidence intervals.

8.6. Discussion for Experiment 10

The results of Experiment 10 confirmed the observed limitation of visual language-based cues coming from the simulator in Experiment 9. These cues had created the longest response times, the least accurate responses and disturbed driving metrics the most. However, the intervention of this study, *i.e.* moving the cues from the simulator to the tablet and adding abstract cue designs, positively influenced metrics and addressed the problem with visual cues in Experiment 9 (mean response times to visual cues were reduced from 6.91 to 2.24 *sec*).

A notable difference to Experiment 9 is the lower handover times observed in this study. This could be partly because all handovers were critical, requiring imminent attention. Another reason could be the simplicity of the task, which, in contrast with Experiment 9, was always the same and did not involve different types of responses. The order of modalities in terms of average response times (H_{7a}) was similar to Experiment 9, which, in combination with the better performance of cues including audio compared to unimodal tactile and visual cues, increases confidence in the advantage of audio cues for signifying handovers in autonomous cars. This extends the findings of (Naujoks et al., 2014) by providing a more elaborate examination of warning modalities for this situation. It also introduces an extensive set of possible cues to be used as warnings during an automation failure, extending the case presented by (Christian Gold & Bengler, 2014).

In terms of handover times across warning designs (H_{7b}), language-based warnings showed a disadvantage, which was mainly observed in the simulator location and the visual modality (interaction between Modality and Information, $H_{7a} – H_{7b}$ Modality and Location, $H_{7a} – H_{7c}$, and Information and Location, $H_{7b} – H_{7c}$). This arguably confirms the findings of Experiment 9 and once again shows that the problem with the visual warnings was ameliorated by moving them to the area of the gaming interaction. This also extends findings of Experiment 7, where abstract and language-based cues showed similar performance in a critical task. In Experiment 10, these cues were delivered from different locations in simulated driving. It was found that Abstract and Language-based cues are equally effective when coming from the game location, while language-based ones present limitations when delivered away from it. This extends findings of (Naujoks et al., 2014; Telpaz et al., 2015; Walch et al., 2015) by investigating a much wider set of modalities to inform about imminent handovers. As a guideline, in a vehicle where the drivers could be inattentive to the road but still expected to intervene periodically, it would be essential to capture their visual attention. Achieving this by interrupting the game on the tablet showed good results in this study, since the saliency of the cues increased. Abstract cues also showed a possibility of capturing attention when delivered from the simulator, possibly because of their pulsating design. Investigating this further by using eye-tracking techniques would be promising. It is noted that, as in previous experiments, Language-based cues were slightly longer and this might have created an advantage for abstract cues. However, reactions were required immediately for all cues, and similar results were achieved for both designs when coming from the tablet.

A further comment related to the location used for informing drivers is that tactile messages delivered on the hand that was interacting with the tablet showed a disadvantage compared to the hand that was assigned for simulator cues (interaction between Modality and Location, H_{7a} – H_{7c}). Additionally, when tactile messages were delivered in combination with visual ones, the bimodal presentation was beneficial when coming from the simulator for language-based cues, but problematic when coming from the tablet for abstract cues (interaction between all three factors, H_{7a} – H_{7b} – H_{7c}). The fact that combining visual and tactile modalities for language-based warnings showed an advantage when coming from the simulator, possibly reveals that this bimodal presentation may have been clarifying the message content which was not salient enough when delivered only visually, as in Experiment 9. In contrast, the limitation of unimodal tactile presentation from the tablet could reveal unfamiliarity of this type of warning, since a novel location for vibration was chosen, even though participants were trained with these cues until they felt confident with them. Future studies could experiment on different locations for tablet vibrations, *e.g.* the finger, and with more extensive training.

The limitations of bimodal tactile and visual presentation from the tablet could also reveal a high cognitive load when being occupied with a non-driving task while still being expected to periodically return to driving. Similar effects were observed in Experiment 3 and in the visual modality when combined with other modalities. Since this study was a more demanding one, requiring attention to both the road and a game, seems to have created this effect of increased attentional demand. In line with Experiment 3, the use of a limited number of modalities in warnings is suggested unless the event to be signified is critical. Even when critical, when a warning is delivered through a tablet, a preference for audio and visual modalities is advised.

The results of ResA showed that the visual modality created the least accurate responses, in line with Experiment 9 (H_{8a}). Abstract cues (H_{8b}) and cues coming from the tablet (H_{8c}) created more accurate responses, possibly indicating the advantage of adding a new cue design and cue location compared to Experiment 9. This further supports the guideline of using the area where interaction takes place to warn the drivers of imminent events, as well as an abstract urgent cue design. It can also inform designs of studies like (Krome et al., 2015; Terken et al., 2013), by combining a gaming interaction in the car with more critical interventions. It is worth noting that, although the described results are significant, the overall ResA is much higher than in Experiment 9 (1.3% of responses were inaccurate here, as opposed to 9.4% in Experiment 9). This improvement could be attributed to the new cue design and cue location, but also to the simpler nature of the response task. Further, the added auditory distraction added by the game sounds did not appear to be problematic for the recognition of the cues, indicating that their saliency was still high.

Results of LDaH confirm the observed disadvantage of the V warnings found in the reaction time analysis (H_{9a}), which is also in line with (Naujoks et al., 2014). The disadvantage was stronger in the simulator and language-based condition, as in ResT (H_{9b}, H_{9c}, interaction between Modality and Information, H_{9a} – H_{9b}, Modality and Location, H_{9a} – H_{9c}, and Information and Location H_{9b} – H_{9c}). This again indicates the benefit of this new setup, which improved LDaH, and thus reduced driver distraction during critical events requiring intervention. When coming from the tablet, language-based tactile cues showed a limitation in terms of LDaH (H_{9c}, and interaction between all three factors, H_{9a} – H_{9b} – H_{9c}), which is in line with the slower responses observed in ResT. This highlights the caution needed when using speech Tactons unimodally, also observed in previous experiments. It is stressed that in the few cases where there was an absence of response, the effects would be catastrophic. This is because the vehicle would be uncontrolled, as the automation failure would have disabled autonomous driving and the enforced handover would have been missed by the driver. Warning designers should aim to eliminate such cases by creating salient handover warnings that will be noticed by drivers.

8.7. General Discussion

A general comment resulting from the experiments in this chapter, is that an effective design of handover warnings for autonomous cars can help overcome distraction that is expected to increase when driving becomes less frequent. Therefore, envisioning scenarios in which such handovers may happen and the appropriate warnings for them is argued to be essential. Further, there is potential in warning drivers not only using the conventional methods available in cars, but also at the area of attention focus. In the studies in this chapter, this was a tablet, but one can easily imagine other locations away from the central area of attention, such as the car centre stack. Synchronising these devices with the car warning mechanism would increase saliency of warnings and enable drivers to return to driving promptly in an autonomous car. If this is not possible, the positive results observed with abstract warnings coming from the simulator show benefit in using multimodal messages to capture peripheral attention in critical situations. The saliency of cues including audio could also be used, by combining visual and audio warnings in the area of attention focus in critical cases. Future work should use shorter speech messages conveying handovers and investigate warning effectiveness will improve.

To explore further locations, future work should also explicitly compare the presentation of warnings on mobile devices versus on the centre stack. Systems such as Apple CarPlay (Apple, 2015) and Android Auto (Google, 2015a) are gaining popularity with users and car manufacturers. These systems link mobile devices to car systems so could potentially capture and display car warnings and messages on phones or tablets in the car. This could be on a device used by the driver, or even devices used by other passengers, that might be connected to the car. The findings of this chapter are relevant to these applications, as well as to app designers who consider an autonomous car driver as a possible part of their user group.

8.8. Conclusions & Statement of Findings

This chapter answers the research question of *"How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?"* A set of scenarios requiring a handover of control was envisioned and a set of multimodal warnings for these scenarios was designed. Experiment 8 provided subjective and Experiment 9 objective results for these warnings. This was in line with experimental methodology applied in previous experiments of this thesis, investigating manual driving scenarios. It was found that drivers clearly identified the urgency of the cues and rated multimodal cues as more urgent and more effective. Unimodal audio and visual cues were rated as less annoying but less effective than multimodal ones. When evaluating the handover time and accuracy for distracted drivers to take control of a vehicle, participants were occupied with a tablet game. This is an activity very likely to occur as drivers become less engaged on the road and driving requires less involvement. It was found that they were quicker to resume driving when warned with multimodal cues of high designed urgency. Unimodal visual cues were especially poor since they did not attract drivers' attention back to the road. The use of multimodal informative cues for critical handovers was therefore suggested and the use of unimodal visuals for such a case is not advised.

Experiment 10 presented a study that focused on critical handovers in an autonomous car. Handovers were signified to distracted participants by multimodal combinations of abstract or language-based cues. Delivering the warnings with abstract cues including audio and visuals from the area of the game captured attention more effectively when signifying a handover of control. Therefore, the utilisation of this area when a driver is distracted in an autonomous vehicle is suggested. Since in a real driving situation there may or may not be a side task, the synchronization of mobile devices used by the driver with the autonomous vehicle is suggested. In this way, warnings and notifications from the car can be presented where the driver's attention is focused, increasing warning saliency.

In summary, the following guidelines can be derived from this chapter:

- When designing autonomous handovers, it is important to consider the situation requiring a transfer of control. Urgency of the situation can vary, and it can be effectively recognised through warnings designed for the appropriate handover situations;
- Handovers by drivers are more rapid when signified by critical messages in critical situations. Drivers seem to recognise that their intervention is essential in these cases and this improves their responses;
- When attention is not on the road, unimodal visuals originating from the direction of the road in front should not be used for autonomous handovers. This strongly compromises driving safety and transition time and accuracy;
- Using the area where a distracting interaction takes place to warn the drivers of imminent events, as well as an abstract urgent cue design can be effective in capturing attention

when not driving. Warnings originating from the area of a distracting interaction are more salient, and therefore drivers return to driving quicker when exposed to them.

This chapter presented a series of experiments concluding this thesis, where the experimental methodologies and the warnings designs applied in Experiments 1 – 7 were adjusted and put to the test for an autonomous driving scenario. The topic of autonomous car handovers stands aside thematically to the topics examined in the previous experiments of this thesis, is however a critical one, with little investigation in terms of effective warning designs to signify it. Therefore, the experimental framework established in previous experiments was utilised in Experiments 8, 9 and 10, investigating the influence of modality, urgency and situation to responses to multimodal driver warnings in autonomous driving handovers, examining both subjective and objective measures when measuring responses to these warnings. In this way, new guidelines on how to effectively design multimodal warnings for this scenario were presented, resulting to a more complete body of work presented by this thesis. As a result, warnings for both manual and autonomous driving scenarios are evaluated, and more complete guidelines are presented.

9. Conclusions

This thesis investigated the use of multimodal displays for drivers with a series of experiments looking into driver responses to the cues designed. The cues used were multimodal, since they utilised all combinations of audio, visual and tactile modalities of various designs. They also varied in urgency, to address scenarios varying in criticality. Starting with a simple set of abstract cues consisting of repeated pulses, responses of participants when exposed to these cues were investigated. These cues were then delivered in varying contexts of situational urgency, to assess their effectiveness both in the presence and the absence of a critical event on the road. Utilising speech in multimodal cues, a set of vibrational warnings deriving from speech (Speech Tactons) were designed. In this way an evaluation of multimodal speech-related cues varying in urgency was possible. The two different designs were then compared to each other, and the effectiveness of abstract versus speech related multimodal warnings was assessed. Finally, these cue designs were used in an autonomous car scenario, to signify handovers of control in such a vehicle. The envisioned handover scenarios were investigated and the utility of truly multimodal warnings varying in urgency and message content was investigated.

9.1. General Discussion of Experimental Results

Based on the findings of this thesis' experiments, a general discussion will be presented in this section. For a summary of this discussion, see Table 9.2 in Section 9.6. A primary observation was the influence of interpulse interval to the perceived urgency of abstract cues (Experiments 1, 2, 3, 6, 7, 9). Findings of studies like (Judy Edworthy, Loxley, & Dennis, 1991a; Pratt et al., 2012) were confirmed multimodally and in all combinations of audio, visual and tactile cues. These findings provide a tool to driver warnings designers in order to create cues that convey varying degrees of urgency, in a relative simple way, *i.e.* by decreasing interpulse interval to increase expected perceived urgency. This expectation holds in all modalities, which affords flexibility of cue design, based on currently occupied resources, following the pointers of Wickens (C. D. Wickens, 2008), whereby delivering cues in a modality other than the one occupied may reduce workload. Guidelines by Stanton (Stanton, 1994) on the expected benefit of matching the criticality of an alarm to the criticality of the signified event were also confirmed.

In terms of language-based warnings, it was found that the urgency of the message utterance can be used as a design parameter to vary their perceived urgency multimodally (Experiments 4, 6, 7, 8, 9). This was a further extension of the guidelines by studies like (C. L. Baldwin & Moore, 2002; Elizabeth Hellier et al., 2002), in cues of multiple modalities. Varying expected perceived urgency by using the urgency of the message utterance is also a relatively easy process, which was achieved in all cases by simple instructions to the actors recording the warnings. In combination with interpulse interval for abstract cues, this thesis offers warning designers a set of guidelines on how to vary expected perceived urgency in two cue designs, *i.e*. abstract and language-based, and in all modality combinations of audio, visual and tactile warnings. In this way, the semantic association to the signified event can be controlled, depending on how essential it is to disclose accurate information on the event (affording the use of language-based warnings), and how obvious is the event is to the driver (affording the use of abstract warnings). The flexibility of cue design with a set of modalities to choose from also holds for language-based warnings, for reasons of effectively managing resources utilised for a task, as described above. An advantage of abstract cues in critical situations, observed in Experiments 6, 7 and 10, is also worth noting, although future work should investigate whether this advantage will hold with shorter language-based warnings.

An important finding, complementing the observations of perceived urgency, is that warnings designed to be urgent also elicited quicker reactions (Experiments 2, 3, 6, 7, 8, 9). This is a crucial and desired quality of an alarm, and one of the main purposes for which they are used in critical scenarios (Stanton, 1994). In order to maintain Situational Awareness, it is argued that an alarm needs to convey the appropriate information on how essential an imminent reaction is. Although not directly assessing Situational Awareness in this thesis' experimental work, the fact that that more critical alarms were both rated as such (observed through subjective measures) and created appropriately quick reactions (observed through objective measures) was encouraging in terms of their suitability as driver warnings.

On a similar note, it was found that alerts of high designed urgency further improved the speed of reactions when delivered multimodally in the presence of critical events (Experiments 3, 7, 9, 10). This increased confidence on the effectiveness of using multiple modalities in critical alarms. This was rarely investigated in prior literature (Bridget A. Lewis et al., 2013) and never in this number of modalities. It is argued that the saliency created with multimodal cues increases their effectiveness in attracting attention (as was speculated by Stanton (Stanton, 1994)), while orienting attention to the task at hand, even if some attentional resources are occupied. This also relates to the investigation performed in Experiments 9 and 10, where moving the cues in the area of the distracting interaction showed a clear advantage, evidently attracting attention more effectively. Taken together, this thesis' findings extend results of studies like (Cristy Ho, Tan, et al., 2005, 2006; Bridget A. Lewis et al., 2013; Scott & Gray, 2008) using a multimodal warning design, and provide cues that elicit appropriately quick reactions, especially when delivered multimodally. This is an important design tool when designing critical alerts, considering however that this will also increase their perceived annoyance (see next paragraph).

Investigating perceived annoyance, it was found that it increased along with perceived urgency, and along with number of modalities used in alarms (Experiments 1, 4, 8). This finding can be viewed in combination with the finding that warnings signifying more urgent events were perceived as more annoying, but also as more effective (Experiments 4, 8). As discussed in the experimental chapters, participants showed a tolerance for cues that could be more annoying, as long as they would signify events worth being informed about, *i.e.* urgent events. This extends findings of studies like (C. L. Baldwin & Moore, 2002; D. C. Marshall et al., 2007), showing how alarms are perceived as more effective if they are urgent, using an evaluation in a multimodal context. It provides warning designers with the guideline that warnings should not be over-utilised for events less worthy of attention. The need to be economical in the use of warnings and the use of multiple modalities, unless an imminent event needs to be signified, was a guideline deriving from this thesis. As also noted by Bliss & Acton (Bliss & Acton, 2003), invasiveness of alarms needs to be minimal, and this thesis confirmed this guideline in multimodal warnings varying in urgency and message content.

As another word of caution, utilising tactile as the only cue to alert drivers also demonstrated an annoying effect (Experiments 1, 4, 5, 8), leading to the suggestion of avoiding tactile cues when they are not signifying critical events. This can be viewed in combination with the limitations observed in reactions to unimodal tactile language-based cues, the Speech-Tactons (Experiments 4, 5, 6, 8 and 9). A resulting guideline was to avoid the use of unimodal tactile cues with more complex designs, since they can be misinterpreted, while the use of unimodal tactile cues of any design can be annoying. This ties to the same discussion on annoyance as in the previous paragraph, suggesting that driver warnings need to be as seamless as possible when not signifying critical events, while their annoyance is tolerated when signifying more imminent dangers. Since the language-based tactile cues designed and evaluated in this thesis were novel, more investigation is needed on how to improve their design so as to reduce their annoyance and improve their performance when used unimodally. However, it is noted that such an investigation would be of higher value for critical contexts, since, as discussed, the tolerance for annoyance and complexity of cues is low for less critical alarms.

The following section with summarise the results of this thesis, by re-posing each research question and summarising how it was answered through the experiments conducted.

9.2. Research Question 1

How do multimodal driver displays varying in urgency affect performance?

To answer this question all multimodal combinations of Audio, Visual and Tactile modalities were utilised to design abstract warnings, consisting of repeated pulses that varied in urgency. In Experiment 1 (Chapter 4), it was found that the cues were clearly identified in terms of perceived urgency, while perceived annoyance was not high, indicating the appropriateness of the cues for the driving context. Experiment 2 (Chapter 4) found a similar reduction of recognition times for highly urgent cues. These findings extended available results in all modality combinations used and in the context of a driving simulator. The strength of cues using the visual modality in conveying messages quickly and accurately, as well as an increased annoyance of tactile cues for warnings were also highlighted. In addition, more modalities meant quicker and more accurate responses, as well as higher perceived urgency, without a large increase in perceived annoyance.

9.3. Research Question 2

How does situational urgency influence responses to multimodal driver displays varying in urgency?

To answer this question, the effects of varying situational urgency on the response times, lateral deviation and steering angle of participants in a simulated driving task were investigated in Experiment 3 (Chapter 5). The set of multimodal warnings varying in urgency used in Experiments 1 and 2 was also used in Experiment 3. Three situations were simulated: a car braking without warnings, warnings without a car braking and both simultaneously. The results showed a clear reduction in response times to warnings when the critical event in the driving scene occurred at the same time as a critical warning. Quicker responses were observed when responding to bimodal and trimodal warnings compared to unimodal ones and to warnings of high urgency compared to medium and low urgency. Further, the use of visual warnings slowed responses in the critical situation, providing evidence of high load in the visual modality. This effect was also observed in lateral deviation and steering angle values, where the benefit in driving metrics when there were either warnings or a critical event, was not present when the event arose together with the warnings. These results extended knowledge of in-car warning design by identifying the effect of situational urgency on participant response times as well as driving metrics. They also verified the benefit of using multimodal displays of varying designed urgency to alert drivers in a context of varying situational urgency, a case not previously simulated. Finally, the evidence of high visual load during a critical event highlighted the limitation of the visual modality for warnings when encountering critical events in the driving scene.

9.4. Research Question 3

How do multimodal driver displays varying in urgency and message content compare to each other in terms of performance?

To answer this question, a set of multimodal language-based warnings varying in urgency was first designed. These used speech and Speech Tactons, the tactile counterparts of speech warnings for drivers. Experiments 4 and 5 (Chapter 6) evaluated these new warnings. Results showed that the addition of these new cues improved subjective responses of drivers to speech warnings (Experiment 4). The warnings were clearly distinguished in terms of urgency, their annoyance was low and their alerting effectiveness changed similarly to urgency, increasing for more urgent messages and for multimodal cues. Recognition accuracy of the tactile cues' urgency was high overall and recognition accuracy of individual messages was higher for longer cues (Experiment 5). Speech Tactons were therefore suggested as an addition to speech warnings in driver alerts. With the technique created, these tactile cues can be easily designed and added to warnings that will improve drivers' responses.

Experiments 6 and 7 (Chapter 7) then used these new warnings and presented a first evaluation of responses to abstract versus language-based multimodal car warnings of varying urgency. All multimodal combinations of audio, tactile and visual warnings were evaluated in a driving simulator. Two tasks were used; a recognition task (Experiment 6), where the cues' urgency was identified with no critical event present, and a response task (Experiment 7), where responses to high urgency warnings were measured in the presence of a critical event. An advantage of abstract warnings and warnings including visuals in the recognition task was observed. Cues including audio performed better in the response task. In both tasks, multimodal cues were the best performing ones, with the exception of unimodal visuals for recognition and unimodal audio for response. Driving behaviour, although slightly worsened by all cues in the critical situation, was marginally better when using language-based cues compared to abstract ones. These results showed the benefit of using abstract cues in non-critical situations and a possible advantage of language-based cues in a critical situation.

9.5. Research Question 4

How do multimodal driver displays varying in urgency and message content affect performance during handovers of control in an autonomous car?

To answer this question, a set of possible scenarios that would require a handover of control in an autonomous car were envisioned. A set of multimodal, language-based warnings for these situations was then introduced. All combinations of audio, tactile and visual warnings for handovers were evaluated in terms of perceived urgency, annoyance and alerting effectiveness in Experiment 8 (Chapter 8). Results showed clear recognition of the warning urgency in this new context, as well as low perceived annoyance overall, and higher perceived effectiveness for critical warnings. In Experiment 8 (Chapter 8) participants were distracted from the road by playing a game on a tablet while using an autonomous car simulator. A handover of control was then simulated, by either requesting or enforcing driver intervention, depending on the criticality of the event on the road. The time of transition from self-driving to manual mode in the presence of the warnings was evaluated. Results showed quicker transitions for highly urgent warnings and poor driving performance for unimodal visual warnings. Finally, Experiment 10 (Chapter 8) presented a comparison of abstract versus language-based multimodal warnings signifying handovers due an automation failure; a rare but very critical situation for the driver to be in. Multimodal abstract or language-based warnings signifying this situation were then delivered, either from the simulator or from the tablet, to discover the most effective location. Abstract cues, cues including audio and cues delivered from the tablet improved handovers. This indicated the potential of moving simple but salient autonomous car warnings to where driver attention is focused.

9.6. Contributions

Summarising the findings of this thesis, Table 9-1 presents the guidelines resulting from each experimental chapter and Table 9-2 outlines the main contributions of this work providing a set of statements and the experiments from which they derived.

Guideline	Rationale	Priority	Chapter
1. Use interpulse interval and number of modalities to vary urgency in multimodal audio, visual and tactile abstract warnings.	Decreasing interpulse interval and increasing number of modalities increases perceived urgency and decreases recognition time of multimodal warnings.	High	4
2. Avoid the use of multimodal abstract warnings including tactile delivered on the in non-critical contexts.	Such warnings lead to slower recognition and higher ratings of annoyance.	Low	
3. Use abstract warnings of high designed urgency to signify critical events.	Using such warnings speed up reactions to critical events.	High	5
4. Use abstract warnings of medium designed urgency to signify non-critical events.	Using such warnings can provide an overall alertness, as well as improved lane keeping and steering behaviour when no critical event is present.	Medium	
5. Use Speech Tactons along with auditory language-based warnings when signifying events of high urgency.	The addition of Speech Tactons helps the recognition of warning urgency, while Perceived Urgency and Effectiveness escalate similarly in ratings, indicating that language-based multimodal warnings are more appreciated in urgent situations.	High	6
6. Avoid the use of low urgency language-based warnings.	Annoyance is higher but more acceptable for high urgency language-based warnings compared to lower urgency ones. Low urgency warnings are perceived as less effective and more annoying.	Low	
7. Use abstract warnings including visuals for low criticality driving tasks.	Such cues have quicker recognition in a low criticality task, *i.e.* recognizing warning urgency with no critical event on the road. This is because participants rely on a visual interpretation of the cues.	Medium	7
8. Use abstract or language-based multimodal cues including audio for high criticality driving tasks.	Multimodal cues including audio create quicker responses in a high criticality task, *i.e.* responding to a car in front braking sharply. Abstract and language-based cues have similar response times when a critical event is presented.	High	

Guideline	Rationale	Priority	Chapter
11. Use abstract warnings of high designed urgency to signify critical handovers of control.	Handovers by drivers are more rapid when signified by abstract warnings of high designed urgency in critical situations. Drivers recognise that their intervention is essential in these cases and this improves their responses.	High	8
12. Avoid the use of unimodal visuals to signify handovers of control.	When attention is not on the road, unimodal visuals originating from the direction of the road in front strongly compromise driving safety, handover time and accuracy.	High	
13. Use the area where a distracting interaction takes place to warn the drivers of imminent events.	This technique, as well as an abstract urgent cue design can be effective in capturing attention when not driving.	High	

Table 9-1: The guidelines presented in each experimental chapter of this thesis, with reference to the rationale of the guidelines (based on the experimental results of this thesis), the priority of the guidelines (High for safety-related, Medium for potentially safety-related, and Low for comfort-related guidelines), and the related chapter.

Contribution Statement	Experiments
Interpulse interval can modulate perceived urgency in abstract multimodal audio, visual and tactile cues.	1,2,3,6,7,9
Urgency of the message utterance can modulate perceived urgency in language-based multimodal audio, visual and tactile cues.	4,6,7,8,9
Perceived annoyance can increase with perceived urgency, and with number of modalities used in the warnings.	1,4,8
Warnings signifying more urgent events can be perceived as more annoying, but also as more effective.	4,8
Warnings of higher designed urgency can lead to quicker reactions.	2,3,6,7,8,9
Unimodal tactile warnings can increase perceived annoyance and can hinder recognition.	1,4,5,8
Using multimodal warnings rather than unimodal ones can improve reaction times to critical events of high situational urgency.	3,7,9,10
Visual warnings can hinder reactions to critical events when they do not originate from the area of visual attention. Moving visual warnings to the area of visual attention can be beneficial for reaction to critical events.	9,10
Speech Tactons can improve reactions when used together with audio or visual cues, but not alone, since they can hinder effective interpretation.	4,5,6,8,9
Abstract cues can improve reactions more than language-based ones when signifying critical events.	6,7,10

Table 9-2: Summary of the contributions of this thesis in the form of statements, with reference to the experiments supporting these statements. The experiments supporting the statements are conducted both in manual and in autonomous car scenarios.

9.7. Limitations & Future Work

This section will provide suggestions on future work based on this thesis. In summary, future work can use a non-simulated driving task, new message designs, road scenarios, user groups and modalities, as well as attempt a modelling approach to user performance.

9.7.1. Using a Real Driving Task

This thesis used a simulated driving task in all the studies performed. Although this limits ecological validity of the results, the safety critical situations simulated could not be easily reproduced in a real driving scenario. This limitation is widely acceptable in the research community, and the vast majority of the studies cited in this thesis are also simulator studies. Further, there are indications that simulated and on-road studies can produce similar results (Wang et al., 2010). Future work could make steps towards increasing the ecological validity of the results by performing on-road studies. The practical limitation of simulating critical events on-road would need to be overcome in such a case. Possible options could be relocating the simulator inside a stationary or a moving vehicle and performing critical tasks in the simulated environment (see (Ahmad et al., 2015; Beattie, Baillie, Halvey, & Mccall, n.d.) for similar examples with non-critical tasks). This would allow for a closer approximation of the environment and the ambient cues while driving, and still maintain safety. It could however introduce new limitations, such as simulator sickness (Brooks et al., 2010).

9.7.2. Using More Message Designs

Although there was extensive investigation of multimodal abstract and language-based warnings in this thesis, there are other message designs that could be investigated. Such an example is Auditory Icons, where an event is associated with a natural sound (Denis McKeown & Isherwood, 2007). This type of cue could be created multimodally, for example by using pictures of the metaphors and investigating what aspects of the cue could be conveyed with vibration. This could increase the repertoire of cues at the designers' disposal. Further, language-based cues could be made shorter in order to convey the appropriate message with less words. Their length was a possible limitation in this thesis and steps could be taken to create messages that are informative and not longer than needed (P. Green, Levison, Paelke, & Serafin, 1994). Subjective evaluations of message text could be a meaningful step towards this direction.

9.7.3. Using Richer Scenarios

The scenarios simulated in this thesis were not complex in terms of road conditions. Although a critical and a non-critical situation were clearly different to each other, there could be more variables to be varied in an on-road scenario. Such variables could be the road curvature, the traffic density, the weather conditions and the road furniture (*e.g.* signs, advertisements, unexpected obstacles, see for example (Horberry, Anderson, Regan, Triggs, & Brown, 2006)). These parameters, although studied before, they have not been examined in the presence of multimodal warnings, to assess their influence in reactions. A higher fidelity simulator or a road study would be a viable direction to address this, and is a possible future step for this research.

9.7.4. Using More User Groups

The participants used in this thesis were typically younger drivers, and there was no specific focus in recruiting diverse age groups. This could be a useful future direction, since age has shown to affect reaction in driving (Horberry et al., 2006; Kramer et al., 2007; Warshawsky-Livne & Shinar, 2002). Further, literature on warning reactions across age groups is limited, which could motivate future work in this direction. Another possible future direction could be using participants of similar ages but with different developmental characteristics. As an example, Experiments 1 and 2 of this thesis were replicated with participants on the Autism spectrum (Shim et al., 2015), and interesting results on different warning perception and reactions across two groups were revealed. This exercise could be extended to other user groups, for example drivers with learning disabilities, motor impairments or mental disorders. In this way, more diverse members of the population could be examined, possibly revealing varying effectiveness of different warning types.

9.7.5. Using New Modalities

Although this thesis examined all combinations of audio, visual and tactile modalities as warnings, new modalities could be investigated in future work. As display technology progresses, the opportunities for such new modalities increase. Examples could be thermal displays (G. Wilson, Halvey, Brewster, & Hughes, 2011) or ultrasound haptics (Graham Wilson, Carter, Subramanian, & Brewster, 2014), which have been used in previous studies out of the context of cars. Creating new multimodal displays having an association to a road event, while being adequately similar to each other is a promising research direction. It would also be interesting to discover what features could be varied multimodally to create these new displays. For example, in a thermal-ultrasonic icon, would more heat or higher ultrasound power mean higher perceived urgency? And would these parameters be perceivable when varied along with interpulse interval?

9.7.6. Modelling User Performance

As mentioned in Section 2.1, there is theoretical work on the reasons behind limited performance during high workload. There is limited work however on predicting driver performance in the presence of warnings. This thesis points to a clear improvement when using warnings to signify critical events, however there are limitations in how salient warnings can be when there is high visual load (see Experiment 3). Models of human performance predicting how effective warnings would be in different modalities, depending for example on present workload in each modality would be an interesting future direction.

9.8. Conclusions

This thesis has presented a series of experiments in multimodal driver warnings, using all combinations of audio, visual and tactile modalities, and both manual and autonomous driving scenarios. It has revealed the role of interpulse interval and message utterance as design parameters that can modulate perceived urgency of the warnings. It has considered the role of perceived annoyance in the cues, which increases as cues become more urgent. It has however shown that perceived alerting effectiveness also increases as cues become more urgent, while it decreases as they become less urgent. Combining these two findings, it has posited that warnings should be used for more critical events and avoided in less critical ones. The improved performance observed when exposed to highly urgent cues during critical events has highlighted the essential role of warnings in that context. The improved performance observed in the presence of multimodal warnings has further highlighted the saliency achieved when the same message is delivered multimodally, which is especially useful in critical situations. Finally, the warning design in terms of message content, warning modality and warning location has been examined, in order to provide guidelines on the utility of abstract versus language-based warnings, originating from the area of a distracting interaction or away from it, and using any combination of audio, visual and tactile modalities. In this way, an extensive set of guidelines for these contexts has been provided, completing the work's contribution in the design of effective multimodal driver warnings varying in urgency and message content, in both critical and non-critical situations and in both manual and autonomous driving scenarios.

Supplemental Materials

All data acquired by Experiments 1 – 10 can be found in the following link:

https://goo.gl/EFKAFi

All warnings created in this thesis can be found in the following link:

https://goo.gl/4lWH6E

Bibliography

Abdi, H. (2007). Signal Detection Theory (SDT). *Encyclopedia of Measurement and Statistics*, 886–889.

Ablaßmeier, M., Poitschke, T., Wallhoff, F., Bengler, K., & Rigoll, G. (2007). Eye gaze studies comparing head-up and head-down displays in vehicles. In *IEEE International Conference on Multimedia and Expo* (pp. 2250–2252). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4285134

Ahmad, B. I., Langdon, P. M., Godsill, S. J., Hardy, R., Skrypchuk, L., & Donkor, R. (2015). Touchscreen usability and input performance in vehicles under different road conditions. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15*, 47–54. http://doi.org/10.1145/2799250.2799284

Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis and Prevention*, *27*(5), 707–715. http://doi.org/10.1016/0001-4575(95)00026-V

ANSI. (2011). *American National Standard: Safety Colors. ANSI Z535.1-2006 (R2011).*

Apple. (2015). CarPlay. Retrieved from http://www.apple.com/uk/ios/carplay/

Aust, M. L., Engström, J., & Viström, M. (2013). Effects of forward collision warning and repeated event exposure on emergency braking. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 34–46. http://doi.org/10.1016/j.trf.2012.12.010

Baldwin, C. L. (2011). Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance. *Ergonomics*, *54*(4), 328–337. http://doi.org/10.1080/00140139.2011.558634

Baldwin, C. L., Eisert, J. L., Garcia, A., Lewis, B., Pratt, S., & Gonzalez, C. (2012). Multimodal urgency coding: auditory, visual, and tactile parameters and their impact on perceived urgency. *Work-Journal of Prevention Assessment and Rehabilitation*, *41*, 3586–3591. http://doi.org/10.3233/WOR-2012-0669-3586

Baldwin, C. L., & Lewis, B. A. (2013). Perceived urgency mapping across modalities within a driving context. *Applied Ergonomics*, *45*(5), 1270–1277. http://doi.org/10.1016/j.apergo.2013.05.002

Baldwin, C. L., & May, J. F. (2011). Loudness interacts with semantics in auditory warnings to impact rear-end collisions. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(1), 36–42. http://doi.org/10.1016/j.trf.2010.09.004

Baldwin, C. L., & May, J. F. (2014). Auditory forward collision warnings reduce crashes associated with task-induced fatigue in young and older drivers. *International Journal of Human Factors and Ergonomics 16*, *3*(2), 107–121.

Baldwin, C. L., & Moore, C. (2002). Perceived Urgency, Alerting Effectiveness and Annoyance of Verbal Collision Avoidance System Messages. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 46, pp. 1848–1852). SAGE. http://doi.org/10.1177/154193120204602215

Beattie, D., Baillie, L., Halvey, M., & Mccall, R. (n.d.). Maintaining a Sense of Control in Autonomous Vehicles via Auditory Feedback.

Biondi, F., Rossi, R., Gastaldi, M., & Mulatti, C. (2014). Beeping ADAS: Reflexive effect on drivers' behavior. *Transportation Research Part F: Traffic Psychology and Behaviour*, *25*, 27–33. http://doi.org/10.1016/j.trf.2014.04.020

Birrell, S. a, Young, M. S., & Weldon, A. M. (2013). Vibrotactile pedals: provision of haptic feedback to support economical driving. *Ergonomics*, *56*(2), 282–292. http://doi.org/10.1080/00140139.2012.760750

Blattner, M. M., Sumikawa, D. A., & Robert, M. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, *4*(1), 11–44. http://doi.org/10.1207/s15327051hci0401

Bliss, J. P., & Acton, S. A. (2003). Alarm mistrust in automobiles: how collision alarm reliability affects driving. *Applied Ergonomics*, *34*(6), 499–509. http://doi.org/10.1016/j.apergo.2003.07.003

Brandenburg, S., & Skottke, E. (2014). Switching from manual to automated driving and reverse : Are drivers behaving more risky after highly automated driving ? In *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), 2014* (pp. 2978–2983). IEEE.

Braun, C. C., & Silver, N. C. (1995). Interaction of signal word and colour on warning labels: differences in perceived hazard and behavioural compliance. *Ergonomics*, *38*(11), 2207–2220. http://doi.org/10.1080/00140139508925263

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 222–227). ACM Press. http://doi.org/http://doi.acm.org/10.1145/169059.169179

Brewster, S., & Brown, L. M. (2004). Tactons: Structured Tactile Messages for Non-Visual Information Display. In *Proceedings of Australasian User Interface Conference-Volume 28* (pp. 15–23). Dunedin, New Zealand: Austalian Computer Society.

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., Koon, B. L., … Wills, R. F. (2010). Simulator sickness during driving simulation studies. *Accident Analysis and Prevention*, *42*(3), 788–796. http://doi.org/10.1016/j.aap.2009.04.013

Brown, L. M., Brewster, S. a., & Purchase, H. C. (2006). Multidimensional tactons for non-visual information presentation in mobile devices. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (pp. 231–238). New York, New York, USA: ACM Press. http://doi.org/10.1145/1152215.1152265

Brown, L. M. M., Brewster, S. A. a, & Purchase, H. C. C. (2005). A First Investigation into the Effectiveness of Tactons. In *Eurohaptics Conference, 2005 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2005. World Haptics 2005. First Joint* (pp. 167–176). IEEE. http://doi.org/10.1109/WHC.2005.6

Brumby, D. P., Davies, S. C. E., Janssen, C. P., & Grace, J. J. (2011). Fast or Safe ? How Performance Objectives Determine Modality Output Choices While Interacting on the Move. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 473–482). ACM.

Brumby, D., & Seyedi, V. (2012). An empirical investigation into how users adapt to mobile phone auto-locks in a multitask setting. In *MobileHCI 2012* (pp. 281–290). New York, New York, USA: ACM Press. http://doi.org/10.1145/2371574.2371616

Burnett, G. E., & Porter, M. J. (2001). Ubiquitous computing within cars: designing controls for non-visual use. *International Journal of Human-Computer Studies*, *55*(4), 521–531.

http://doi.org/10.1006/ijhc.2001.0482

Cambridge Dictionary. (2016). Retrieved from dictionary.cambridge.org

Cao, Y., Mahr, A., Castronovo, S., Theune, M., Stahl, C., & Müller, C. A. (2010). Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 239–248). ACM Press. Retrieved from http://dl.acm.org/citation.cfm?id=1720004

Cao, Y., Theune, M., & Müller, C. (2010). Multimodal Presentation of Local Danger Warnings for Drivers : A Situation-dependent Assessment of Usability. In *Professional Communication Conference (IPCC), 2010 IEEE International* (pp. 226–229). IEEE.

Cao, Y., van der Sluis, F., Theune, M., op den Akker, R., & Nijholt, A. (2010). Evaluating informative auditory and tactile cues for in-vehicle information systems. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 102–109). New York, New York, USA: ACM Press. http://doi.org/10.1145/1969773.1969791

Carsten, O., Lai, F. C. H., Barnard, Y., Jamson, A. H., & Merat, N. (2012). Control Task Substitution in Semiautomated Driving: Does It Matter What Aspects Are Automated? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 747–761. http://doi.org/10.1177/0018720812460246

Chan, A. H. S., & Ng, A. W. Y. (2009). Perceptions of implied hazard for visual and auditory alerting signals. *Safety Science*, *47*(3), 346–352. http://doi.org/10.1016/j.ssci.2008.06.003

Chapanis, A. (1994). Hazards associated with three signal words and four colours on warning signs. *Ergonomics*, *37*(2), 265–275. http://doi.org/10.1080/00140139408963644

Chun, J., Han, S. H., Park, G., Seo, J., Lee, I., & Choi, S. (2012). Evaluation of vibrotactile feedback for forward collision warning on the steering wheel and seatbelt. *International Journal of Industrial Ergonomics*, *42*(5), 443–448. http://doi.org/10.1016/j.ergon.2012.07.004

Chun, J., Lee, I., Park, G., Seo, J., Choi, S., & Han, S. H. (2013). Efficacy of haptic blind spot warnings applied through a steering wheel or a seatbelt. *Transportation Research Part F: Traffic Psychology and Behaviour*, *21*, 231–241. http://doi.org/10.1016/j.trf.2013.09.014

Cruise. (2015). Meet the Cruise. Retrieved August 7, 2015, from http://www.getcruise.com/

Cummings, M. L., Kilgore, R. M., Wang, E., Tijerina, L., & Kochhar, D. S. (2007). Effects of Single Versus Multiple Warnings on Driver Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(6), 1097–1106. http://doi.org/10.1518/001872007X249956

Davidsson, S., & Chen, S. (2014). Method and arrangement for handover warning in a vehicle having autonomous driving capabilities.

de Rosario, H., Louredo, M., Díaz, I., Soler, A., Gil, J. J., Solaz, J. S., & Jornet, J. (2010). Efficacy and feeling of a vibrotactile Frontal Collision Warning implemented in a haptic pedal. *Transportation Research Part F: Traffic Psychology and Behaviour*, *13*(2), 80–91. http://doi.org/10.1016/j.trf.2009.11.003

de Winter, J. C. F. F., Happee, R., Martens, M. H., & Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 196–217. http://doi.org/10.1016/j.trf.2014.06.016

Di Stasi, L. L., Contreras, D., Cañas, J. J., Cándido, a., Maldonado, a., & Catena, a. (2010). The consequences of unexpected emotional sounds on driving behaviour in risky situations. *Safety Science*, *48*(10), 1463–1468. http://doi.org/10.1016/j.ssci.2010.07.006

Dingus, T. a, McGehee, D. V, Manakkal, N., Jahns, S. K., Carney, C., & Hankey, J. M. (1997). Human factors field evaluation of automotive headway maintenance/collision warning devices. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 216–229. http://doi.org/10.1518/001872097778543930

Doshi, A., Cheng, S. Y., & Trivedi, M. M. (2009). A novel active heads-up display for driver assistance. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *39*(1), 85–93. http://doi.org/10.1109/TSMCB.2008.923527

Dredge, S. (2015). Elon Musk : Self-driving cars could lead to ban on human drivers. Retrieved August 7, 2015, from http://www.theguardian.com/technology/2015/mar/18/elon-musk-self-driving-cars-ban-human-drivers

Edworthy, J., Hellier, E., Walters, K., Clift-Mathews, W., & Crowther, M. (2003). Acoustic, semantic and phonetic influences in spoken warning signal words. *Applied Cognitive Psychology*, *17*(8), 915–933. http://doi.org/10.1002/acp.927

Edworthy, J., Hellier, E., Walters, K., Weedon, B., & Adams, A. (2000). The Relationship between Task Performance, Reaction Time, and Perceived Urgency in Nonverbal Auditory Warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 674–677). SAGE. http://doi.org/10.1177/154193120004402247

Edworthy, J., Loxley, S., & Dennis, I. (1991a). Human Factors : The Journal of the Human Factors and Ergonomics Society. http://doi.org/10.1177/001872089103300206

Edworthy, J., Loxley, S., & Dennis, I. (1991b). Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *33*(2), 205–231. http://doi.org/10.1177/001872089103300206

Edworthy, J., Walters, K., Hellier, E., & Weedon, B. (2000). Comparing Speech and Nonspeech Warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 746–749). SAGE. http://doi.org/10.1177/154193120004402265

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. http://doi.org/10.1518/001872095779049543

Enriquez, M., Afonin, O., Yager, B., & Maclean, K. (2001). A pneumatic tactile alerting system for the driving environment. In *Proceedings of the 2001 workshop on Percetive user interfaces - PUI '01* (pp. 1–7). New York, New York, USA: ACM Press. http://doi.org/10.1145/971478.971506

Eriksson, A., Marcos, I. S., Kircher, K., Västfjäll, D., & Stanton, N. A. (2015). The Development of a Method to Assess the Effects of Traffic Situation and Time Pressure on Driver Information Preferences. In *Engineering Psychology and Cognitive Ergonomics* (pp. 3–12). http://doi.org/10.1007/978-3-642-02728-4

Erp, J. B. F. Van, & Veen, H. A. H. C. Van. (2001). Vibro-Tactile Information Presentation in Automobiles. In *Proceedings of Eurohaptics* (pp. 99–104).

Fagerlönn, J., Lindberg, S., & Sirkka, A. (2012). Graded auditory warnings during in-vehicle use: using sound to guide drivers without additional noise. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 85–91). http://doi.org/10.1145/2390256.2390269

Fenton, R. E. (1966). An Improved Man-Machine Interface for the Driver-Vehicle System. In *IEEE Transactions on Human Factors in Electronics* (Vol. HFE-7, pp. 150–157). IEEE. http://doi.org/10.1109/THFE.1966.232662

Fenton, R. E., Cosgriff, R. L., Olson, K. W., & Blackwell, L. M. (1968). One approach to highway automation. *Proceedings of the IEEE*, *56*(4), 556–566. http://doi.org/10.1109/PROC.1968.6344

Fenton, R. E., & Mayhan, R. J. (1991). Automated highway studies at the Ohio State University-an overview. *IEEE Transactions on Vehicular Technology*, *40*(1), 100–113. http://doi.org/10.1109/25.69978

Fenton, R. E., & Montano, W. B. (1968). An Intervehicular Spacing Display for Improved Car-Following Performance. In *IEEE Transactions on Man-Machine Systems* (Vol. 9, pp. 29–35). IEEE. http://doi.org/10.1109/TMMS.1968.300005

Field, A. (2009). *Discovering Statistics Using SPSS. Statistics* (Vol. 58). SAGE Publications.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. *Control*. SAGE Publications. http://doi.org/10.1016/B978-0-12-384727-0.00002-1

Fitch, G. M., Hankey, J. M., Kleiner, B. M., & Dingus, T. A. (2011). Driver comprehension of multiple haptic seat alerts intended for use in an integrated collision avoidance system. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(4), 278–290. http://doi.org/10.1016/j.trf.2011.02.001

Freeman, E., Wilson, G., Vo, D.-B., Politis, I., & Brewster, S. (2017). Handbook of Multimodal-. In S. Oviatt (Ed.), *Handbook of Multimodal-Multisensor Interfaces (In Press)*. Morgan-Claypool.

Gaver, W. W. (1986). Auditory Icons: Using Sound in Computer Interfaces. *Human-Computer Interaction*, *2*(2), 167–177. http://doi.org/10.1207/s15327051hci0202

Geiselman, E. E., Johnson, C. M., & Buck, D. R. (2013). Flight Deck Automation: Invaluable Collaborator or Insidious Enabler? *Ergonomics in Design: The Quarterly of Human Factors Applications*, *21*(3), 22–26. http://doi.org/10.1177/1064804613491268

Gilliland, K., & Schlegel, R. E. (1994). Tactile Stimulation of the Human Head for Information Display. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(4), 700–717. Retrieved from <Go to ISI>://WOS:A1994QG37100009

Gold, C., & Bengler, K. (2014). Taking Over Control from Highly Automated Vehicles. In *Advances in Human Aspects of Transportation: Part II* (pp. 64–69).

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). “Take over!” How long does it take to get the driver back into the loop? In *HFES Annual Meeting* (Vol. 57, pp. 1938–1942). SAGE Publications. http://doi.org/10.1177/1541931213571433

Gonzalez, C., Lewis, B. A., Roberts, D. M., Pratt, S. M., & Baldwin, C. L. (2012). Perceived Urgency and Annoyance of Auditory Alerts in a Driving Context. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1684–1687. http://doi.org/10.1177/1071181312561337

Google. (2015a). Android Auto. Retrieved from https://www.android.com/auto/

Google. (2015b). Google Self-Driving Car project. Retrieved August 7, 2015, from http://www.google.com/selfdrivingcar/

Google. (2015c). Google Self-Driving Car Project - Monthly reports. Retrieved August 7, 2015, from http://www.google.com/selfdrivingcar/reports/

Graham, R. (1999). Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application. *Ergonomics*, *42*(9), 1233–1248. http://doi.org/10.1080/001401399185108

Gray, R. (2011). Looming Auditory Collision Warnings for Driving. *Human Factors*, *53*(1), 63–74. http://doi.org/10.1177/0018720810397833

Gray, R., Ho, C., & Spence, C. (2014). A comparison of different informative vibrotactile forward collision warnings: does the warning need to be linked to the collision event? *PloS One*, *9*(1), e87070. http://doi.org/10.1371/journal.pone.0087070

Green, A. C., Salinger, J. A., Tsimhoni, O., & Raphael, E. L. (2013). Steering wheel device for indicating required supervisory control of a vehicle and method for use. Washington, DC: U.S. Patent and Trademark Office.

Green, D., & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Wiley.

Green, P., Levison, W., Paelke, G., & Serafin, C. (1994). Suggested Human Factors Design Guidelines for Driver Information Systems, *1993*(November 1993).

Haas, E. C., & Casali, J. G. (1995). Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise. *Ergonomics*, *38*(11), 2313–2326. http://doi.org/10.1080/00140139508925270

Haas, E. C., & Edworthy, J. (1996). Designing urgency into auditory warnings using pitch, speed and loudness. *Computing & Control Engineering Journal*, *7*(4), 193–198. http://doi.org/10.1049/cce:19960407

Haas, E. C., & Erp, J. B. F. Van. (2014). Multimodal warnings to enhance risk communication and safety. *Safety Science*, *61*, 29–35. http://doi.org/10.1016/j.ssci.2013.07.011

Hellier, E., Edworthy, J., Weedon, B., Walters, K., & Adams, A. (2002). The Perceived Urgency of Speech Warnings: Semantics versus Acoustics. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 1–17. http://doi.org/10.1518/0018720024494810

Hellier, E. J., Edworthy, J., & Dennis, I. (1993). Improving Auditory Warning Design: Quantifying and Predicting the Effects of Different Warning Parameters on Perceived Urgency. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(4), 693–706. http://doi.org/10.1177/001872089303500408

Hellier, E., Weedon, B., Edworthy, J., & Walters, K. (2000). Using Psychophysics to Design Speech Warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 698–701). SAGE. http://doi.org/10.1177/154193120004402253

Hirst, S. J., & Graham, R. (1997). The Format and Presentation of Collision Warnings. *Ergonomics and Safety of Intelligent Driver Interfaces*, 203–219. http://doi.org/ISBN:0-8058-1956-8

Ho, C., Reed, N., & Spence, C. (2006). Assessing the effectiveness of “intuitive” vibrotactile warning signals in preventing front-to-rear-end collisions in a driving simulator. *Accident Analysis and Prevention*, *38*(5), 988–996. http://doi.org/10.1016/j.aap.2006.04.002

Ho, C., Reed, N., & Spence, C. (2007). Multisensory In-Car Warning Signals for Collision Avoidance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(6), 1107–1114. http://doi.org/10.1518/001872007X249965

Ho, C., & Spence, C. (2005). Assessing the effectiveness of various auditory cues in capturing a driver’s visual attention. *Journal of Experimental Psychology. Applied*, *11*(3), 157–174. http://doi.org/10.1037/1076-898X.11.3.157

Ho, C., & Spence, C. (2008). *Multisensory Driver*. Ashgate.

Ho, C., & Spence, C. (2009). Using Peripersonal Warning Signals to Orient a Driver’s Gaze. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(4), 539–556. http://doi.org/10.1177/0018720809341735

Ho, C., Spence, C., & Gray, R. (2013). Looming Auditory and Vibrotactile Collision Warnings for Safe Driving. In *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.* (pp. 551–557).

Ho, C., Spence, C., & Tan, H. Z. (2005). Warning Signals Go Multisensory. In *Proceedings of HCI International* (pp. 1–10).

Ho, C., Tan, H. Z., & Spence, C. (2005). Using spatial vibrotactile cues to direct visual attention in driving scenes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(6), 397–412. http://doi.org/10.1016/j.trf.2005.05.002

Ho, C., Tan, H. Z., & Spence, C. (2006). The differential effect of vibrotactile and auditory cues on visual spatial attention. *Ergonomics*, *49*(7), 724–738. http://doi.org/10.1080/00140130600589887

Hogema, J. H., De Vries, S. C., Van Erp, J. B. F., & Kiefer, R. J. (2009). A Tactile Seat for Direction Coding in Car Driving: Field Evaluation. *IEEE Transactions on Haptics*, *2*(4), 181–188. http://doi.org/10.1109/TOH.2009.35

Hoggan, E. E., & Brewster, S. A. (2006). Crossmodal icons for information display. In *CHI’06 extended abstracts on Human factors in computing systems* (pp. 857–862). ACM Press. http://doi.org/10.1145/1125451.1125619

Hoggan, E., Raisamo, R., & Brewster, S. S. A. (2009). Mapping information to audio and tactile icons. *ICMI-MLMI ’09*, *0*(1), 327–334. http://doi.org/10.1145/1647314.1647382

Horberry, T., Anderson, J., Regan, M. a, Triggs, T. J., & Brown, J. (2006). Driver distraction: the effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident; Analysis and Prevention*, *38*(1), 185–191. http://doi.org/10.1016/j.aap.2005.09.007

Horrey, W. J., & Wickens, C. D. (2004). Driving and Side Task Performance: The Effects of Display Clutter, Separation, and Modality. *Human Factors*, *46*(4), 611–624. http://doi.org/10.1518/hfes.46.4.611.56805

Hulst, M. Van Der. (1999). Anticipation and the adaptive control of safety margins in driving. *Ergonomics*, *42*(2), 336–345. http://doi.org/10.1080/001401399185694

Inuzuka, Y., Osumi, Y., & Shinkai, H. (1991). Visibility of Head Up Display (HUD) for Automobiles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1574–1578). SAGE Publications.

Kanter, Z. (2015). How Uber's Autonomous Cars Will Destroy 10 Million Jobs and Reshape the Economy by 2025. Retrieved December 8, 2015, from http://zackkanter.com/2015/01/23/how-ubers-autonomous-cars-will-destroy-10-million-jobs-by-2025/

Kaptein, N., Theeuwes, J., & Van Der Horst, R. (1996). Driving Simulator Validity: Some Considerations. *Transportation Research Record: Journal of the Transportation Research Board*, *1550*, 30–36.

Kaufmann, C., Ohg, F., Risser, R., Geven, A., & Sefelin, R. (2008). Effects of simultaneous multi-modal warnings and traffic information on driver behaviour. In *Proceedings of European Conference on Human Centred Design for Intelligent Transport Systems* (pp. 33–42).

Kern, D., Marshall, P., Hornecker, E., & Rogers, Y. (2009). Enhancing Navigation Information with Tactile Output Embedded into the Steering Wheel. In *Pervasive Computing* (pp. 42–58). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-01516-8_5

Kidd, D. G., Nelson, E. T., & Baldwin, C. L. (2010). The Effects of Repeated Exposures to Collision Warnings on Drivers' Willingness to Engage in a Distracting Secondary Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(24), 2086–2090. http://doi.org/10.1177/154193121005402414

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2014). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 1–7. http://doi.org/10.1007/s12008-014-0227-2

Kramer, A. F., Cassavaugh, N., Horrey, W. J., Becic, E., & Mayhugh, J. L. (2007). Influence of Age and Proximity Warning Devices on Collision Avoidance in Simulated Driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 935–949. http://doi.org/10.1518/001872007X230271

Krome, S., Goddard, W., Greuter, S., Walz, S. P., & Gerlicher, A. (2015). A Context-Based Design Process for Future Use Cases of Autonomous Driving: Prototyping AutoGym. In *AutomotiveUI 2015* (pp. 265–272). ACM Press.

Kyriakidis, M., Happee, R., & Winter, J. C. F. De. (2014). *Public opinion on automated driving : Results of an international questionnaire among 5,000 respondents*. *SSRN 2506579*.

Lai, J., Cheng, K., Green, P., & Tsimhoni, O. (2001). On the road and on the Web?: comprehension of synthetic and human speech while driving. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 206–212). http://doi.org/10.1145/365024.365100

Lee, J. D., Bricker, D. L., & Hoffman, J. D. (2008). A Dynamic Programming Algorithm for Scheduling In-Vehicle Messages. *IEEE Transactions on Intelligent Transportation Systems*, *9*(2), 226–234. http://doi.org/10.1109/TITS.2008.922876

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 631–640. http://doi.org/10.1518/001872001775870340

Lee, J. D., Hoffman, J. D., & Hayes, E. (2004). Collision warning design to mitigate driver distraction. *Proceedings of the 2004 Conference on Human Factors in Computing Systems - CHI '04*, *6*(1), 65–72. http://doi.org/10.1145/985692.985701

Lee, J. D., McGehee, D. V., Brown, T. L., & Reyes, M. L. (2002). Collision Warning Timing, Driver Distraction, and Driver Response to Imminent Rear-End Collisions in a High-Fidelity Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(2), 314–334. http://doi.org/10.1518/0018720024497844

Lees, M. N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*(8), 1264–1286. http://doi.org/10.1080/00140130701318749

Lemieux, C., Stinchcombe, A., Gagnon, S., & Bédard, M. (2014). Comparison of simulated driving performance across platforms: from'Low-Cost Desktop'to'Mid-Level'driving simulators. *Advances in Transportation Studies*, *34*.

Lewis, B. A., & Baldwin, C. L. (2012). Equating Perceived Urgency Across Auditory, Visual, and Tactile Signals. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, pp. 1307–1311). SAGE Publications. http://doi.org/10.1177/1071181312561379

Lewis, B. A., Eisert, J. L., & Baldwin, C. L. (2014). Effect of Tactile Location, Pulse Duration, and Interpulse Interval on Perceived Urgency. *Transportation Research Record: Journal of the Transportation Research Board*, (2423), 10–14. http://doi.org/10.3141/2423-02

Lewis, B. A., Penaranda, B. N., Roberts, D. M., & Baldwin, C. L. (2013). Effectiveness of Bimodal Versus Unimodal Alerts for Distracted Drivers. In *Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 376–382). Retrieved from http://drivingassessment.uiowa.edu/sites/default/files/DA2013/Papers/058_Lewis_0.pdf

Li, K. A. (2009). *Enabling eyes-free interaction with tactile messages based on human experience. PhD Dissertation.* UC San Diego.

Li, Y., Burns, C., & Li, Y. (2013). Perceived Urgency of Tactile Warnings. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 530–535. http://doi.org/10.1109/SMC.2013.96

Liebermann, D. G., Ben-David, G., Schweitzer, N., Apter, Y., & Parush, A. (1995). A field study on braking responses during driving. I. Triggering and modulation. *Ergonomics*, *38*(9), 1894–1902. http://doi.org/10.1080/00140139508925237

Lindgren, A., Angelelli, A., Mendoza, P. A., & Chen, F. (2009). Driver behaviour when using an integrated advisory warning display for advanced driver assistance systems. *IET Intelligent Transport Systems*, *3*(4), 390–399. http://doi.org/10.1049/iet-its.2009.0015

Liu, Y. C. (2001). Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems. *Ergonomics*, *44*(4), 425–442. http://doi.org/10.1080/00140130010011369

Liu, Y.-C., & Jhuang, J.-W. (2012). Effects of in-vehicle warning information displays with or without spatial compatibility on driving behaviors and response performance. *Applied Ergonomics*, *43*(4), 679–86. http://doi.org/10.1016/j.apergo.2011.10.005

Liu, Y.-C., & Wen, M.-H. (2004). Comparison of head-up display (HUD) vs. head-down display (HDD): driving performance of commercial vehicle operators in Taiwan. *International Journal of Human-Computer Studies*, *61*(5), 679–697. http://doi.org/10.1016/j.ijhcs.2004.06.002

Llaneras, R. E., Salinger, J., & Green, C. A. (2013). Human factors issues associated with limited ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. In *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design* (pp. 92–98). Retrieved from http://drivingassessment.uiowa.edu/sites/default/files/DA2013/Papers/015_Llaneras_0.pdf

Mahr, A., Cao, Y., Theune, M., Dimitrova-Krause, V., Schwartz, T., & Müller, C. A. (2010). What if it Suddenly Fails? Behavioral Aspects of Advanced Driver Assistant Systems on the Example of Local Danger Alerts. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)* (pp. 1050–1052).

Maltz, M., & Shinar, D. (2004). Imperfect In-Vehicle Collision Avoidance Warning Systems Can Aid Drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 357–366. http://doi.org/10.1518/hfes.46.2.357.37348

Maltz, M., & Shinar, D. (2007). Imperfect in-vehicle collision avoidance warning systems can aid distracted drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(4), 345–357. http://doi.org/10.1016/j.trf.2007.01.002

Marshall, D. C., Lee, J. D., & Austria, P. A. (2007). Alerts for In-Vehicle Information Systems: Annoyance, Urgency, and Appropriateness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 145–157. http://doi.org/10.1518/001872007779598145

Marshall, D., Lee, J. D., & Austria, A. (2001). Annoyance and Urgency of Auditory Alerts for In-Vehicle Information Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1627–1631). SAGE.

Martin-Emerson, R., & Wickens, C. D. (1997). Superimposition, Symbology, Visual Attention, and the Head-Up Display. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(4), 581–601.

Mayhew, D. R., Simpson, H. M., Wood, K. M., Lonero, L., Clinton, K. M., & Johnson, A. G. (2011). On-road and simulated driving: Concurrent and discriminant validation. *Journal of Safety Research*, *42*(4), 267–275. http://doi.org/10.1016/j.jsr.2011.06.004

McKeown, D., & Isherwood, S. (2007). Mapping Candidate Within-Vehicle Auditory Displays to Their Referents. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(3), 417–428. http://doi.org/10.1518/001872007X200067

McKeown, D., Isherwood, S., & Conway, G. (2010). Auditory Displays as Occasion Setters. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(1), 54–62. http://doi.org/10.1177/0018720810366861

Medenica, Z., Kun, A. L., Paek, T., & Palinko, O. (2011). Augmented reality vs. street views: a driving simulator study comparing two emerging navigation aids. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '11)* (pp. 265–274). ACM Press.

http://doi.org/10.1145/2037373.2037414

Meng, F., Gray, R., Ho, C., Ahtamad, M., & Spence, C. (2014). Dynamic Vibrotactile Signals for Forward Collision Avoidance Warning Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. http://doi.org/10.1177/0018720814542651

Meng, F., Ho, C., Gray, R., & Spence, C. (2014). Dynamic vibrotactile warning signals for frontal collision avoidance: towards the torso versus towards the head. *Ergonomics*, *58*(3), 411–425. http://doi.org/10.1080/00140139.2014.976278

Merat, N., Jamson, A. H., Lai, F. F. C., Daly, M., & Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 274–282. http://doi.org/10.1016/j.trf.2014.05.006

Miller, D., Sun, A., Johns, M., Ive, H., Sirkin, D., Aich, S., & Ju, W. (2015). Distraction Becomes Engagement in Automated Driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1676–1680). SAGE.

Mohebbi, R., Gray, R., & Tan, H. Z. (2009). Driver Reaction Time to Tactile and Auditory Rear-End Collision Warnings While Talking on a Cell Phone. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(1), 102–110. http://doi.org/10.1177/0018720809333517

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, automation off: unstructured transition timing for distracted drivers of automated vehicles. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE.

Mok, B. K., Johns, M., Lee, K. J., Ive, P. H., Miller, D., & Ju, W. (2015). Timing of Unstructured Transitions of Control in Automated Driving. In *2015 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1167–1172). IEEE.

Mollenhauer, M. a., Lee, J., Cho, K., Hulse, M. C., & Dingus, T. a. (1994). The Effects of Sensory Modality and Information Priority on In-Vehicle Signing and Information Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *38*(16), 1072–1076. http://doi.org/10.1177/154193129403801617

Mulder, M., Abbink, D. A., & Boer, E. R. (2008). The effect of haptic guidance on curve negotiation behavior of young, experienced drivers. In *IEEE International Conference on Systems, Man and Cybernetics* (pp. 804–809). IEEE. http://doi.org/10.1109/ICSMC.2008.4811377

Mulder, M., Abbink, D. a., Van Paassen, M. M., & Mulder, M. (2011). Design of a haptic gas pedal for active car-following support. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 268–279. http://doi.org/10.1109/TITS.2010.2091407

Mulder, M., Mulder, M., van Paassen, M. M., & Abbink, D. a. (2008). Haptic gas pedal feedback. *Ergonomics*, *51*(11), 1710–1720. http://doi.org/10.1080/00140130802331583

Murata, A., Kanbayashi, M., & Hayami, T. (2012). Effectiveness of automotive warning system presented with multiple sensory modalities. In *SICE Annual Conference* (pp. 920–925). IEEE Comput. Soc. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6318572

National Highway Traffic Safety Administration. (2013). *Preliminary Statement of Policy*

Concerning Automated Vehicles. Washington, DC.

Naujoks, F., Mai, C., & Neukum, A. (2014). The effect of urgency of take-over requests during highly automated driving under distraction conditions. In *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014* (pp. 431–438).

Nees, M. a., & Walker, B. N. (2011). Auditory Displays for In-Vehicle Technologies. *Reviews of Human Factors and Ergonomics*, *7*(1), 58–99. http://doi.org/10.1177/1557234X11410396

Neubauer, C., Matthews, G., & Saxby, D. (2014). Fatigue in the Automated Vehicle: Do Games and Conversation Distract or Energize the Driver? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 2053–2057). SAGE.

Oskarsson, P. A., Eriksson, L., & Carlander, O. (2011). Enhanced Perception and Performance by Multimodal Threat Cueing in Simulated Combat Vehicle. *Human Factors*, *54*(1), 122–137. http://doi.org/10.1177/0018720811424895

Pfromm, M., Khan, R., Oppelt, S., Abendroth, B., & Bruder, R. (2015). Investigation of Take-Over Performance of Driving Tasks by the Driver due to System Failure of Semi-Automated and Assisted Driving. In *Proceedings of the 19th Triennial Congress of the IEA* (pp. 1–8).

Pratt, S. M., Lewis, B. A., Penaranda, B. N., Roberts, D. M., Gonzalez, C., & Baldwin, C. L. (2012). Perceived Urgency Scaling in Tactile Alerts. *HFES Annual Meeting*, *56*(1), 1303–1306. http://doi.org/10.1177/1071181312561378

Prokhorov, D., & Uehara, Y. (2015). Auditory Interface for Automated Driving System. Washington, DC: U.S. Patent and Trademark Office.

Purchase, H. (2012). *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press.

Redelmeier, D. a, & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *The New England Journal of Medicine*, *336*(7), 453–458. http://doi.org/10.1056/NEJM199702133360701

Riener, A. (2012). Subliminal Persuasion and Its Potential for Driver Behavior Adaptation. *IEEE Transactions on Intelligent Transportation Systems*, *13*(1), 71–80. http://doi.org/10.1109/TITS.2011.2178838

Riener, A., Zia, K., Ferscha, A., Ruiz Beltran, C., & Minguez Rubio, J. J. (2012). Traffic flow harmonization in expressway merging. *Personal and Ubiquitous Computing*, *17*(3), 519–532. http://doi.org/10.1007/s00779-012-0505-6

Rusch, M. L., Schall, M. C., Gavin, P., Lee, J. D., Dawson, J. D., Vecera, S., & Rizzo, M. (2013). Directing driver attention with augmented reality cues. *Transportation Research. Part F: Traffic Psychology and Behaviour*, *16*, 127–137. http://doi.org/10.1016/j.trf.2012.08.007

SAE. (2002). Its In-Vehicle Message Priority Standard, *J2395*, 1–7.

SAE J2944: Operational Definitions of Driving Performance Measures and Statistics. (2015).

SAE J3016, & J3016, S. (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*.

Saket, B., Prasojo, C., Huang, Y., & Zhao, S. (2013). Designing an effective vibration-based notification interface for mobile phones. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13*, 1499–1504. http://doi.org/10.1145/2441776.2441946

Salminen, K., Surakka, V., Lylykangas, J., Rantala, J., Ahmaniemi, T., Raisamo, R., … Kildal, J. (2012). Tactile Modulation of Emotional Speech Samples. *Advances in Human-Computer Interaction*, *17*, 1–13. http://doi.org/10.1155/2012/741304

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*(2), 362–380. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16884055

Schoettle, B., & Sivak, M. (2014). *A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia*.

Schweitzer, N., Apter, Y., Ben-David, G., Liebermann, D. G., & Parush, A. (1995). A field study on braking responses during driving. II. Minimum driver braking times. *Ergonomics*, *38*(9), 1903–1910. http://doi.org/10.1080/00140139508925238

Science : Radio Auto. (1925, August 10). *TIME*.

Scott, J. J., & Gray, R. (2008). A Comparison of Tactile, Visual, and Auditory Warnings for Rear-End Collision Prevention in Simulated Driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 264–275. http://doi.org/10.1518/001872008X250674

Serrano, J., Di Stasi, L. L., Megías, A., & Catena, A. (2011). Effect of directional speech warnings on road hazard detection. *Traffic Injury Prevention*, *12*(6), 630–635. http://doi.org/10.1080/15389588.2011.620661

Shim, L., Liu, P., Politis, I., Regener, P., Brewster, S., & Pollick, F. (2015). Evaluating Multimodal Driver Displays of Varying Urgency for Drivers on the Autistic Spectrum. *AutoUI '15*, 133–140. http://doi.org/10.1145/2556288.2556988

Sivak, M., Post, D. V, & Olson, P. L. (1981). Driver Responses to High-Mounted Brake Lights in Actual Traffic. *Human Factors*, *23*(2), 231–235.

Spence, C., & Ho, C. (2008). Multisensory interface design for drivers: past, present and future. *Ergonomics*, *51*(1), 65–70. http://doi.org/10.1080/00140130701802759

Spens, K. E., Huss, C., Dahlgvist, M., & Agelfors, E. (1997). A hand held two-channel vibro-tactile speech communication aid for the deaf : characteristics and results. *Scandinavian Audiology. Supplementum.*, *47*, 7–13.

Stanton, N. (1994). *Human Factors in Alarm Design*. CRC Press. Retrieved from http://books.google.com/books/about/Human_Factors_in_Alarm_Design.html?id=NNF6iVgvAksC&pgis=1

Steele, M., & Gillespie, R. B. (2001). Shared Control between Human and Machine: Using a Haptic Steering Wheel to Aid in Land Vehicle Guidance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, pp. 1671–1675). SAGE. http://doi.org/10.1177/154193120104502323

Sullivan, J. M., & Buonarosa, M. L. (2009). Acquisition, response, and error rates with three suites of collision warning sounds. In *Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 284–290).

Summers, I. R., Whybrow, J. J., Gratton, D. a., Milnes, P., Brown, B. H., & Stevens, J. C.

(2005). Tactile information transfer: A comparison of two stimulation sites. *The Journal of the Acoustical Society of America*, *118*(4), 2527. http://doi.org/10.1121/1.2031979

Suzuki, K., & Jansson, H. (2003). An analysis of driver's steering behaviour during auditory or haptic warnings for the designing of lane departure warning system. *JSAE Review*, *24*(1), 65–70. http://doi.org/10.1016/S0389-4304(02)00247-3

Szybalski, A., Gomez, L. R. P., Nemec, P., Urmson, C. P., & Thrun, S. (2014). User Interface for Displaying Internal State of Autonomous Driving System. Washington, DC: U.S. Patent and Trademark Office.

Telpaz, A., Rhindress, B., Zelman, I., & Tsimhoni, O. (2015). Haptic Seat for Automated Driving : Preparing the Driver to Take Control Effectively. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 23–30). ACM Press.

Terken, Z., Haex, R., Beursgens, L., & Szostak, D. (2013). Unwinding After Work: An In-Car Mood Induction System for Semi-Autonomous Driving. *AutomotiveUI 2013*, 246–249. http://doi.org/10.1145/2516540.2516571

Tuuri, K., Eerola, T., & Pirhonen, A. (2010). Leaping across modalities: Speed regulation messages in audio and tactile domains. In *Haptic and Audio Interaction Design* (pp. 10–19). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-15841-4_2

Underwood, G., Crundall, D., & Chapman, P. (2011). Driving simulator validation with hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(6), 435–446. http://doi.org/10.1016/j.trf.2011.04.008

United Nations. (1959). Agreement Concerning the Adoption of Uniform Technical Prescriptions for Wheeled Vehicles, Equipment and Parts which can be Fitted and/or be Used on Wheeled Vehicles and the Conditions for Reciprocal Recognition of Approvals Granted on the Basis of these . *United Nations Treaty Series, U.N.T.C.*

United Nations. (1968). Convention on Road Traffic. *Treaty Series*. http://doi.org/E/CONF.56/16/Rev.1/Amend.1

van Erp, J. B. F., Toet, A., & Janssen, J. B. (2015). Uni-, bi- and tri-modal warning signals: Effects of temporal parameters and sensory modality on perceived urgency. *Safety Science*, *72*, 1–8. http://doi.org/10.1016/j.ssci.2014.07.022

Van Erp, J. B. F., & Van Veen, H. a. H. C. (2004). Vibrotactile in-vehicle navigation system. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(4–5), 247–256. http://doi.org/10.1016/j.trf.2004.09.003

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). Autonomous Driving: Investigating the Feasibility of Car-Driver Handover Assistanc. In *AutomotiveUI '15* (pp. 11–18). ACM Press. http://doi.org/10.1145/2799250.2799268

Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L. A., & Coughlin, J. F. (2010). The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics*, *53*(3), 404–20. http://doi.org/10.1080/00140130903464358

Warnock, D., McGee-Lennon, M., & Brewster, S. (2011). The Role of Modality in Notification Performance. In *Interact* (pp. 572–588). Springer.

Warshawsky-Livne, L., & Shinar, D. (2002). Effects of uncertainty, transmission type, driver age and gender on brake reaction and movement time. *Journal of Safety Research*, *33*(1), 117–128. http://doi.org/10.1016/S0022-4375(02)00006-3

Weihrauch, M., Meloeny, G. G., & Goesch, T. (1989). *The first head up display introduced by General Motors*. *SAE Technical Paper*. http://doi.org/10.4271/890288

Werneke, J., & Vollrath, M. (2013). How to present collision warnings at intersections?--a comparison of different approaches. *Accident Analysis and Prevention*, *52*, 91–99. http://doi.org/10.1016/j.aap.2012.12.001

White, T. L. (2011). *The Perceived Urgency of Tactile Patterns*.

White, T. L., & Krausman, A. S. (2012). *Perceived Urgency of Tactile Patterns During Dismounted Soldier Movements*.

White, T. L., & Krausman, A. S. (2015). Effects of inter-stimulus interval and intensity on the perceived urgency of tactile patterns. *Applied Ergonomics*, *48*, 121–129. http://doi.org/10.1016/j.apergo.2014.11.010

Wickens, C. D. (1980). *The structure of attentional resources*. *Attention and performance* (Vol. VIII).

Wickens, C. D. (1992). *Engineering psychology and human performance*. HarperCollins Publishers. http://doi.org/10.1002/per.556

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. http://doi.org/10.1080/14639220210123806

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. http://doi.org/10.1518/001872008X288394

Wickens, C. D., Dixon, S. R., & Seppelt, B. (2005). Auditory preemption vs. multiple resouce: who wins in interruption managnement. *Human Factors*, (3), 463–467.

Wiese, E. E., & Lee, J. D. (2004). Auditory alerts for in-vehicle information systems: the effects of temporal conflict and sound parameters on driver attitudes and performance. *Ergonomics*, *47*(9), 965–986. http://doi.org/10.1080/00140130410001686294

Wiese, E., & Lee, J. D. (2001). Effects of Multiple Auditory Alerts for in-Vehicle Information Systems on Driver Attitudes and Performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, pp. 1632–1636). SAGE. http://doi.org/10.1177/154193120104502315

Wilson, F. A., & Stimpson, J. P. (2010). Trends in Fatalities From Distracted Driving in the United States, 1999 to 2008. *American Journal of Public Health*, *100*(11), 2213–2219. http://doi.org/10.2105/AJPH.2009.187179

Wilson, G., Carter, T., Subramanian, S., & Brewster, S. A. (2014). Perception of Ultrasonic Haptic Feedback on the Hand: Localisation and Apparent Motion. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 1133–1142. http://doi.org/10.1145/2556288.2557033

Wilson, G., Halvey, M., Brewster, S. a., & Hughes, S. a. (2011). Some Like it Hot: Thermal Feedback for Mobile Devices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '11*, 2555–2564. http://doi.org/10.1145/1978942.1979316

Wogalter, M. S., Kalsher, M. J., Frederick, L. J., Magurno, A. B., & Brewster, B. M. (1998). Hazard Level Perceptions of Warning Components and Configurations. *International*

Journal of Cognitive Ergonomics.

Young, M. S., & Stanton, N. a. (2007). Back to the future: brake reaction times for manual and automated vehicles. *Ergonomics*, *50*(1), 46–58. http://doi.org/10.1080/00140130600980789

Zhao, S., Brumby, D. P., Chignell, M., Salvucci, D., & Goyal, S. (2013). Shared Input Multimodal Mobile Interfaces: Interaction Modality Effects on Menu Selection in Single-Task and Dual-Task Environments. *Interacting with Computers*, *25*(5), 386–403. http://doi.org/10.1093/iwc/iws021

Zlotnik, M. A. (1988). Applying electro-tactile display technology to fighter aircraft-flying with feeling again. In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, NAECON 1988* (pp. 191–197). IEEE. http://doi.org/10.1109/NAECON.1988.195013