Designing Speech Output for In-car Infotainment Applications Based on a Cognitive Model of Attention Allocation

vorgelegt von
M.Sc.
Julia Niemann
aus Berlin

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -
genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Olaf Hellwich
Gutachter:	Prof. Dr.-Ing. Sebastian Möller
Gutachter:	Prof. Dr. Mark Vollrath

Tag der wissenschaftlichen Aussprache: 4. März 2013

Berlin 2013
D 83

Table of Content

Preface ..... **iv**

Acknowledgements ..... **v**

Abstract ..... **vi**

Zusammenfassung ..... **vii**

Abbreviations ..... **ix**

1. Introduction **1**

1.1 Infotainment Systems: Driver Distraction 2

1.2 Eye Gazes on the GUI Despite the Presence of a VUI: Dominance of Visual Output 3

1.3 Research Question & Thesis Overview 4

2. Theoretical Background and Further Analysis: SEEV Model and Specification of Parameters **6**

2.1 SEEV Model: Attention Allocation in Dynamic Environments 6

2.2 Specification of SEEV Model Parameters 8

2.2.1 Effort: Differentiating $Effort_{between}$ and $Effort_{within}$ 8

2.2.2 Expectancy and Value: Relevant Information Content 11

2.2.3 Summary 14

2.3 EXCURSION: Methodological Approach - Using PDT Display as an Indicator of $P(A_{Speech})$ 14

3. Comparison of Speech and Visual Output on the SEEV Model Parameters **18**

3.1 Disadvantage 1: Trade off Between $Effort_{within}$ and Relevant Information Content 18

3.1.1 Experiment 1: The Influence of $Effort_{within}$ and Relevant Information Content on Attention Allocation 19

3.1.2 Hypotheses 20

3.1.3 Operationalization 21

3.1.4 Method 22

3.1.5 Results 27

3.1.6 Discussion 30

3.2 Disadvantage of Speech 2: Less Controllability 32

3.2.1 Experiment 2: Influence of $Effort_{between}$ on Attention Allocation 37

3.2.2 Hypotheses 38

3.2.3 Operationalization 38

3.2.4 Method 39

3.2.5 Results 44

3.2.6 Discussion 49

3.3 Summary 50

4. Transfer to Applied Context: Design Implications for In-Car Speech Output 53

4.1 Connected Life and Drive (CLD) Prototype 53

4.1.1 Advanced Voice Framework (AVF) 53

4.1.2 General Approach for the CLD VUI 55

4.1.3 Infotainment Applications 57

4.2 Design Recommendations for Infotainment Applications Based on the SEEV Model 59

4.2.1 Increasing Controllability and Relevant Information Content 59

4.2.2 Minimizing Time Effort ($Effort_{within}$) 62

4.2.3 Summary 69

4.2.4 Experiment 3: Evaluation of the Influences of Deducted Design Recommendations on Attention Allocation 70

4.2.5 Hypotheses 72

4.2.6 Method 72

4.2.7 Results 78

4.2.8 Discussion 81

4.3 Influence of Shortening Speech Prompts on Users' Mental Model 83

4.3.1 Experiment 4: Effects on Users' Mental Models 83

4.3.2 Measuring Users' Mental Models 86

4.3.3 Method 88

4.3.4 Results 92

4.3.5 Discussion 96

5. Beyond Pragmatic Qualities: Influence of Hedonic Aspects on Attention Allocation 99

5.1 Increasing the Hedonic Value 100

5.1.1 Experiment 5: The Influence of Hedonic Aspects on Attention Allocation 102

5.1.2 Hypotheses 102

5.1.3 Method 103

5.1.4 Results 105

5.1.5 Discussion 109

6. Consolidation and Future Work 112

6.1 Future Work 118

7. References 121

APPENDIX 135

Preface

This work has been prepared and written between September 2009 and November 2012. Between these dates, I was working as a researcher at the faculty of Electrical Engineering and Computer Science, in the Quality and Usability Lab department at the TU Berlin. The Quality and Usability Lab is part of the Deutsche Telekom Innovation Laboratories (T-Labs). The research for this thesis was integrated in the T-Labs project Connected Life and Drive (CLD), which offered me the chance to conduct experiments with all the facilities, materials and test subjects described in this work.

Earlier versions of Chapters 4.2, 4.3 and 5 have been published previously. This is marked in the respective sections.

Acknowledgements

Firstly I would like to thank Prof. Möller for kindly giving me the opportunity to write my doctoral thesis at his chair. Despite his many obligations and duties, he always found the time to help whenever asked. Further to this, he motivated me to finish this work and not to give up, especially when I began my new job.

I would like to thank Prof. Marc Vollrath for not hesitating in taking on the role as my second supervisor, and also for his extensive engagement in the field of traffic psychology.

A big thank you also to my colleagues at the chair in the T-Labs, who were always forthcoming with a helping hand. In particular, I thank Stefan Schmidt, Klaus Peter-Engelbrecht and Ina Wechsung. I would like to thank Ina Wechsung for not only being a colleague who supported me through my work, but also for becoming such a great friend.

I'd also like to thank Marcus Heitmann for bringing me into the CLD Project and T-Labs in the very first place, and for supporting my research project. I'd like to thank the Multimodal Interaction team for all of the team work during the CLD Project. In particular, Felix Burkhard and Stefan Seide must be highlighted for their contribution to the development and running of my experiments, and for their continued patience and support in facilitating the project, even when my requests and changes were given at short notice.

I must also thank my hardworking students: my Masters student, Jessika Reissland (with whom I had the chance to sharpen my topic through many discussions), Kati Schulz and Sara Bongartz, as well as the student research assistants Maik Mann and Anna Zoe Krahnstöver.

I would also like to take this opportunity to thank Prof. Dietrich Manzey, who wasn't directly involved in this project, however helped and supported me continuously through the course of my studies and for being the person I have to thank for my fascination with the study of Human Factors.

For the proof reading, valuable discussions and help in the final stage of the project, I'd like to thank Grace O'Malley, Jasmin Osinki, Lutz Lorenz, Günter Horna and Rebecca Wiczorek.

Finally, I'd especially like to thank my family. I am very grateful that you always support me in everything I do, unconditionally.

Abstract

Drivers tend to glance at the display of in-car infotainment systems despite the presence of speech output. The SEEV Model by Wickens et al. (2003) defines parameters influencing attention allocation towards events in dynamic environments. Analysing the SEEV Model provides insights on which of the parameters of the SEEV Model speech output has disadvantages compared to visual output. In two driving simulator experiments, it was tested whether increasing or decreasing the deducted parameters of the SEEV Model for speech output by means of improving the speech output in various respects actually decreases attention allocation to the display. It was shown that increasing the relevant information content (corresponding to the parameters expectancy and value) for speech as well as decreasing the time effort (which corresponds to the parameter effort) of speech compared to a baseline condition leads towards lower percentage dwell time to the display. Next, it was shown that a conscious motor action performed to request for speech output (corresponding to the parameter effort) decreases attention allocation towards the display in situations whereby the secondary task gets interrupted by a highly demanding driving task. Based on theses insights, design recommendations for speech output were deducted and implemented in a prototype with several infotainment applications. In another driving simulator experiment, it was observed that the design recommendations actually reduced attention allocation towards the display compared to a common speech output design of in-car infotainment systems. The design recommendations to reduce the time effort of speech output were again evaluated regarding their influence on the development of users' mental models. Finally, it was tested whether increasing the hedonic quality of speech output also leads towards less time glancing at the display.

On the one hand, the conducted experiments showed which parameters of the SEEV Model could be influenced for speech output to decrease attention allocation to the display of a multimodal in-car infotainment system. On the other hand, the results provided insights regarding the applicability of specific SEEV Model parameters to the auditory modality since so far the model had only been evaluated with respect to the visual modality. Finally, it was shown that also hedonic aspects of speech output do indeed influence attention allocation.

Zusammenfassung

Trotz Verfügbarkeit von Sprachausgaben schauen Fahrer vermehrt auf das Display von Infotainmentsystemen im Fahrzeug. Das SEEV-Modell von Wickens et al. (2003) definiert Parameter, welche die Aufmerksamkeitsausrichtung in dynamischen Umgebungen vohersagen sollen. Eine Analyse des SEEV-Modells liefert Hinweise darauf, welche dieser Parameter als Indikatoren für Nachteile einer Sprachausgabe gegenüber einer visuellen Darstellung herangezogen werden können. In zwei Fahrsimulatorstudien wurde getestet, ob durch eine Veränderung der Parameterwerte des SEEV-Modells in Richtung einer Verbesserung der Sprachausgabe, die Aufmerksamkeitsausrichtung zum Display gesenkt werden kann. Es hat sich gezeigt, dass eine Erhöhung des relevanten Informationsgehalts (entspricht den Parametern Expectancy und Value) sowie eine Verringerung des zeitlichen Aufwands (entspricht dem Parameter Effort) bei Sprachausgaben tatsächlich die prozentuale Blickdauer zum Display gegenüber einer Baseline-Bedingung reduzieren können. Darüber hinaus wurde gezeigt, dass eine bewusste motorische Handlung, um die Sprachausgabe anzufordern (entspricht dem Parameter Effort), ebenfalls die Aufmerksamkeitsausrichtung auf das Display minimieren kann. Dies zeigt sich allerdings nur in Situationen in denen die Zweitaufgabe durch eine hoch beanspruchende Fahraufgabe unterbrochen wird. Basierend auf den gewonnenen Erkenntnissen wurden Gestaltungshinweise für Sprachausgaben abgeleitet und in einem Prototyp mit unterschiedlichen Infotainmentapplikationen implementiert. In einer weiteren Fahrsimulatorstudie wurde nachgewiesen, dass die Integration der Gestaltungshinweise gegenüber der herkömmlichen Auslegung von Sprachausgaben bei Infotainmentsystemen im Fahrzeug, die Blickzuwendungen zu einem Display reduzieren können. Die Gestaltungshinweise zur Reduzierung des zeitlichen Aufwands von Sprache wurden darüber hinaus hinsichtlich ihres Einflusses auf die Entwicklung des mentalen Modells von Nutzern evaluiert. Schlussendlich wurde getestet ob die Verbesserung hedonischer Aspekte (entspricht dem Parameter Value) ebenfalls dazu führt, dass prozentual weniger Zeit auf das Display des Prototypen geschaut wird.

Einerseits zeigen die durchgeführten Experimente welche Parameter des SEEV-Modells für Sprachausgaben beeinflusst werden können, um die Aufmerksamkeitsausrichtung auf visuelle Ausgaben von multimodalen Infotainmentsystemen im Fahrzeug zu minimieren. Andererseits liefern die Ergebnisse den Beweis, dass einzelne Parameter des SEEV-Modells auch auf die

akustische Modalität anwendbar sind. Außerdem wurde gezeigt, dass hedonische Aspekte die Aufmerksamkeitsausrichtung beeinflussen können.

Abbreviations

ASR	automatic speech recognition
AVF	Advanced Voice Frame Work
CLD	Connected Life and Drive
LPA	long-term push to activate
GUI	graphical user interface
HCI	human computer interaction
IDC	International Data Corporation
IMI	Intrinsic Motivation Inventory
ISO	International Organisation for Standardization
NASA TLX	NASA Task Load Index
NLU	natural language understanding
PDT	percent dwell time
PND	portable navigation device
QoE	quality of experience
TTS	text-to-speech
SASSI	Subjective Assessment of Speech System Interfaces
SDS	spoken dialogue system
STT	speech-to-text
VUI	voice user interface

1. Introduction

Since its invention at the end of the 19th century, the automobile has been primarily accountable for locomotion of the human race (Dragon, 2007). Similar to this development, mobile telecommunication is credited with having forever changed modern communication, with all of its many extensive services (e.g., phone, text messages, e-mails). Congruent to market researchers of IDC (International Data Corporation), the smartphone market grew by 55 percent in 2011[1] – four times the speed of the overall mobile phone industry – and consequently resulted in a tremendous integration of mobile communication services, in terms of how we both send and receive information. The use of secondary portable devices to retrieve e-mails, digitally read news or even to 'google' restaurant reviews is now common practice. As such, it seems perfectly natural that automobile users could also benefit from the development of quality mobile communication devices tailored to their driving needs. According to a survey conducted by Deutsche Telekom AG (2008), 41.7 percent of people use smartphones while driving (even in cases where legal regulations would actually hinder it). Hence, tailoring telecommunication devices to drivers' unique needs seems like a natural development within the market. The ability to access the Internet while 'on the go' is a desired function by drivers who spend a lot of time in their vehicle, regardless of whether driving privately or professionally. Of critical importance, however, in conceptualizing and designing in-car infotainment applications, is automotive safety. It is well known that driver distraction is "clearly a major highway safety problem" (Wickens & Horrey, 2009, p.54). The NHTSA (1997) have reported one third of all traffic accidents to be caused by poor driver attention. For this reason, it is extremely important to consider limitations of human cognitive resources when developing in-car infotainment applications in order to minimize the likelihood of distracting the driver from what is clearly the most important cognitive task while in the car: driving safely.

[1] IDC Worldwide Quarterly Mobile Phone Tracker, June 9, (2011)

1.1 Infotainment Systems: Driver Distraction

Driver distraction is defined as "attention away from activities critical for safe driving towards a competing activity" (Regan et al., 2009, p.3).[2]

Bubb (2003) subdivided in-car tasks into the following categories:

- Primary task: the actual driving task (longitudinal and lateral guidance)
- Secondary tasks: tasks that support safe driving, such as switching on or off the headlights or using the horn
- Tertiary tasks: tasks not directly related to driving, such as interacting with the climate control, entertainment applications (e-mail, radio) or information applications (navigation systems)

For the purposes of clarity in this thesis, driving-related tasks – both primary and secondary tasks – will be referred to as primary tasks, while non-related driving tasks (tertiary tasks), which can distract the driver from driving safely, will be referred to as secondary tasks. In order to ensure a safe journey, driver distractions due to secondary tasks should be minimized. Information perception during the driving task is primarily operationalized through the visual system, while execution of reactions tends to occur via motor activity. According to Wierwille and Tijerina (1995; as well as Fastenmeier & Gstalter, 1998; Green, 2000; Medenica & Kun, 2007), a huge number of car accidents are caused by visual distractions. The authors suggested the benefits of extending infotainment systems via visual-haptic interfaces with speech input and output to provide sufficient resources for the visual-motor driver task. Thus, studies on distractive potential of in-car infotainment systems could indicate an advantage of acoustic over visual information input (e.g., Vollrath & Totzke, 2000). Using speech as an interaction modality causes 50% fewer driving errors compared to manual interaction (Gärtner et al., 2002, as cited in Bayly et al., 2008). This finding is in line with the multiple resource theory by Wickens (2002). The multiple resource model (see Figure 1) defines four dimensions which all respectively demand different resources: a) information processing

[2] This chapter reuses text fragments from Niemann et al. (2010b).

stage (encoding, central processing vs. response), b) processing code (verbal vs. spatial), c) input modality (visual vs. auditory), and d) response level stage (manual vs. vocal). One of the main assumptions of the model is that two tasks would interfere with each other to a lesser degree if they utilize different mental resources (Wickens, 2002). Thus, it may be fruitful for in-car infotainment systems to use speech as opposed to vision as an in- and output modality.

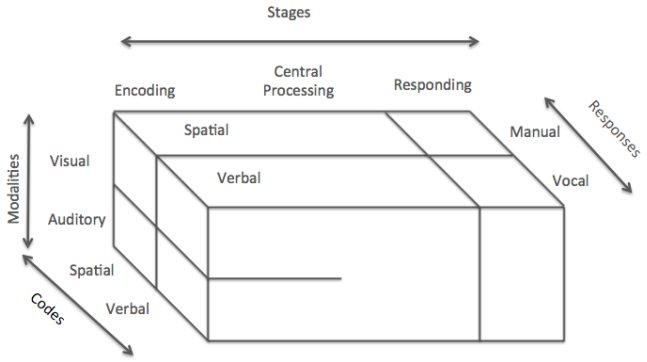


Figure 1. Multiple Resource Model (redrawn from Wickens, 2002).

1.2 Eye Gazes on the GUI Despite the Presence of a VUI: Dominance of Visual Output

To use speech as an in- and output modality, spoken dialogue systems are utilized. "Spoken dialogue systems possess speech recognition, speech understanding, dialogue management and speech generation capabilities, and enable a more-or-less natural spoken interaction with the human." (Möller, 2005, ix). Via the voice user interface (VUI), the human is enabled to interact with devices through a spoken dialogue system. Hierarchically structured menu-based multimodal systems whereby the user can interact either via the graphical user interface (GUI) or via the VUI are currently the most commonly used systems in in-car infotainment systems. Commercial in-car infotainment systems with speech as an in- and output modality are therefore characterized by the user navigating through a menu of options and by the use of commands. The VUI dialogue concept is then commonly determined by the structure of the GUI.

The main objective in using VUIs for in-vehicle systems is to minimize distractions and to avoid eye movements away from the road. It is questionable as to whether VUIs can sufficiently replace graphical user interfaces (GUIs), and the possible difficulties related to the use of speech outputs also remain unclear.

Young, Regan and Hammer (2003) suggested that while visual and psychomotor interactions distract from driving, they are not the only forms of distraction, as auditorily and cognitively demanding secondary tasks may also pose a problem. Hence, the acoustical modality would not represent a panacea (Lee et al., 2001). Spoken dialogue systems should lend themselves to being operated without any visual contact (Bengler et al., 2000), however, users still seem to prefer carrying out tasks visually via the display (Vilimek, 2007). In order to be able to manage the secondary task, test subjects tend to acquire information via the display more often than via acoustical output, which consequently leads to impaired driving performance (e.g., lane keeping; Brumby et al., 2011). During a study using a driving simulator, Kun, Paek, Medenica and Palinko (2009) demonstrated that, while using a Portable Navigation Device (PND) with acoustical support, test subjects averted their gaze from the road more often when a display was available than when it was covered. Consequently, information was gathered visually, even though the secondary task could have been successfully achieved by using speech output and without looking at the display. Furthermore, looking away from the road leads to variations of lane keeping and steering wheel angle. Basic studies have also shown that adults (compared to children) have a greater preference for visual displays (Robinson & Sloutsky, 2004). Reese (1984) demonstrated that during simultaneous presentation of visual and acoustic information, verbal output tends to be ignored.

Thus, it appears to be ‘easier’ or more preferential for the driver to gain information visually than via speech output support, despite the associated potential for neglectful driving. Furthermore, reports indicate that visual feedback represents an important component for interaction with an information system for the driver, despite the option for speech output (Kun et al., 2009). Concealing a display would not provide an adequate solution, and cognitive distraction would remain an issue (Brumby et al., 2011; Salvucci & Beltowska, 2008). As such, it is essential that speech outputs be improved and revised in order to avoid eye movements away from the roads and potential accidents.

1.3 Research Question & Thesis Overview

The aim of this thesis was to examine why, during a driving task, individuals draw more attention to visual output of a secondary task, despite the presence of speech output, and thus avert their gaze from the road. Secondly, I pose the question: can speech outputs be improved based on knowledge on attention allocation without covering the display?

Firstly, parameters related to the prediction of one's attention allocation during a dynamic task were theoretically analysed (Chapter 2), and disadvantages of speech compared to visual output on specific attention allocation parameters were explored (Chapters 3.1 and 3.2). Laboratory experiments were then detailed: Next to the more basic insights regarding the determination of parameters for speech output influencing attention allocation, it was tested whether likelihood of allocating attention to a display can be decreased by increasing the value of the deducted parameters in favour of speech output. Insights obtained from Chapter 3 were then transferred and applied to the in-car infotainment context (Chapter 4). Based on the identified parameters, design recommendations were made for speech output of a multimodal menu-based infotainment system and implemented in a prototype with several infotainment applications. Further evaluations then examined whether glancing at the display of in-car infotainment systems can actually be reduced (Chapter 4.2.4). To ensure that the intended structural measures do not contribute to a deterioration considering the development of users' mental model, an examination by a laboratory experiment was conducted and introduced in the present thesis (Chapter 4.3). Lastly, the hedonic quality of the acoustical output was enhanced, and hedonic quality influences on attention allocation towards the display were thereafter examined (Chapter 5).

2. Theoretical Background and Further Analysis: SEEV Model and Specification of Parameters

A cognitive model of attention allocation for dynamic environments will be introduced and more elaborately analysed to later compare the two modalities (visual output and speech output) regarding specific characteristics influencing attention allocation.

2.1 SEEV Model: Attention Allocation in Dynamic Environments

Wickens and McCarley (2008) define attention as having two characteristics: a filter and a fuel.[3] The filter characteristic refers to the process of attention allocation towards certain information (events or stimuli), also known as selective attention. The fuel, on the other hand, enables one to process the selected information. When processing two stimuli, task difficulty and resources needed determine whether it's possible to perform the processing simultaneously. Figure 2 shows the association between selective attention and the multiple resource model described by Wickens and McCarley (2008).

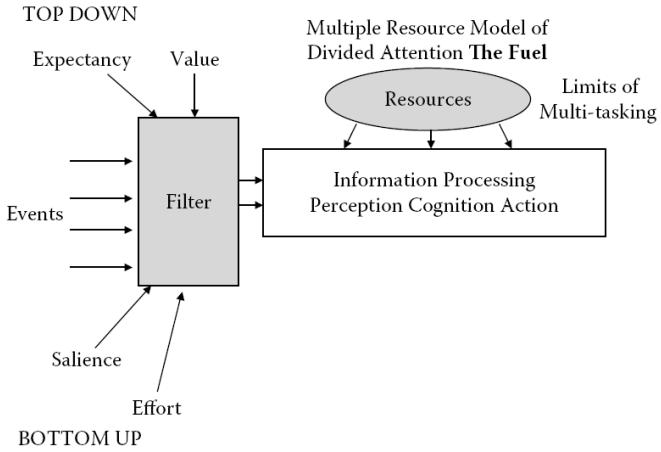


Figure 2. The filter and the fuel: attention selection and diverted attention ("A Simple Model of Attention", Figure cited from Wickens and McCarley, 2008, p.3)

Since visual-manual and auditory-verbal tasks are more easily processed simultaneously in divided attention, the tendency for drivers to select the display (visual output) to encode the information remains unexplained. It is possible that it is driven by the selective attention

[3] This chapter is a revised and extended version of a chapter in Niemann et al. (2010a), Niemann et al. (2010b) & Niemann et al. (2010c).

allocation. Selective attention (or attention allocation) is influenced by four factors, as defined in the so-called SEEV Model.

The SEEV Model (Wickens et al., 2003) allows predictions of operators' attention allocation in dynamic environments and will therefore be used to explain the effect of information acquisition via the display despite the presence of speech output. This model also provides a theoretical basis for the deducted design recommendations for in-car infotainment systems developed in this thesis, which aim to improve the speech output. Four factors influencing the probability of attention allocation to an area[4] (P(A)) are defined in the model: salience (S), effort (EF), expectancy (EX) and value (V). Salience and effort represent what are referred to as 'bottom up' factors, which influence attention allocation through environmental stimuli. Salience describes the strength of a stimulus; a strongly salient stimulus (e.g., a brightly flashing light) is more likely to attract attention. Probability decreases however if the effort for information access grows (e.g., when larger distances need to be covered by long scanning paths and head movement). Expectancy and value, in comparison, are regarded as 'top down' factors. Based on the knowledge of the operator regarding the bandwidth of information presented, as well as the value of this information, the probability of attention allocation will be influenced. These are referred to as knowledge driven factors. As such, the subjective experiences related to operators' expectation in detecting an event in an area (or information source) is highly influenced by the frequency of events occurring in a specific period of time and the information content in the language of information theory (bits per event; Senders, 1964, 1980) – or so-called expectancy. If the events are highly relevant for task accomplishment (value), the probability of glances would increase accordingly. This is further represented in the following equation (Wickens & McCarley, 2008):

$$P(A_{IS}) = sS - efEF + exEX + vV \qquad (1)$$

$P(A_{IS})$: probability of attention allocation to an information source
S: level of salience in a particular task and information source combination
s: general strength of factor salience on attention allocation
EF: level of effort in a particular task information source combination
ef: general strength of factor effort on attention allocation
EX: level of expectancy in a particular task information source combination

[4] As the term area is not applicable for a systems with acoustic output; from now on, the focus will be on probability of attention allocation to information sources ($P(A_{IS})$: probability of attention allocation towards an information source).

ex: general strength of factor expectancy on attention allocation
V: level of value in a particular task information source combination
v: general strength of factor value on attention allocation

Note that Wickens et al. (2003) do not make a definite statement as to whether multiplication or addition is the right mathematic operation to combine expectancy and value (see Chapter 2.2.2 for further discussion).

As well as in the context of aviation, the evaluation of the SEEV Model and the prediction of percent dwell time (PDT) on defined areas of interests (AOIs) when using the SEEV Model have also occurred in the automobile context. The percent dwell time on display (PDT Display) refers to the percentage of time that the participant spent looking at the screen during the infotainment task and represented the operationalization of attention allocation to the display. Horrey, Wickens and Consalus (2006) varied the parameter expectancy in a driving simulator study for a driving task as well as for a secondary task. Moreover, prioritization of tasks occurred in the experiment (variation of the factor value). A significant influence of the factor value as well as of expectancy variation on attention allocation to the secondary task was found. Across the various studies (Wickens et al., 2003; Wickens et al., 2008; Horrey et al., 2006), the predicted dwell times and the actual scanning (PDTs) were correlated between r= 0.65 to 0.95 (Wickens & McCarley, 2008). The predicted dwell times were based on the SEEV Model. No metric values were inserted into the formula for predicting dwell times. Rather, the relation to the other AOIs were defined by simply "rank ordering the terms on the lowest integer values possible (e.g. 1, 2, 3)" (Wickens & McCarley, 2008, p. 66). According to Wickens and McCarley (2008), the SEEV Model is also relevant in terms of display design; for example, the theory proposes that highly relevant displays should be made more salient, and displays with a high bandwidth of information should be allocated close to each other.

2.2 Specification of SEEV Model Parameters

In the following, the SEEV Model parameters effort, expectancy and value will be more closely analysed.

2.2.1 Effort: Differentiating $\text{Effort}_{\text{between}}$ and $\text{Effort}_{\text{within}}$

The effort entailed in the SEEV Model is defined as 'information access effort'. This includes the effort of allocating attention from one stimulus to another stimulus. This factor will be

referred to as $effort_{between}$ in the following. The factor $effort_{between}$ was operationalized as spatial distance, and therewith associated variation of scanning paths in the studies of model evaluation (Wickens & McCarley, 2008). “People will tend to avoid longer scans or other information access travels when shorter ones can be made” (Wickens & McCarley, 2008, p. 54). If more complex motor head movements are necessary instead of eye movements to allocate attention on a display, it will result in significantly less eye movements to the display (Ballard, Hayhoe & Pelz, 1995). It is possible to switch attention without eye movements and without interruption within the visual field (Eimer, 1999), however, if motor reaction is necessary, a phase with no information acquisition occurs (Latour, 1962). It is possible to perform two tasks in parallel if the $effort_{between}$ is minimal. $Effort_{between}$ can also be referred to as ‘task switching costs’. Jersild (1927) observed ‘alternation costs’ when test subjects performed different tasks after another compared to performing the same tasks in a row. According to Spector and Biedermann (1976), the reaction time increases when we switch from one mental process to another.

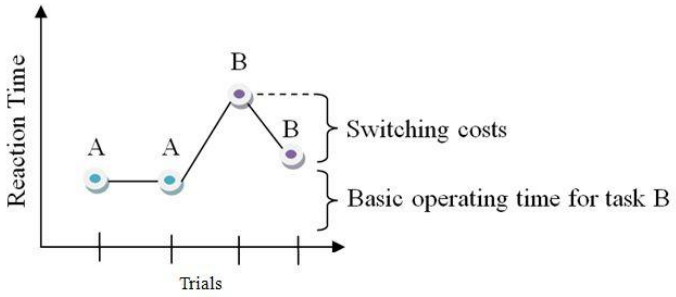


Figure 3. Increase in reaction time by switching between mental processes (Spector & Biedermann, 1976)

Figure 3 shows that the time needed to perform task type B after previously performing task type A is higher than it would be for performing task type B after task type B. Juola and Botella (2004) differentiate between task switch and location switch. For visual output in case of eye movements among the task switch, a location switch needs to be performed. For perceiving and processing auditory information, usually no location switch is necessary. The effort of switching from one object to another for the auditory output is therefore not as high as for the visual output (if the same task is performed with information of visual and acoustical output). Compared to resolving long scanning paths by allocating attention from one visual stimulus to another, auditory attention allocation is not necessarily linked to a motor reaction (e.g., turning one’s head, moving to the stimulus). Therefore the acoustic output of a secondary task during a driving task has even higher benefits as it means that

allocating attention from the street (primary task) towards the secondary task is not aligned with high location switch costs on the factor $effort_{between}$.

Besides the influence of $effort_{between}$, the effort to capture and perceive the relevant information has a decisive influence on the prediction of attention allocation. The time effort to perceive task-relevant information when the task switch is already performed will be called $effort_{within}$ in the following. $Effort_{within}$ represents the general requirement in terms of duration or cognitive load, which constitutes the operator's reception of information to cope with the task. While $effort_{between}$ describes the switching costs, $effort_{within}$ can be viewed as part of the operating time for task (see Figure 4).

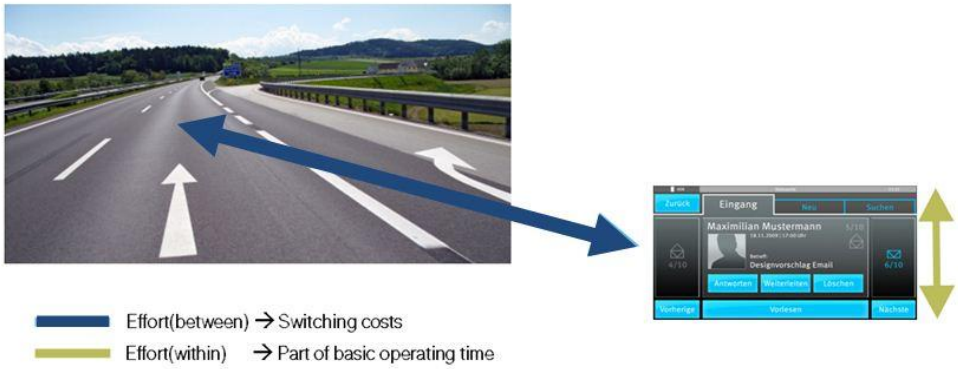


Figure 4. Differentiation between $effort_{(within)}$ and $effort_{(between)}$.

Freed (2000) also differentiates between 'duration' (complies with $effort_{within}$) and 'switching cost/interruption' ($effort_{between}$) amongst influences on task switching, next to importance and urgency. A display with many information units requires more time exposure than a display with few information units. According to Gray, Fu and Schoelles (2006), distribution of cognitive, perceptive and motor resources is based on cost optimization measured in terms of time; "time is a resource that is to be preserved" (p.416). Minimal time changes during processing of tasks always have an effect on the operator's behaviour (Gray & Fu, 2004). Furthermore, Wickens and McCarly (2008) mention an inhibition influence of the 'clutter' (accumulation of irrelevant information units) on attention allocation, which involve a more time-consuming visual search. This factor becomes particularly relevant for events or areas of interest with complex and rich informational content. Displays or speech outputs of infotainment systems involve an accumulation of items and information within an event, which requires a visual search (display) or filtering of relevant information (speech output).

To the best of my knowledge, no study has evaluated this factor of the SEEV Model so far (either for acoustical and visual information).

For the SEEV Model, a more detailed differentiation between $effort_{within}$ and $effort_{between}$ is recommended. An advantage for acoustical output of secondary tasks performed during a visual primary task compared to visual output can be seen for the switch of attention allocation ($effort_{between}$), while $effort_{within}$ is a decisive factor for huge benefits of visual output. This will be discussed in more detail in Chapter 3.

2.2.2 Expectancy and Value: Relevant Information Content

In terms of content, pictures tend to be more informative than written text (Bauer, 1995). However, to understand fully what this means, it raises the question: what exactly does 'more informative' mean?

Wickens and McCarley (2008) dedicated information content to the factor expectancy: The expectancy of an AOI increases with a higher bandwidth as well as contextual cueing. Wickens and McCarley (2008) suggest that the message's information content in the language of information theory (Shannon & Weaver, 1949) as well as the number of events occurring in a specific time period has to be considered to define the bandwidth of an information source. The message's information content (entropy) is the minimum number of binary bits per symbol required to encode a message. "[…] - Entropy increases with randomness" (Havaldar & Medioni, 2009, p.152). Next to the information content in the language of information theory, the more often events tend to occur/change in a given window of time ('event rate'), the higher the probability of attention allocation. "When information from a given channel is refreshed more often (i.e., a higher bandwidth), uncertainty regarding the current status will increase more rapidly. Observers will therefore need to sample this channel more often, in order to attain the information they expect at that particular location" (Horrey et al., 2006, p.8). In a driving simulator experiment the bandwidth for the AOI 'driving environment' was manipulated by the frequency of wind turbulences which increases the uncertainty of lane position: "In the low wind condition, wind occurred every 5 to 7 seconds (approximately 0.16 Hz). In the high wind condition, the interval between wind bursts was reduced to 2 to 4 seconds (approximately 0.33 Hz)" (Horrey et al., 2005, p.26). For GUIs or VUIs a 'refresh' or 'change' does not take place with every symbol/item presented. Refresh for the information sources GUIs and VUIs usually appear with different views of a display or

speech prompts containing several symbols/items. In the case of infotainment tasks, these changes/refreshes are on the one hand often initiated by a user's input but can also be influenced by system-initiated changes, e.g., new incoming email notifications.

The definition of information content in terms of information theory completely excludes the consideration of semantic and pragmatic factors. In the subjective sense, no consideration is given to the subjective importance (or meaning) of a message for the receiver. For example, the help button is of critical importance in a situation where the user is unable to proceed any further within the interaction, while it is unimportant in a situation where the user knows exactly what to do. Thus, information content in terms of information theory does not as such take the specific characteristics of perceived relevance into account. This issue of relevance is however resolved by the factor value in the SEEV Model. According to Wickens and McCarley (2008) the value for task with a visual output is set as followed: a) "define tasks served by an AOI and their relative value or importance within the overall multitask context" and b) "establish the relative relevance of each AOI to each task" (p.57). Thus, for calculating the probability of attention allocation towards a specific AOI, Horrey et al. (2006) suggest differentiating between relevance of the AOI for solving a task and overall value of the task.

Relevance of information for task completion is especially important for comparing visual and acoustical output of the same infotainment task since the overall value of the infotainment task within the multitask context of driving is the same for the two modalities; i.e. the email task itself compared to other tasks has the same overall importance for the VUI as for the GUI. What matters in this context is the average task-relevant information content of the two information sources (GUI and VUI). The time to perceive the task-relevant information ($effort_{within}$) of a speech prompt for example usually increases with every new item occurring (will be discussed again later in Chapter 3.1). A content-heavy display involves a high cost factor (high number of symbols and thus high temporal effort to perceive the information), however at the same time, if the content is highly relevant it could be argued to be of a high value and thus increases probability of attention allocation. With relatively irrelevant items and the same temporal effort, in comparison, the probability of attention allocation decreases. Thus the time effort needs to be set in relation to the information content. However, not only in terms of information theory; the task-relevancy of each symbol presented by an event (new view of a display or a speech prompt) needs to be multiplied by the information content in the language of information theory and added up over all symbols presented by the view or speech prompt. This will be referred to as relevant information content in the following.

According to Wickens and McCarley (2008) the value is multiplied by, or added up to, expectancy (bandwidth). As mentioned before, Wickens and Horrey (2009) do not define which of the mathematical operations should be accomplished. Although the overall relevance for solving a task is multiplied with the bandwidth for solving a task in Horrey et al. (2005) - they do express preference for summing as opposed to multiplying, in order to "signal the independency of the two factors" (p.62). In this thesis, I will assume that they should be multiplied. Wickens and Horrey (2009) argue that an empty autobahn should be observed more often although there are no hazards to be detected. No hazards would result in a very low expectancy rate, which in the case of multiplying, results in a low predicted percent dwell time for the AOI empty autobahn – which contradicts the assumption that an empty autobahn should be observed often. I argue that the assumption of high scanning frequency of an empty autobahn is justified by the fact that an empty autobahn is actually a frequently changing environment, and due to driver knowledge, that hazards could be detected (as Wickens & McCarley, 2008, states that these are knowledge driven parameters). Those two factors represent the parameter expectancy – which will therefore not be of low value. Since the value of expectancy is not low, a higher scanning of an empty autobahn is predicted which again would rather meet the assumption of high percent dwell time for the AOI empty autobahn. Nonetheless, the value should be set high to meet the tremendous consequences of not detecting a possible hazard in order to avoid an overall low probability of attention allocation. Further, in decision theories it is common that the factors expectancy and value are multiplied (Tolman, 1959). Thus in the present work, it is suggested that the value (here relevance for solving a task) is linked multiplicatively, however not with the overall bandwidth for solving a task: Compared to Wickens & McCarley (2008), the relevancy of each symbol for solving the task will be multiplied to each symbol's information content in terms of information theory. In addition, to calculate the bandwidth the number of refreshes needs to be considered. To sum up, subjective relevant bandwidth increases with:

- the information content as defined in information theory (Shannon & Weaver, 1949)
- relevance of each symbol for solving the task
- increase of refreshes/changes within a specific time period

2.2.3 Summary

The following specifications of the SEEV Model for describing probability of attention allocation to complex information sources like a GUI or VUI (to enable comparison of the two different modality outputs) is suggested:

- a differentiation between $effort_{(within)}$ and $effort_{(between)}$.
- $effort_{within}$ has to be considered in relation to the relevant information content (corresponds to the factor expectancy and value) to balance the cost benefit ratio

The relevant information bandwidth (see previous chapter for description) consists of the informational content of each presented symbols, linked with the specific relevance of each symbol for solving the task and the number of changes/refreshes (new screens or new speech prompt occurring). Since it is most often the case that changes/refreshes for the GUI and the VUI of the same infotainment task with the same dialogue concept of a multimodal menu based system are initiated simultaneously, in the following, preference will be given to conducting the relevant information content without taking the frequency of refreshes into account, in order to compare the two modalities, visual and speech output.

2.3 EXCURSION: Methodological Approach - Using PDT Display as an Indicator of $P(A_{Speech})$

One of the goals of this thesis is to examine characteristics of speech output, compared to visual output, based on the SEEV Model parameters, in order to explore why drivers tend to gain information via the display instead of via speech output. Design recommendations for improving speech output in in-car infotainment systems are hoped to be developed, and thereby decrease likelihood of glancing towards the display. Subsequently, $P(A_{Speech})$ (the probability of attention allocation towards the speech output) while driving should be increased by positively influencing the SEEV Model parameters characteristics of speech output. To evaluate this assumption, I needed to first measure $P(A_{Speech})$.

Although the SEEV Model is applicable to other modalities, only the predictability of visual attention allocation has been evaluated so far. This is due to the difficulties regarding the measurement of acoustical attention allocation: “How to measure such selection within other perceptual modalities (e.g., hearing) remains a challenge” (Wickens & McCarley, 2008, p.61).

An acoustical attention allocation switch is not linked to a motor reaction and according to this, acoustical attention allocation can only be detected via performance measurements. Performance measurement is not as highly frequent as eye gazes measurement. By measuring eye gazes, a sampling rate of about 500hz is provided. This sampling rate cannot be provided by performance measurement. Additionally, it is difficult to find a task whereby only attention allocation determines the performance rate and whereby no other higher cognitive processes are activated.

To overcome this problem, measuring PDT Display (percent dwell time on display) was used as a reversed measurement for $P(A_{Speech})$ in this thesis. This will be explained in the following:

Rötting (2001) defines PDT as the relative amount of eye gazes on certain areas of interest over a specific time period. Important objects receive more eye gazes than other objects. Figure 5 shows the relations between the probabilities of paying attention towards the display ($P(A_{Display})$), the street ($P(A_{Street})$) and speech ($P(A_{Speech})$ within the paradigm of a multimodal secondary task (e.g., infotainment task with visual and speech output performed while driving). $P(A)_{Street}$ in the presented figure represents all driving relevant information sources which the driver needs to allocate attention to to ensure safe driving. If $P(A_{Speech})$ increases, the probability of attention allocation to the display $P(A_{Display})$ will decrease. Thus, the probability of paying attention to the display (measured by PDT Display) if speech outputs are available decreases if:

a) properties of speech outputs encourage attention allocation towards it

b) properties of visual output cause only low potential attention allocation

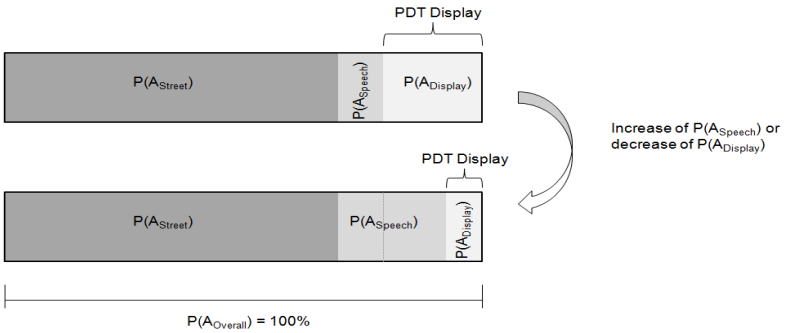


Figure 5. Relation of $P(A_{Display})$, $P(A_{Street})$ and $P(A_{Speech})$ in the secondary task paradigm driving and performing a secondary task. It is assumed that PDT Display can be reduced by increasing $P(A_{Speech})$ or decreasing $P(A_{Display})$. By keeping the other entire parameters stable, with the same secondary task, PDT Display can be used as an indicator for $P(A_{Speech})$

By using this secondary task paradigm (interacting with a non-related driving task with a visual and speech output while performing a driving task) and by keeping the display information of the secondary task as well as the driving task the same, the percent dwell time display (PDT Display) can be used as a method of data collection for $P(A_{Speech})$. Measuring PDT Display is therefore an indirect and reverse measurement of $P(A_{Speech})$.

At this point it should be mentioned that it is not hereby proposed that one can draw completely certain conclusions regarding acoustical attention allocation by examining PDT Display in conditions of speech output. It is possible that a person may look at the display while their attention is actually allocated to the speech output. Scharf (1998) distinguishes between looking and seeing, as do Wickens and Horrey (2009), who propose the phenomenon of inattentional blindness, i.e., looking at but not processing the information. While 'looking' is an action indicative of orientation to stimulus, 'seeing' refers to the actual perception and processing of a stimulus. In the case of looking at the display but listening to the speech prompt, the actual percentage of attention allocation towards the speech output by measuring PDT Display would be underestimated. Detection of eye movements is not entirely informative if the presented event in this AOI has actually been perceived and processed. Another restriction should be mentioned: in studies evaluating the model, no absolute measurable values for the parameters can be determined; only the relative value (i.e., in comparison to the other AOIs or information sources such as VUIs) of particular parameters can be defined. As mentioned before, Wickens and McCarley (2008) also set the values of the SEEV Model parameters for predicting the dwell time based in relation to other AOI or tasks and do not define absolute values for the particular parameters, e.g., for determining the value

of the parameter expectancy they suggest rank ordering the AOIs from "most rapidly changing to the slowest" (p. 57).

Despite these limitations, the influence of the SEEV Model parameters for speech output (and therefore for the acoustical modality) on attention allocation were tested in the present thesis by varying the parameters for speech output. It was examined if the predicted lowering or increase of attention allocation for $P(A_{Speech})$ would lead to actual lowering or increase (in a reverse manner) of PDT Display. In other words, based on considered variation of the particular SEEV Model parameter characteristics for speech output, the predicted probability of attention allocation to display ($P(A_{Display})$) was tested in light of whether it will lead to higher (or lower) attention allocation. These hypotheses were experimentally tested via a driving simulation study: driving on a highway while performing a secondary task with a visual and speech output. The SEEV Model parameters for the driving task and for the visual output of a secondary task were kept stable while the SEEV Model parameters for the speech output were actually varied. The dependent variable and reverse measurement for attention allocation on speech ($P(A_{Speech})$) was PDT Display. According to Horrey et al. (2006), for information acquisition via the display foveal vision is required to perform an in-car infotainment task while the driving task can also be performed by peripheral vision. "The model does a much better job of predicting scans to in-vehicle displays than to the roadway in front in the driving application" (Wickens & McCarley, 2008, p.58). Thus it was chosen to use PDT Display and not PDT Street as an indicator for $P(A_{Speech})$. PDT Display was collected in all of the experiments in this thesis.

3. Comparison of Speech and Visual Output on the SEEV Model Parameters

The parameters of the SEEV Model specified in the previous Chapter ($effort_{between}$, $effort_{within}$ and relevant information content) serve as useful tools in analysing the advantages and disadvantages of speech compared to visual output. Potential disadvantages of speech include high time effort to perceive relevant information, as well as a lack of controllability. Both aspects will be addressed in the following subsection.

3.1 Disadvantage 1: Trade off Between $Effort_{within}$ and Relevant Information Content

A difference between visual output and speech output is $effort_{within,}$ which mainly refers to the time required for perceiving the relevant information (see Chapter 2.2.1); a significant benefit of displays being their ability to detect information more rapidly than voice outputs.[5] Because of the sequential nature of speech outputs, serial listening from the beginning to the end is required. Therefore, relevant information cannot be 'picked out' as easily as with visual output. For visual search phenomena, an example is the pop-out effect, whereby the target differs from the clutter (irrelevant information) by a unique visual characteristic. In this sense, the target (relevant information) can be detected very quickly (Treisman, 1985). Not every item on the screen has to be scanned and more clutter (irrelevant information) will not result in longer duration to detect the target stimulus. The following conclusion can be drawn: While for speech output the time effort definitely increases with every new symbol or proposition, this is not mandatory for visual information. Furthermore, listening to a text tends to be more time consuming than reading a text (Kozma, 1991).

In practice, as with every new information item, the $effort_{(within)}$ increases for speech, speech-output of in-car infotainment systems is attempted to be kept short. For this reason Jeschke (2008) recommends speech output not to be any longer than necessary. She further suggests simple formulations, short sentences and clear sentence constructions to be advantageous. In a study concerning distraction effect of long speech outputs compared to keyword based speech outputs, Villimek (2007) found that the presentation of extra, irrelevant information that could essentially be omitted lead to performance deficits on the primary task. Thus, aesthetic

[5] This chapter reuses text fragments from Niemann et al. (2010b) & Niemann et al. (2010c).

language could be excluded if this were to be considered an unnecessary temporal extension of speech output. To reduce the $effort_{within}$, developers of in-car speech dialogue systems commonly use the strategy 'speak-what-you-see' (Vilimek, 2007; Weinschenk & Barker, 2000): Interaction options are only presented on the GUI. Interaction options for VUIs are speech commands with which users initiate the next interaction step. To clarify again, commands that can be used (options) are presented (outsourced) on the display instead of by speech output. According to Lee, Caven and Brown (2001), remembering the interaction option in command-based speech systems negatively influences driving performance. Since speech commands are of high relevance (especially to help novices know what to say), information items with significant task relevance are omitted for the speech prompt.

The relevant information content also increases with each new symbol, depending on the subjective relevance of each symbol and the information content of each symbol. Inference occurs either way, either the number of information items is reduced for speech or the time effort is increased. This is a trade off problem between decreasing the $effort_{within}$ and providing the necessary amount of relevant information. It is especially pertinent if with every new information item added, the time effort increases. This effect is particularly pronounced for acoustic output and is less of a concern with visual output.

To sum up, there are considerable drawbacks to speech output in terms of either:

a) less relevant information content (to keep the time effort low), or

b) increased time effort (aligned to provide same relevant information content).

3.1.1 Experiment 1: The Influence of $Effort_{within}$ and Relevant Information Content on Attention Allocation

The aim of the first study is to evaluate and assess the parameters' influence for $effort_{within}$ and relevant information content on attention allocation. Does decreasing the $effort_{within}$ for speech but not display result in higher attention allocation towards speech? Does increasing relevant information content for speech compared to display result in increased attention allocation towards speech?

So far, the construct of effort from the SEEV Model has been operationalized as the variation of spatial distance between two areas of interest (Steltzer & Wickens, 2006). This leads to the

parameter $effort_{between}$. The influence of temporal effort on relevant information extraction and subsequently on attention allocation has not been the object of investigation, however Wickens and McCarley (2008) stress the inhibiting control of "effort, particularly time, required to search through information to locate a desired target" (p.62). In the following experiment, it was investigated whether $effort_{within}$ truly has an influence on attention allocation. Furthermore it was examined if increasing the relevant information content in terms of presenting more relevant semantic information for completing a task results in increased attention allocation to that given source of information. In addition to yielding insights into which parameters can be influenced to reduce attention allocation towards the display of infotainment systems, it should be investigated as to whether the specific SEEV Model parameters are applicable to the auditory modality, as well as whether time effort characteristics to perceive relevant information ($effort_{within}$) indeed influence attention allocation.

A driving simulator study with test subjects was conducted. While driving (primary task), a secondary task with visual (GUI) and acoustic (VUI) system output was performed. The parameters $effort_{within}$ and relevant information content were tailored for speech output and display in order to increase $P(A_{Speech})$. The two parameters of the information source display to decrease $P(A_{Display})$ and of the information source speech output to increase $P(A_{Speech})$ were thus both influenced (see Chapter 2.3). Based on the variation of the parameters, the predicted attention allocation to the VUI was investigated by examining the actual percent dwell time on the display (PDT Display).

3.1.2 Hypotheses

H1: Varying the $effort_{within}$ by changing the time to perceive information will lead to different percentage dwell times on display.

H1.1: Lowering the $effort_{within}$ for speech output by lowering the amount of irrelevant speech output information content and increasing the amount of irrelevant information content on the display will lead to a significantly shorter percentage dwell time for the display than the baseline condition.

H1.2: Lowering the $effort_{within}$ for speech output by using different voices for relevant and irrelevant information content (faster discrimination is provided) will lead to a shorter percentage dwell time for the display than the baseline condition.

H2: Increasing the relevant information content for speech output by increasing the amount of relevant information in relation to the visual output will lead to a significantly lower percent dwell time on display than the baseline condition.

3.1.3 Operationalization

In the first driving simulator study, the influence of the parameter relevant information content and $effort_{within}$ on the attention allocation for the acoustic modality was examined. The operationalization of varying these parameters will be described in the following.

The secondary task represented a free recall task. Words were presented graphically, as well as verbally coded in the form of a list (see Figure 6). Some of these words were animals and some were not (non-animal words). After the presentation of animals and non-animal words, subjects were asked to recall the presented animals verbally. Each reproduced animal was rewarded with money.

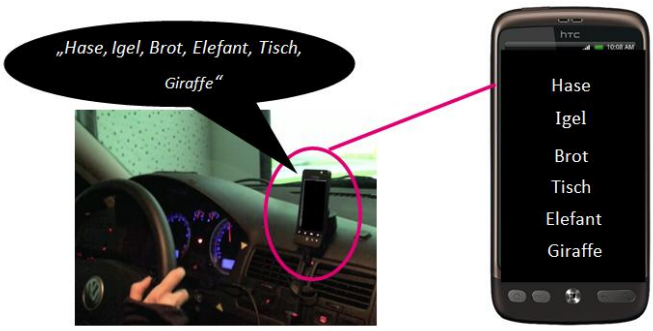


Figure 6. Visual and speech output during the secondary task.

Concerning speech outputs, variation of the factor $effort_{within}$ was achieved through an increase in the amount of non-animal words presented compared to in the baseline condition. Non-animal words represented irrelevant information (value almost zero): reproduction of these words was not part of the task, nor was their reproduction rewarded. Thus, only the factor $effort_{within}$ increased considerably as due to the low value of task-relevance for the added non-animal words, the factor relevant information content remained almost constant as in the baseline condition. Only time effort to perceive the relevant information was increased.

Additionally, the factor $effort_{within}$ was adjusted in a new way which was assumed to result in more speedy communication of relevant information: Animals were presented in a different voice (male voice) from non-animal words (female voice). Therefore, the decision was

facilitated by whether the uttered word was an animal, while this was not the case in the baseline condition. Consequently, acquisition of relevant information was encouraged. Thus, relevant information content remained constant, while the factor $effort_{within}$ was modified. The properties of the visual output were not changed whatsoever.

Finally, compared to the visual output, relevant information content was increased for the speech output. This was achieved by increasing the amount of relevant information during speech output while the relevant information content for visual output was kept the same as at baseline. For example, one animal may have been presented visually, while three were presented acoustically. In this case, reproducing every animal – because attention was allocated to speech output – would result in higher monetary profit. However, the total amount of words presented remained constant in order not to increase the factor $effort_{within}$.

3.1.4 Method

The experiment was carried out in a driving simulator of the Centre for Human Machine Systems at the TU Berlin.

Experimental Design

A two-factor repeated measures experimental design was used for the present study. One independent variable was the degree of $effort_{within}$ in relation to the display compared to a baseline condition. This factor varied in two conditions. The extent of relevant information content in relation to the display compared to baseline represented the second independent variable (see table 1).

Table 1. Experimental design of the first experiment

IV1: level $effort_{within}$			**IV2: level relevant information content**	
Baseline	low1	low2	baseline	high1

In the baseline condition for both independent variables, the same amount of information as well as the same level of relevant information content was presented visually and acoustically. Low1 describes a decrease in $effort_{within}$ for acoustical information in relation to visual information, as compared to the baseline, by increasing the clutter (irrelevant information) for visual information (and decreasing it for speech), while low2 represents a decrease in $effort_{within}$ by using different voices to speed up recognition of the relevant information. High1

(for value) was operationalized by presenting less relevant information visually compared to acoustically, while still presenting the same amount of items overall as in the baseline.

Test Subjects

Thirty-four subjects participated in the study (17 men and 17 women). In order to participate, subjects were required to be native German speakers, hold a valid driver's license (class B), and possess good visual acuity (without wearing glasses or contact lenses). The mean age of subjects was 23.88 years with a standard deviation of 2.98. On average, subjects reported driving a vehicle for 2.63 hours a week (SD = 3.97).

Materials

Driving Simulator. The driving simulator at the Centre for Human Machine Systems was a static, real car (Volkswagen Bora) with all relevant in-car interior equipment (e.g., pedal, steering wheel). A screen was placed in front of the car. Processors provided by KMW (Krauss-Maffel Wegmann GmbH & Co. KG, Munich) included one computer visualizing the driving environment, one master entity to transform driver input into movements of the virtual car, and one sound computer to simulate driving sounds. The virtual environment was presented on the screen in front of the car (see Figure 7).

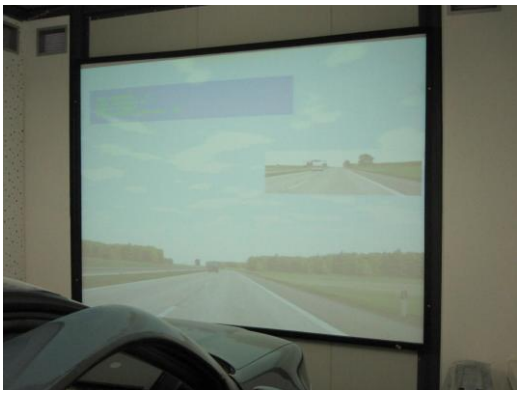


Figure 7. Driving simulator: Virtual driving environment on screen with car in front.

The simulation and operation of the various processors was controlled from a separate experimenter control room, which neighboured the simulator room. The use of a microphone and loudspeakers enabled the experimenter to communicate and instruct test subjects from the other room.

Driving Task. For the driving task, test subjects were instructed to follow a controlled vehicle on the highway, while always maintaining a consistent distance from that vehicle. First, the test subject was instructed to drive on the right lane of the highway at a speed of 120 km/h. After approximately one minute a vehicle appeared onscreen. The test subject had to follow the controlled vehicle at a constant distance. The vehicle's speed changed from time to time, so that higher visual attention was necessary to handle the task. Furthermore, other traffic appeared in the left lane from time to time. To stay in the terminus of the SEEV Model, the driving task related to a relative high expectancy of events (event 1: speed variation of the controlled vehicle, event 2: number of vehicles in the left lane, i.e., other traffic) by which the probability of attention allocation on the street ($P(A_{Street})$) increased. It is important to note that the driving task and driving situation always stayed the same across all conditions and trials. As such, $P(A_{Street})$ was assumed to be constant, and thus could not have influenced $P(A_{Display})$.

Secondary Task. As previously mentioned, the secondary task was a recall task. It was operationalized using a smartphone type HTC Desire (display size = 3.7 inches, Android 2.1) in the form of two Java applications (Framework Android SDK). The first application generated dynamic speech outputs from xml-lists parallel to displaying visual lists through a speech synthesis from SVOX (SVOX German Petra Voice, processing of TTS: non-uniform unit-selection [6]) installed on the smartphone. The second application for implementing different voices for (non-) animals played pre-synthesized prompts. Thus, audio files were pre-generated. Irrelevant information (non-animals) was generated as audio files with the same speech synthesis (SVOX German Markus Voice) and merged together with audio files for the actual animals (generated by another SVOX voice; SVOX German Markus Voice).

Test subjects were instructed to reproduce the names of animals. The relation of relevant information over irrelevant information (number of animals/non-animals) is shown for each condition in Table 2, separated by output modaility (i.e., graphical (display) versus acoustical (speech) presentation). Furthermore, whether or not different voices are used for differentiation of information is also shown.

---

[6] "**Non-uniform unit-selection:** best fitting chunks of speech from large databases get concatenated, minimizing a double cost-function: best fit to neighbour unit and best fit to target prosody. Sounds most natural (similar to original speaker), but inflexible with respect to out-of-domain words and large footprint." (Burkhardt et al. 2010, p. 263)

Table 2. Relation of relevant information to irrelevant information (number of animals/non-animals) as well as the various different voices (yes/no) for the first experiment's four conditions. Red and bold marked information refer to which characteristics are assumed to cause either lowering of $effort_{(within)}$ or increase of relevant information content for speech output but not the display in the four conditions.

		baseline	**IV1: $effort_{within}$**		**IV2: subjective relevant information content**
		baseline	**low1**	**low2**	**high1**
	Different voices	no	no	**yes**	No
Display	Animals/non-animals	3/3	3/**6**	3/3	**1/5**
Speech output	Animals/non-animals	3/3	3/**0**	3/3	**3/3**

Four lists were produced in xml-files for the two respective modalities (i.e., eight lists in total). After the presentation of two lists (graphic and acoustic), subjects recalled the animals' names. As such, animals' names were reproduced in each of the six trials per condition. For each reproduced animal, the test subject was rewarded ten cents. The reproduction of a name more than once was not rewarded again. At the beginning of the experiment, test subjects were informed that in some conditions more animals would be presented visually than acoustically, and vice versa.

To rule out the possibility of learning effects, there were no repetitions across trials. In each condition and trial, three different animals were presented. Each non-animal word was also presented only once. Thus, it was ensured that the information content was equal for all words in terms of different probability of occurrence.

The maximum reward for participation was 7.20 Euros (4 conditions * 6 trials * 3 animals * 10 cents). Subjects' input (i.e., recalled animal names) was recorded by them pressing a button on the steering wheel, which began the recording. The driving task was not interrupted while recalling the animals.

Once the participant was finished recalling the information, he or she simply pushed the same button again to stop the recording. At this point the visual and acoustical presentation of the next two lists commenced. After all trials, test subjects were instructed to stop the vehicle. Another condition was loaded in the form of a new list, and the driving task began again for the second time.

Across all conditions, the display was kept at the same distance from the steering wheel. To start the audio recording, participants pressed a button on the headset. To use the button comfortably, it was connected and soldered with buttons on the steering wheel in the driving simulator.

Performance Recording in the Secondary Task. Subjects' answers were recorded as audio files on the smartphone in order to record their utterances of animal names. Because test subjects received their reward after the experiment, after each trial the experimenter noted the number of animals named. Furthermore, operating time was obtained by pressing the button. This information was also saved by means of a log file.

Eye Tracking. Test subjects' eye movements were recorded with an eye tracker, iViewX head mounted recording System (iView X™ HED) by SMI (SensoMotoric Instruments). This instrument is a mobile, head-based, non-invasive system, which is composed of a baseball cap, a scene camera (to record test subjects' view), and two other cameras (one infrared camera) to capture the pupil position (see Figure 8). The sampling rate is 50Hz.

Figure 8. Eye tracking system, SMI

Raw data was analysed with the Be gaze software. The following AOIs were defined: Street, dashboard and display (of the secondary task). Due to head movements, dynamic AOIs had to be adjusted manually.

Procedure

After subjects were greeted, they were invited to complete a demographic questionnaire. Subjects were then brought to the investigator's room where they were given detailed instructions on the secondary task. Participants then performed a test drive on the driving task. Subjects were informed that they could stop the experiment at any time (e.g., because of driving simulator sickness). Once subjects were familiarized with the simulator, the eye-

tracking instrument (see Figure 8) was put on and calibrated with a 5 points template. Furthermore, the interface allowed the experimenter to check whether pupils were actually being captured during the experiment. In cases where gazes were not captured properly, the experimenter was able to correct this at a later point. After the calibration the four test conditions started. The test conditions (Baseline, low1, low2 and high1) were presented in a randomized order. For every test condition the driving task was started again from the beginning.

3.1.5 Results

In the following the results of experiment one will be presented. In trials where dwell time for the whole scene did not exceed 75% due to poor tracking recognition of the pupil, cases were excluded from analysis.

PDT Display and performance data were analysed for only four of the six trails in each condition, as the first two trials were allocated for subjects familiarizing themselves with the task (i.e., the relation and difference between animals and non-animals).

An exploratory data analysis was conducted for eye tracking data and the number of correctly named animals. All data was tested for normal distribution and outliers were extracted. For repeated ANOVAs, sphericity was tested and in the case of non-sphericity the Greenhouse-Geisser (1959) correction was used.

Eye Tracking Data

H1: Varying the effort$_{within}$ by increasing the time to perceive information will lead to different percentage dwell times for the display.

A one-way repeated measures ANOVA was conducted. A significant effect for the independent variable effort$_{within}$ was found ($F(2,54)=2.63$, $p=0.04$).

H1.1: Lowering the effort$_{within}$ for speech output by lowering the amount of irrelevant speech output information content and increasing the amount of irrelevant information content on the display will lead to a significantly shorter percentage dwell time for the display than the baseline condition.

The baseline condition was compared with condition low1 (a decreased $effort_{within}$ in relation to visual output). A two-tailed paired *t*-test was conducted, and a significant difference was observed (*t*(28)1.90, *p*=0.028, *d*=0.372). As illustrated in Figure 9, lowering the $effort_{within}$ for speech prompts in relation to the visual output leads to shorter percentage dwell time for the display (M=6.53, SD=6.90) compared to baseline (M=9.00, SD=7.18). As such, hypothesis 1.1 is supported.

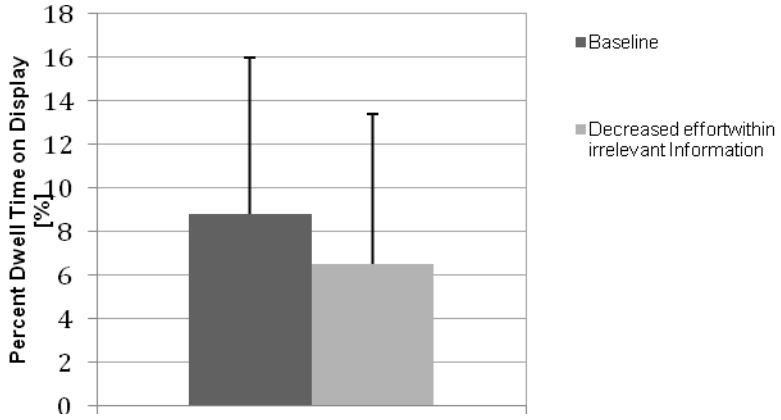


Figure 9. Means and standard deviation of percent dwell time on display for baseline and condition low1.

H1.2: Lowering the $effort_{within}$ *for speech output by using different voices for relevant and irrelevant information (faster discrimination is provided) will lead to a shorter percentage dwell time for the display than the baseline condition.*

The baseline condition was compared with condition low2 (see Table 2). A paired *t*-test was conducted. Although a significant difference was not shown (*t*(28)1.62, *p*=0.056, *d*=0.300), a tendency in the expected direction was observed. Figure 10 shows the means of the two conditions (Decreased $effort_{within}$_different voices: M=7.50, SD=6.57, baseline: M=9.00, SD=7.53).

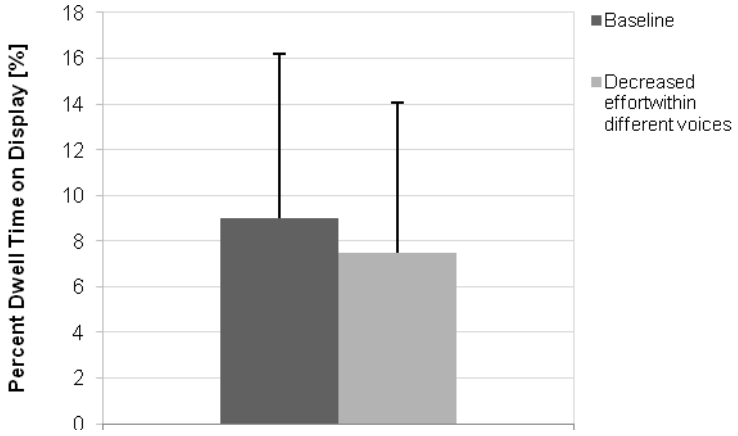


Figure 10. Means and standard deviations of percent dwell time on display for baseline and condition low2

H2: Increasing the relevant information content for speech output by increasing the amount of relevant information in relation to the visual output will lead to a significantly shorter percentage dwell time for the display than the baseline condition.

In the baseline condition every trial involved three relevant items. In the comparative condition for the speech prompt three relevant items were presented, while the visual output consisted of only one relevant item. The relevant information content was thus increased for the speech output. The overall amount of information stayed the same to hold the parameter $effort_{within}$ constant. Again, a two-tailed paired *t*-test was carried out. The results showed a significant effect ($t(28)=1.90$, $p = 0.035$, $d=0.352$). As such, hypothesis 2 was also confirmed (see Figure 11, baseline: M=9.00, SD=7.53, Increased subjective value: M=7.00, SD=5.50).

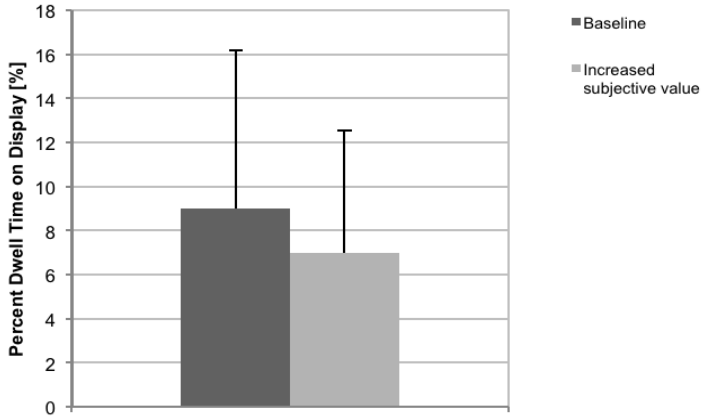


Figure 11. Means and standard deviations of percent dwell time on display for baseline and condition high1

Performance Data

The number of animals recalled was recorded. For the two conditions where the $effort_{within}$ was reduced for speech, no significant difference in the performance (animal recall ability) was found compared to at baseline.

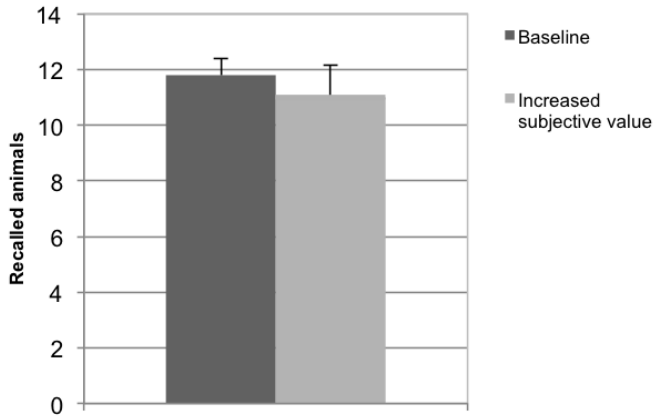


Figure 12. Means and standard deviations of performance data for baseline and condition high1

The Wilcoxon test however showed that animal recall ability was significantly higher in the baseline condition (M=11.78, SD=0.61) as compared to the condition with more presented animals for the speech output in relation to the graphical output (M=11, SD=1.16), p=0.03, see Figure 12.

Driving Data

No significant effects were found for lane keeping performance, variance in distance with varying $effort_{within}$, or for increasing the relevant information content compared to the baseline (all p's >.05).

3.1.6 Discussion

The first experiment was conducted to evaluate two defined parameters' effects on attention allocation: $effort_{within}$ and relevant information content. Test subjects were asked to perform a driving task in a driving simulator, and also to perform a secondary multimodal task involving speech and visual output. Based on the theoretical assumptions in Chapter 3.1, it was expected that the percentage of time spent glancing towards the secondary task display could be reduced by lowering the $effort_{within}$ or by increasing the relevant information content of the speech output in relation to the visual information. Both of these parameters were found to have an influence on PDT Display. $Effort_{within}$ was varied by the amount of irrelevant

information (amount of relevant items were kept the same) compared to baseline, as well as by facilitating participants in judging information relevancy by the use of the different voices. The relevant information content was augmented by increasing the amount of relevant information in relation to irrelevant information of speech output, as compared to visual output. However the overall number of presented items (relevant and irrelevant items) stayed the same as at baseline and in relation to the visual output.

Results indicate that the relevant information content parameter had a significant influence on attention allocation. Higher relevant information content of speech, in relation to graphical information, caused a decrease in percentage dwell time for the graphical output, as compared to a condition where relevant information content for visual and speech output is the same. Participants seemed to allocate less attention to the display if the parameter relevant information content is negatively influenced for visual information (or positively enhanced for speech output). However, in both conditions the same number of animals (three) could be detected via the speech output, although lower performance was observed for having a lower amount of relevant information content presented visually. It was not expected that the probability of attention allocation would be reduced to zero for the display by improving speech output. Further, there may have been trials where participants glanced at the display to gain information visually, whereby there happened to be fewer detectable animals, and as such, they could not possibly have perceived the full amount of relevant information. It is also possible that this effect may be due to the loss in information redundancy between the two modalities. According to Paivio (1971), the memory performance increases by providing the information in terms of more memory codes. Further, it has been proposed that presenting different information acoustically and visually at the same time leads to impaired information processing (Grimes, 1990). Nonetheless, some conclusions may be drawn from the present study: Increasing the relevant information content of speech output to be equal to relevant information content provided by the display is recommended. However, increasing the relevant information content for speech output in such a way that there may end up being more relevant information provided acoustically than visually has two further effects, namely; a positive influence of not encouraging glances at the display, and a negative influence on performance caused by a lack of redundancy.

Results would also indicate that reducing the $effort_{within}$ for speech in relation to visual output leads to a significantly lower percentage dwell for one of the strategies (decreasing the

$effort_{within}$ by lowering the clutter). The effect of adding different voices did not reach significance although a tendency in the expected direction was observed. Therefore it has been confirmed that the time to extract the relevant information ($effort_{within}$) is a parameter that needs to be taken into account for predicting attention allocation. This does however question the way in which this factor differs from the parameter salience. In the presented experiment, relevant items 'popped out' because the voices were different. This is in line with Treisman and Gelade's (1980) definition, which states that more salient stimuli tend to pop out (see also Chapter 3.2). Lower $effort_{within}$ can therefore also be seen as being of higher salience for relevant information. However, in the SEEV model, salience describes the strength of the presented stimuli of an information source in comparison to presented stimuli of other information sources, while $effort_{within}$ in this thesis will be advantaged if only the relevant items are more salient.

It must be considered that due to the fact that the parameter relevant information content for both modalities was influenced ($P(A_{Display})$ and $P(A_{Speech})$), it is not possible to tell if the decreased attention allocation towards the display occurred because of the decrease in $P(A_{Display})$ or the increase in $P(A_{Speech})$. For developing design recommendation, the aim has to be to increase $P(A_{Speech})$ based on the defined parameters in order to enhance speech output of infotainment systems and not to deteriorate the GUI.

To sum up, adding information will increase the factor expectancy of Wickens et al.'s model (2003), but always needs to be set in relation to the $effort_{within}$, and with consideration of the value (relevance) of the information. Adding relevant information increases relevant information content (expectancy and value) and the $effort_{within}$. Adding irrelevant information with no value on the other hand increases the $effort_{within}$ alone. As redundancy of relevant information is obviously an important factor for the performance of the secondary task and relevant information content for attention allocation, it is recommended to make sure that the same amount of relevant items of infotainments systems are presented in both modalities, and at the same time, to keep the time effort low for speech output.

3.2 Disadvantage of Speech 2: Less Controllability

After analysing the characteristics of speech output compared to visual output on the parameters relevant information content and $effort_{within}$, in the following the bottom up triggered parameters salience and $effort_{between}$ will be reviewed more closely.

The objective of early research on selective attention was to explore the influence of salience on bottom up attention allocation. In Treisman's feature integration theory (1980), the pop out effect was first proposed. Faster detection occurs if objects are being differentiated by one feature alone, such as colour, orientation or movement. Itti, Koch and Niebur (1998) developed the concept of saliency maps to predict attention allocation based on salience. According to this theory, different features of colour, intensity and movements are extracted and laid out in one conspicuity map. Furthermore, the intensity of how much the information differs compared to its surrounding environment is analysed and integrated in the saliency map. In addition to salience maps for the visual modality, others were also developed for acoustical attention allocation: intensity, temporal frequency contrast, orientation, pitch, as well as the centre surrounding differentiation (Kalinli & Narayanan, 2007, Kayser, Petkov, Lippert and Logothetis, 2005). Salience is not one of the factors for which visual information is assumed to be extra beneficial if the salience is not intentionally decreased by the developer. Contrariwise, auditory information can be more salient (Ortega et al., 2009). Acoustical stimuli seem to catch attention more easily, as they are omnidirectional (Proctor & Proctor, 2006). Compared to the visual modality, no lower value can be detected for salience of the auditory modality. Given the intrusive nature of acoustic information, acoustical stimuli can often be used as a means to indicate alarms.

$Effort_{between}$ was defined as switching costs: the effort to allocate attention from one information source to another (see Chapter 2.2.1). Wickens and McCarley (2008) propose that with higher distance from the fovea, salience of visual stimuli decreases rapidly. Indeed, the factors salience and $effort_{between}$ are not completely independent. Intrusiveness and whether information is omnidirectional can therefore also be considered in terms of the factor $effort_{between}$. As acoustical stimuli are omnidirectional only a task switch and no location switch (see Chapter 2.2.1) has to be performed allocating attention from the street towards the acoustical output of the secondary task. For visual output both a location switch has to be performed. Thus, $effort_{between}$ is lower for acoustical stimuli within the secondary paradigm with the same task performed.

To sum up, for attention allocation, low $effort_{between}$ and high salience of stimuli theoretically increase attention allocation: hence, in-car infotainments systems with speech output have a benefit over visual information regarding the two parameters $effort_{between}$ and salience due to their omnidirectional character. However, it was assumed that in the special case of in-car

infotainment systems (which are not part of the primary task of driving safely), the benefit of acoustical information being omnidirectional can also become a disadvantage.

The absence of a location switch and therefore the absence of a consciously performed motor action to get access to the information for the acoustical modality lead to a low controllability of speech output. Therefore, the speech output requires attention in the very moment the information is accessible (system-initiated). In contrast, for visual information the driver can initiate an attention shift at any time he or she wants, due to the location switch (increased $effort_{between}$). The nature of this kind of information retrieval involves driver initiation, which means that the driver is able to strategically decide the start of the second stimuli, based on the characteristic of the primary task, when and whether to allocate attention to the secondary task. In the following it will be more closely described which cognitive processes and their limitations support the benefit of driver-initiated start of secondary task output information.

According to the one channel theory, information is processed in series (Broadbent, 1958). Pashler and Johnston (1998) showed that a decrease of stimulus onset asynchrony (SOA) between two stimuli leads to longer reaction times for the second stimulus, while reaction time to the first stimulus will be unaffected. SOA defines the time between the start of a first stimulus and the start of a second stimulus. The longer reaction times are due to the so-called bottleneck that occurs in the serial processing of information. The second stimulus has to wait until the first stimulus has been processed. According to Posner, Snyder and Davidson (1980), the bottleneck occurs on the level of response selection of human information processing (human information processing; Wickens, 1992). The perception of the second stimulus can be accomplished although the first stimulus still needs to be processed. The more complex the processing of the stimulus the longer the second stimulus has to wait. In the case of a complex primary task (complex processing of the first stimulus), the perceived information of the secondary task has to be stored in short-term memory. In Baddeley's working memory model (1986), two different short-term memory stores are defined: the phonological loop and the visual-spatio sketchpad. Both are proposed to be regulated by the central executive. The central executive communicates with long-term memory and is responsible for higher order information processing. While the phonological loop is responsible for short-term speech storage, the visual sketchpad holds visual and spatial information. Phonological capacity is on the one hand determined by temporal duration of the information that needs to be stored (word length effect, Baddeley, Thomson & Bachanan, 1975). On the other hand, information stored in the phonological loop is lost within about two seconds if it is not refreshed

(Baddeley et al., 1975). The following assumptions are made: if secondary task information must be stored for too long in the phonological loop because of a very complex first task being performed, then the secondary task (appearing shortly after the first stimulus) can not be accomplished. In the context of this thesis (whereby the primary task is driving, and the secondary task is using an infotainment device with speech and visual output), that indicates that in the case of a very difficult driving task (where time of response selection is increased) and a short time period of system-initiated output appearing after that task, the secondary task information can not be processed auditorily and is therefore interrupted. Due to the location switch (and thus increased $effort_{between}$), the start of the visual output of an in-car infotainment system (second stimuli) is driver-initiated while the start of speech output is system-initiated. As the 'start' of the second stimuli for visual output of the in-car infotainment task can be influenced by the driver neither interruption nor longer short-term storage occur.

With regard to the question as to which driving tasks put high demand on processing, Donges (1982) categorized driving in terms of vehicle control, guidance and navigation, and arranged these categories into these groups dependent on the degree of cognitive load they entail for the driver (see Figure 13). In Figure 13 you can also find Rasmussen's (1983) three level model, which can be aligned to the three categories of driving. Vehicle control describes skill-based tasks where automated sensori-motor patterns are retrieved; this refers to stabilization of the vehicle in terms of longitudinal and lateral stabilization. Guidance implies a task such as lane changing or interaction with other vehicles; tasks that are knowledge, rule or skills based. Navigation tasks require planning and are knowledge-based, e.g., finding a new route or organizing the sequence of a route.

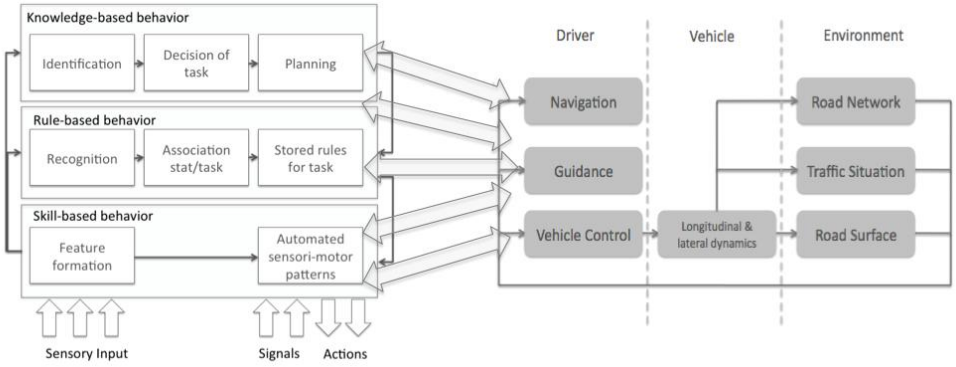


Figure 13. Rasmussen's (1983) three level model and Donges' (1982) driving task model – allocated to one another according to Donges (2009)

In a driving simulator test, it could be shown that for highly automated tasks (braking according to hazards appearing), the reaction time increases when the stimulus onset asynchrony of a secondary choice task decreases (Levy et al., 2006). It is assumed by Levy et al. (2006) that in the case of more demanding tasks (e.g., navigation and guidance) as compared to automated tasks, an even stronger bottleneck effect will occur. Therefore the time to start the second stimuli (information presentation of the secondary task) should be during phases of vehicle control or the guidance task which requires less difficult response selection compared to knowledge or rule-based tasks. Janssen and Brumby (2010) observed that "people can strategically control the allocation of attention in multitask settings to meet specific performance criteria" (p.1548). As previously mentioned, the start of the second stimulus for the visual output can be consciously influenced by glancing away from the street towards the display during phases of lowly demanding driving tasks. For acoustical (omnipresent) stimuli such as speech prompts however, due to the low $effort_{between,}$ the start is system-initiated and in consideration of sequences of highly demanding driving tasks must be held in the phonological loop. In Baddley's working memory model, the demands involved in using the phonological loop and driving at the same time negatively influences driving performance, as well as the performance of the secondary task (as compared to no driving condition; Radeborg et al., 1999).

Thus, it is assumed that the attention allocation towards the display will increase if a difficult driving task (like navigation or guidance) is shortly followed by a speech prompt and therefore increases the time till the speech output can be processed (secondary task is interrupted). If it were possible that access to speech prompts is only given via a consciously performed motor action it would be assumed (according to the SEEV Model) that with this increase of $effort_{between}$ (a more demanding motor action to access the speech output) the attention allocation on the display would increase. However, for situations where a highly complex driving task is shortly followed by a speech prompt (interruption of the secondary task), an increase of $effort_{between}$[7] higher than zero would then lead to less percent dwell time on display. The $effort_{between}$ needs to be increased until controllability is achieved. After controllability is achieved further increase of $effort_{between}$ is assumed to again result in a linear distribution: the higher the $effort_{between}$ the higher the percent dwell time on display.

[7] In the following by $effort_{between}$ it is only referred to the strength of the motor action that needs to be performed to get access to the information source (such as a location switch between two visual tasks). Task switch is the same for visual and acoustical output in the presented context therefore it is disregarded.

The relationship between workload and unpredictability in adaptable and adaptive automation compiled by Miller and Parasuraman (2007) describes a similar concept. The concept is shown in Figure 14. Workload describes the amount of "attentional and cognitive energy" of a user to operate a system and therefore can be compared to $effort_{between}$ (Miller & Parasuraman, 2007, p.60). Unpredictability is described as the opposite of "being in control" (Miller & Parasuraman, 2007, p.60). With an increase in controllability, workload also increases. For adaptive automation (i.e., where the system sets the level of automation) the workload is low but the operator in turn has less control. For adaptable automation (i.e., human delegation and selecting the level of automation), workload increases as controllability increases. Applying this approach to the outlined example of secondary task and driving, a driver-initiated (adaptable) request for speech output is possible by means of controlling the level of effort to request the speech output on an adequate level. Although it is assumed that this leads to a higher $effort_{between}$ to get information access, an $effort_{between}$ close to zero for speech is not ideal as this may extend the duration of glances at the display. Nevertheless, it should be taken into account that by increasing the $effort_{between}$ or the energy to get access to the information 'too much', it may result in a higher probability of attention allocation towards the display.

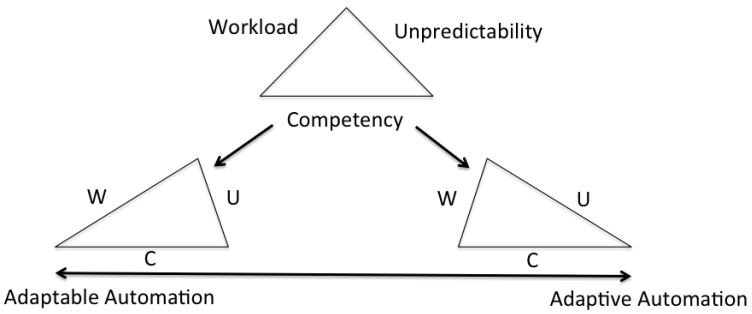


Figure 14. Relationship between workload and unpredictability in adaptable and adaptive automation (Miller & Parasuraman, 2007)

3.2.1 Experiment 2: Influence of $Effort_{between}$ on Attention Allocation

One goal of the following experiment was to examine if increasing the $effort_{between}$ for speech will lead to an decrease of $P(A_{Speech})$. The percent dwell time on the display was again measured as a reverse parameter for $P(A_{Speech})$. $Effort_{between}$ for speech output was increased by introducing a button for accessing the speech prompt (whereby a motor action had to be performed). The second goal was to examine the special situation for which the increased $effort_{between}$ was of benefit and therefore resulted in an increased $P(A_{Speech})$: complex driving task shortly followed by a speech prompt which leads to interruption of the secondary task.

Increased $effort_{between}$ for speech via an introduced conscious motor action increases controllability of speech and thus the time to start the speech output is then driver-initiated. Thirdly it was of interest to measure the effects of increasing the $effort_{between}$ 'too much': Would this imply that it would then, again, lead to more time spent looking at the display (decreased $P(A_{Speech})$), in line with the assumption of Wickens et al.'s (2003) SEEV Model, and in line with the inhibiting nature of the factor $effort_{between}$?

3.2.2 Hypotheses

H3: For different levels of $effort_{between}$, there are differences in the tendency to allocate attention towards the display. Further, there will be an interaction effect between the variables $effort_{between}$ and interruption.

H3.1: As compared to a system-initiated speech output ($effort_{between}$ low) a driver-initiated speech output ($effort_{between}$ medium) increases the percentage dwell time to the display if a highly complex driving task is shortly followed by the speech prompt and thus an interruption of the secondary task is materialized.

H3.2: Further increasing the $effort_{between}$ to access speech output of secondary tasks ($effort_{between}$ high) in turn increases the percentage dwell time to the display in the case of a highly complex driving task being shortly followed by a speech-prompt and thus leads to an interruption of the secondary task.

H3.3: In the case of a consistently lowly-demanding driving task, a decreased percentage dwell time to the display is expected for an increased $effort_{between}$ medium and high compared to a low $effort_{between}$.

3.2.3 Operationalization

As was the case in our first experiment (to test the effects of relevant information content and $effort_{within}$), a secondary task was performed while driving. Again, a list of words with animals and non-animal words was presented visually and acoustically, and subjects were asked to recall the animals. The $effort_{between}$ was increased for acoustical output by providing a user-initiated request for speech output. For this purpose, a functional button was attached to the steering wheel. This button could be pressed once to begin the prompting. In the condition of low $effort_{between,}$ the system initiates the speech output by itself. To further increase the $effort_{between}$, the button had to be pressed three times, upon which access to the speech output

was granted. Thus for the independent variable $effort_{between}$ three variations were operationalized.

As mentioned in Chapter 3.2, the effect of a small $effort_{between}$ compared to no $effort_{between}$ is especially relevant in situations whereby the secondary task is interrupted by the driving task because a complex driving task is shortly followed by the secondary task stimulus. In our experiment, the effects of different levels of $effort_{between}$ on attention allocation were measured under the conditions of interruption and no interruption. As such, in the interruption condition, mathematical tasks had to be completed while both driving and conducting the secondary (animal) task. The mathematical task simulated a highly cognitive demanding task that requires a response selection as navigation and guidance tasks. The mathematical task was presented during the secondary task output and thus, like a highly cognitively demanding driving task (e.g., navigating in an unfamiliar city), will presumably delay processing the secondary task output. Mathematical tasks place demands on the phonological loop as well as the central executive, and therefore can be said to be of increased difficulty. Further, Manzey (1988) showed that mathematical tasks interfere strongly with verbal memory tasks (Sternberg-Paradigm, 1966). As the secondary task was of a verbal recognition nature, the mathematical task presented was expected to interfere strongly with the secondary task.

3.2.4 Method

The experiment was conducted in the laboratories of the Telekom Innovation Lab.

Experimental Design

The independent variables were all tested using a within-group design. Six driving trials with different variations of the two independent variables $effort_{between}$ and interruption were completed by the participants (see Table 3). In order to rule out learning effects, the order was randomized.

Table 3. Experimental design: IV1 and IV2. The abbreviations within the cells describe the level of $effort_{between}$ and if the secondary task is interrupted by the mathematical task ([level of $effort_{between}$]_[interruption or no interruption])

	IV1: $effort_{between}$			
		Low	**Medium**	**High**
IV 2: Interruption	**No interruption**	Eflow_nI	Efmed_nI	Efhigh_nI
	Interruption	Eflow_I	Efmed_I	Efhigh_I

Test Subjects

Twenty-eight people participated in the study. Due to invalid data, 3 subjects' data were excluded from analyses. This left 13 male and 12 female subjects' data to be analysed. The mean age was 28.2 years (SD = 4.44) with a minimum age of 20 and a maximum age of 37. The mean duration of holding a driver's license was 9.4 years with a standard deviation of 4.34. Mean hours per week driving was 6.2 hours, SD= 8.95. The test subjects indicated to be in mean "highly" technical affine.

Materials

Driving Task. The Lane Change Task (LCT) by DaimlerChrysler AG, Research and Technology was used as the driving task. According to Mattes (2003), the task is equally cognitively demanding to driving, and therefore a good alternative for driving simulators, despite the fact that it is not possible to create different driving situations or programming occurrence of other vehicles. Figure 15 shows the experimental set up. A monitor as well as a Logitech driving force steering wheel (Driving Force GT) with a foot panel was placed in front of the test subject. The PC-based simulation ran on a laptop in the back. The Lane Change Task is freeware and an instrument that is frequently used to simulate and measure driving performance. The route of the Lane Change Task consists of a three-lane motorway. Next to the motorway, 18 street signs are set with assignments of which lane to choose.

Figure 15. Experiment setup – Lane Change Task

Subjects were prompted by signs regarding when to change lane. The viewing distance for the signs was set at 50 meters. In other words, 50 meters before reaching the sign, the target lane was presented. The speed limit for the duration of the experimental trial was set at 50 km/h. The track length was 3000m, which allowed for 2 min of data collection (Mattes & Hallen, 2009). The secondary task commenced when a start sign was shown at the beginning of the route. If the secondary task was completed, the experimental conductor set a marker. All participants in all conditions successfully completed the task. If the full route was completed after the 18 signs, a u-shape curve would appear and the route would then start over again. The resulting raw data were analysed using LCT Analysis software. Driving performance was compared against an optimal lane changing behaviour, and the mean deviation as well as the driving duration for each trial was calculated.

Participants were informed that the driving task should always be treated with higher priority than recalling animals (i.e., the secondary task) and completing mathematical tasks. As the subjects gained monetary rewards for every recalled animal and every correct mathematical task, subjects were told that they would lose a significant amount of money by failing to drive in the correct lane[8].

Secondary Task. The secondary task involved participants being asked to recall animals presented both visually and acoustically during the driving task. The HTC Desire java application introduced in Chapter 3.1.4 was used for this purpose. For all of the six

[8] Note that failing to drive the correct lane was not actually analyzed afterwards and was not punished with a loss of money. Test subjects were only able to gain money.

conditions, two equal xml-lists (12 lists in total) were prepared and uploaded to the smartphone's memory card. As in the first experiment, in one of the conditions the TTS SVOX German Petra Voice ($effort_{between}$ low) read out the xml-lists automatically whilst the information was concurrently presented visually (on the display, see Figure 16). Participants were told to press the record button on the steering wheel before recalling relevant information, and to press it again to start a new list. The Logitech steering wheel did not have buttons available to connect the headset keys via a solder junction, so a stylus was attached to the steering wheel for this purpose instead (see Figure 16). The headset buttons were connected with the headset keys of the mobile device.

Figure 16. Experimental setup with an up-close view of the steering wheel and buttons (left) and the position of the smartphone (right)

Two further applications were also developed. A function whereby the TTS read out the lists was implemented and activated by pressing a button on the headset either once or thrice. For this purpose, another button on the headset (next to the record button) was connected with a stylus on the steering wheel. In the increased $effort_{between}$ ($effort_{between}$ medium) condition, participants were told to press the blue button (Figure 16) to request the speech output (synthetic output read aloud) while the visual output was again presented automatically. For the high $effort_{between}$ condition, the blue button had to be pressed three times. Participants were not specifically instructed to press this button, but were simply informed that it was an option. Where the second button application was integrated, a yellow button was used as the record button.

All conditions (see Table 3), and therefore all driving transitions, consisted of seven trials. After each condition, the driving task restarted from the beginning. Each trial represented two equal lists (visual and auditory) with and without animals. After each list was presented, participants' verbal responses regarding perceived animals was recorded. To do this, participants pressed the yellow button on the steering wheel and responded. After pressing the

button again, a new list would load and either begin to read aloud automatically or upon pressing the blue button (either once or thrice). One of the experimental coordinators recorded the number of correctly recalled animals in an excel sheet. For each correctly recalled animal, ten cents was rewarded.

Holding information theory in mind, each animal appeared only once, in order not to influence the probability of occurrence and thereby change the informational content.

Eye Tracking. For measuring eye-tracking data, SMI's (SensoMotoric Instruments) iViewX headmount recording system (iView X™ HED) with a 50Hz sample rate was used again. A 5-point cardboard was used to calibrate the eye tracking (see Figure 8 & 17).

Figure 17. Calibration (left) SMI's iView X™ HED screen shot (middle) presentation of mathematical tasks (right)

Mathematical Tasks. The mathematical tasks used for the interruption condition were presented on the wall using a beamer behind the monitor as (see Figure 17). The mathematical tasks started in the first trial one second after starting the secondary task. In the six other trials, the math task started one second after pressing the record button the second time (which ended the recording). As such, the seven math tasks were consistently presented at the same point in time across trials. The math task disappeared after three seconds. Participants were asked to give their answer verbally while the math task was still present. If participants failed to answer correctly or in time, a note was made of this. Each correct answer was rewarded with 1Euro to make sure that the mathematical task was of higher value then recalling the animals (representation of higher relevance of the primary task compared to the secondary task).

Questionnaire. In order to measure self-reported mental workload experienced while driving and completing the secondary task, the SEA scale (Eilers et al., 1986) was completed after

each trial (condition). The SEA scale is a one dimensional measurement that uses verbal anchors.

Procedure

After filling out the demographic questionnaire, participants were given general instructions and specific instructions for the driving task. After training to do the lane change task, the eye tracking system was mounted and calibrated. Subjects were then introduced to the secondary tasks. In randomized order, the six trials were then completed. The SEA scale was filled out after each trial. At the end of the experiment, participants were given their reward, as well as a fixed sum for their participation.

3.2.5 Results

Data was analysed for normal distribution and outliers. Outliers were extracted. In trials where dwell time for the whole scene did not exceed 75%, data were excluded from analyses.

In cases where a repeated ANOVA was conducted, sphericity was tested using Mauchly's test. In the case of significant findings, the Greenhouse-Geisser Test was used.

Eye Tracking Data

H3: For different levels of $effort_{between}$ *(main effect), there are differences in the tendency to allocate attention towards the display. Further, there will be an interaction effect between the variables* $effort_{between}$ *and interruption.*

A two-way repeated ANOVA was conducted. The main effect, $effort_{between}$, was found to be significant ($F(2,40)=3.710$, $p=0.033$, $\eta^2=0.156$), as was the interaction effect ($F(2,40)=5.03$, $p=0.011$, $\eta^2=0.201$). The main effect for interruption was not found to be significant. As such, the null hypothesis was rejected. For means and standard deviations see Figure 18 and Table 4.

Table 4. Means and standard deviations PDT Display for the three level of $effort_{between}$ in the condition of no interruption and interruption

Effort	Interruption	Mean	Standard Deviation
Low	No Interruption	2.76	2.70
	Interruption	4.97	4.49
Medium	No Interruption	2.76	2.91
	Interruption	2.59	2.73
High	No Interruption	3.60	3.24
	Interruption	3.08	3.20

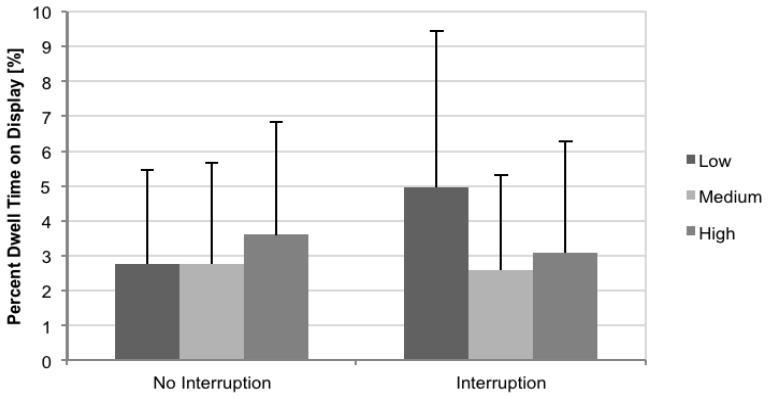


Figure 18. Means and standard deviations PDT Display for the three level of $effort_{between}$ in the condition of no interruption and interruption

H3.1: As compared to a system-initiated speech output ($effort_{between}$ low) a driver-initiated speech output ($effort_{between}$ medium) increases the percentage dwell time to the display if a highly complex driving task is shortly followed by the speech prompt and thus an interruption of the secondary task is materialized.

According to the significant main effect $effort_{between}$ observed, a paired *t*-test was conducted post-hoc. An interruption in the secondary task by a mathematical task was found to lead to a significantly higher percentage dwell time towards the display when the $effort_{between}$ is very low compared to at a medium level ($t(22)=3.22$, $p=0.002$, $d=0.671$). As such, Hypothesis 3.1 was supported.

H3.2: Further increasing the $effort_{between}$ to access speech output of secondary tasks ($effort_{between}$ high) in turn increases the percentage dwell time to the display in the case of a highly complex driving task being shortly followed by a speech prompt, and thus leads to an interruption of the secondary task.

Another post-hoc *t*-test was completed for the dependent variables. There was no significant increase in percentage dwell time towards the display when $effort_{between}$ was increased from medium to high (p=0.571). A significant difference occurred between high and low levels of $effort_{between}$ in terms of an increased percentage dwell time towards the display for the lower $effort_{between}$ (t(20)=2.1, p=0.049, d=0.457), see Table 4. However, the effect power was higher for the difference between low $effort_{between}$ and medium $effort_{between}$ (d=0.671), as compared to the difference between low effort and high $effort_{between}$ (d=0.457). Additionally, the test for contrast effects showed a significant effect for a squared contrast effects (p=0.01).

H3.3: In the case of a constant lowly-demanding driving task, a decrease in percentage dwell time to the display is expected for an increased $effort_{between}$ medium and high compared to a low $effort_{between}$.

No significant effect was shown for the percent dwell time towards the display if the $effort_{between}$ was increased by means of a driver-initiated request for speech output. However, pressing the button three times ($effort_{between}$ high) leads to a significantly higher percentage dwell time towards the display (t(23)=1.98, p= 0.03, d=0.404) compared to low $effort_{between}$. No significant effect for the linear contrast effects was found. Hypothesis 3.3 was thus partly supported.

Subjective Workload

A main effect for interruption was observed (F(1,24)=19.45, p<0.001, η^2=0.448), as well as an interaction effect between interruption and $effort_{between}$ (F(2,48)=5.01, p=0.01, η^2=0.173). As seen in Figure 19 and Table 5, subjective perceived workload was higher for a driving task with an interruption in the secondary task than for driving with a secondary task but without an interruption. The main effect $effort_{between}$ was not significant.

A tendency was found to occur in the expected direction, and an interaction effect was found for the dependent variable. Also testing for linear contrast showed a significant U-shaped distribution for interruption (p=0.033) and linear distribution for no interruption (p=0.018) concerning the $effort_{between}$. See Tables 6 and 7 for means and p-values.

Table 5. Means and standard deviations SEA scale for the three level of $effort_{between}$ in the condition of interruption and no interruption.

Effort	Interruption	Mean	Standard Deviation
Low	No Interruption	8.34	3.55
	Interruption	12.53	4.45
Medium	No Interruption	8.79	3.34
	Interruption	10.54	3.51
High	No Interruption	9.71	4.04
	Interruption	10.79	4.40

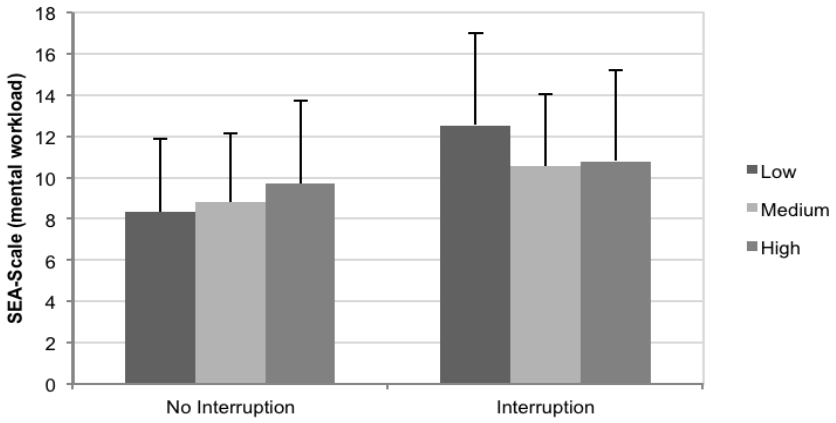


Figure 19. Means and standard deviations SEA scale for the three level of $effort_{between}$ in the condition of interruption and no interruption.

Table 6. Mean differences, *p*-values and power for the comparison of the three level of $effort_{between}$ for the no interruption condition.

$effort_{between}$		mean difference	p-value	η^2
low	medium	-0.44	0.344	0.19
low	high	-1.37	0.046	0.45
medium	high	-0.93	0.141	0.30

Table 7. Mean differences, *p*-values and power for the comparison of the three level of $effort_{between}$ for the interruption condition.

$effort_{between}$		mean difference	p-value	η^2
Low	medium	1.99	0.011	0.55
Low	high	1.74	0.045	0.42
Medium	high	-0.25	0.727	0.07

Driving Performance

A two-way repeated ANOVA for the independent variables $effort_{between}$ and interruption was conducted for the mean deviation from the ideal lane keeping behaviour. No significant effects were found (all *p*'s >.05).

Task Duration

Along with mean lane deviation, the lane change task analysis software provided us with the length of distance covered, which in turn represents the operating time of the secondary task. Again, a two-way repeated ANOVA was conducted. A significant main effect was found for the independent variable interruption ($F(1,24) = 12.99$, $p = 0.001$; $\eta^2 = 0.351$). According to the means (Table 8), it would appear that interruption occurrence lead to longer duration of operating tasks, as compared to when there was no interruption. No interaction effect and no significant main effect for $effort_{between}$ were shown.

Table 8. Means and standard deviations for length of distance (in kilometres) covered for the three levels of $effort_{between}$ in the condition interruption and no interruption.

Effort	Interruption	Mean	Standard Deviation
Low	No Interruption	2.63	1.99
	Interruption	2.78	4.72
Medium	No Interruption	2.64	3.47
	Interruption	2.88	5.25
High	No Interruption	2.71	4.13
	Interruption	2.94	6.54

Performance on the Secondary Task and Mathematical Task

For performance on the secondary task, no differences were found for the two independent variables $effort_{between}$ and interruption. Further, no interaction effect was observed (all *p*'s >.05).

A one-way ANOVA was conducted to test if differences in the mathematical tasks are observed for different levels of $effort_{between}$. No effect was observed (all *p*'s >.05).

3.2.6 Discussion

According to the SEEV Model, an increased effort of switching attention to a stimulus would decrease the probability of allocating attention towards it (Wickens et al., 2003). In the dual task paradigm of driving and conducting a secondary task, the effort of switching attention from the street to a display appears to be linked to a motor reaction in terms of eye movements. Due to the omnipresence of acoustical information, less effort for attention allocation switch to speech output is assumed since no location switch needs to be performed. However, this is not necessarily always a benefit for speech in the special situation of a dynamic primary task. In the present study it was assumed that in the case of higher cognitive demand of the driving task shortly followed by the speech output, which causes a delay of the processing of the secondary task, a self-initiated attention allocation switch related with a higher $effort_{between}$ for acoustical information would increase the controllability. The attention allocation to speech output would thus in turn increase, and less glancing at the display would occur. Information acquisition of the secondary tasks requires being adaptable to the demands of the primary task.

It was shown that in a driving task mostly characterized by lane keeping and lane changing (vehicle control and guidance), a low $effort_{between}$ of acoustical information in the sense of a system-initiated speech output causes the lowest percent dwell time on display. This is in line with the assumptions of the SEEV Model. Notice that there was no significant disadvantage for the medium level of $effort_{between}$ (driver-initiated speech prompt by pressing a button once) while a high level of $effort_{between}$ (driver-initiated speech prompt, achieved by pressing a button thrice) actually increased the tendency to allocate attention towards the display. Information acquisition was ‘easier’ from visual output and by looking away from the street than to press the TTS list-reading initiation button three times. The same distribution was found for the subjective perceived workload. Compared to high $effort_{between}$, low $effort_{between}$ resulted in lower workload. Again, a medium level of $effort_{between}$ did not cause any disadvantages compared to a low level of $effort_{between}$ in terms of a higher perceived workload.

As assumed, if the secondary task is interrupted by a highly cognitively demanding third task, a benefit was observed for a medium-level $effort_{between}$, as compared to low-level $effort_{between}$. People spent significantly less time looking at the display with a driver-initiated request function for speech output, as compared to a system-initiated speech output. Low $effort_{between}$ led to a higher percent dwell time on the display compared to high $effort_{between}$ in interruption

conditions. Again, the same was found for subjective workload. Against the assumptions, no significant increase in percent dwell time on the display was found for increasing the $effort_{between}$ beyond a medium level compared to a high level. For the percent dwell time on the display in the interruption condition however, a tendency in the expected direction was observed, as well as a significant squared contrast effect.

For driving performance, no differences could be observed for the different levels of $effort_{between}$. This could be due to a lack of measurement sensitivity. Further, it was observed that moderate levels of cognitive load lead to lower lane-keeping variance (Engström et al., 2005; Greenberg et al., 2003, both in Young et al., 2009).

Another critical aspect that needs to be considered is the assignability of increasing effort and the concept of controllability or unpredictability as in Miller and Parasuraman‘s model (2007). According to the DIN ISO 9241 Dialogue Strategies, controllability is described as how “the user should be able to control the pace and sequence of the interaction”. In the present study, this is indeed what was achieved by increasing the $effort_{between}$. However it is important to point out that it cannot be confirmed whether a high value of $effort_{between}$ is precisely the same as controllability. $Effort_{between}$ for example also includes task switch costs, which were not examined in the present study.

To conclude, the results indicate the recommendation of a driver-initiated request for speech output of secondary tasks. This would reduce the tendency to glance at the display in situations where the secondary task is interrupted by a highly demanding primary task. Additionally, in the driving task with guidance and vehicle control (lower cognitive demand), no disadvantage was observed for driver-initiated requests. However, in the case of a high $effort_{between}$ to access information (pressing the button for speech output three times), the percent dwell time towards the display again increases. Therefore, it is also recommended to take care that the $effort_{between}$ is of an adequate level. Next to these insights it could be proven that effort of switching attention to an information source is also relevant for acoustical attention allocation.

3.3 Summary

In the first experiments, it was examined whether increasing time effort to perceive the relevant information of an event or area of interest has an inhibiting effect on attention

allocation. It was shown that decreasing the $effort_{within}$ for speech in relation to graphical information by decreasing the amount of irrelevant information for speech and increasing the amount of irrelevant information for graphical information leads to less percent dwell time on the display. According to these results, $effort_{within}$ should be taken into account for predicting attention allocation. $Effort_{within}$ should be seen relatively to the relevant information content. Increasing the relevant information content of speech output is only recommended when concurrently taking care of the time effort. If an increase of relevant information content for speech is obtained by not increasing the time effort, it leads to less percent dwell time on the display (also shown in experiment 1). Speech output, as compared to the same output but delivered visually, is very time consuming. In particular, increasing the information content for speech output could have negative effects on the value of $effort_{within}$. Next to that in the first experiment the importance of presenting the same amount of information visually and acoustically (redundancy of different modality outputs) was shown.

In the second experiment it was shown that an increase of effort to switch attention towards a stimulus ($effort_{between}$) is also relevant for acoustical attention allocation; however, in the context of in-car infotainment systems, this effect does not have a simple linear inhibiting influence on the probability of attention allocation. Thus, differentiation of $effort_{between}$ and $effort_{within}$ is recommended; especially for comparing the benefits and disadvantages of speech and displays for in-car infotainment systems by manipulating the aforementioned parameters. Switching attention from the street towards the display of an infotainment system is not only associated with a task switch, but also a location switch. This motor reaction is not necessary if attention needs to be switched from the driving task towards the speech output. However, the decreased $effort_{between}$ for speech is related to a lack of controllability. This controllability is essential in situations where the secondary task gets interrupted and the point of time presenting the speech output should better be driver-initiated. $Effort_{between}$ (and thus controllability) was increased by introducing a button at the steering wheel. This resulted in a decreased attention allocation to the display: the start of the speech output could be adapted in terms of the driving situation.

The following consideration has not so far been made in this thesis: Introducing the button at the steering wheel could not only result in an increase of $effort_{between}$ and increased controllability; As the speech output could then be repeated by pressing the button again, the availability of speech was thereby increased. Next to the trade off between $effort_{within}$ and

relevant information content, as well as the decreased controllability for speech output, the lack of availability of speech output could also potentially be a characteristic that leads to increased attention allocation towards the display. Speech output is non-permanent and presented only once while visual output is constantly presented (i.e., permanently). The decreased availability could be allocated to the factor expectancy of the SEEV Model. With the button the frequency of speech output presentation can be increased. However, note that this frequency is different from the frequency of refreshes or changes (see Chapter 2.2.2): Frequency only increases with every new event. Thus the Model's definition of expectancy eventually needs to be extended by $expectancy_{availability}$. $Expectancy_{availability}$ can be described as expectancy that the relevant information is constantly presented until the user/driver has perceived all of the relevant information. In the future, it should be investigated whether the lack of availability is one of the factors that causes increased attention allocation towards the display. This was not addressed in the present thesis. As the attention allocation towards the display was not decreased in the second experiment for introducing the button in situations were no interruption occurred, I assume that either the increased $effort_{between}$ as an inhibiting factor lead to no decrease, or alternatively, that the factor availability is only relevant in special situations. As mentioned before, phonological capacity is determined by temporal duration of the information that needs to be stored (Baddeley et al., 1975). For secondary tasks where the amount of relevant information presented is increased, an increased $expectancy_{availability}$ could be of benefit as it relieves the phonological loop. Next to evaluating these assumptions (by increasing the amount of relevant information or by repeating the speech prompt without the need of pressing a button), it could be tested whether decreasing the actual availability of visual output in fact decreases attention allocation towards the display.

In the following part, the findings shall be adapted to an applied context of performing an infotainment task while driving.

4. Transfer to Applied Context: Design Implications for In-Car Speech Output

The following sections detail design recommendations deducted for an actual infotainment system (Chapter 4.1. & 4.2), based on the insights derived from the first two experiments. The deducted design recommendations were evaluated in a driving simulator study to find out if they actually result in lower tendency to allocate attention towards the display (Chapter 4.2.4). In light of this, the object of the study was to evaluate which methods of parameter manipulation for in-car infotainment systems avoid visual distraction without explicit use of only acoustical output instead of visual output. The implemented speech output design recommendations for reducing attention allocation to the display were then tested separately in terms of their influence on users' mental models (Chapter 4.3.). In addition to deducting more pragmatic design recommendations (as will be described in Chapter 4.2), it was also investigated as to whether the influence of increasing hedonic value leads to spending less time glancing at the display (Chapter 4.4). In the next chapter, the prototype used for implementing the design recommendations based on the SEEV Model will be presented.

4.1 Connected Life and Drive (CLD) Prototype

Within the Telekom Innovation Laboratories project, 'Connected Life and Drive' (CLD), a prototype with in-car information and communication applications that can be used while driving was build (see Figure 20). Android-based smartphones were used as hardware devices.

4.1.1 Advanced Voice Framework (AVF)

An integration of automated speech recognition (ASR), text-to-speech (TTS), as well as a speech-to-text (STT) synthesis engine was designed for the information and communication services in CLD. In conclusion, the user was able to fully control the infotainment services through speech, and consequently get system feedback via speech outputs (generated dynamically or pre-processed). Furthermore, the Advanced Voice Framework (AVF) architecture software used facilitated the recording and playing of audio files including the pre-processed prompts (i.e., voice outputs).

Figure 20. Left: HTC Desire with e-mail inbox of the CLD prototype. Right: Apps implemented in CLD prototype

The Advanced Voice Framework (AVF) for Android has Java-Runtime available. The functions can be easily integrated with other programs such as automated speech recognition (ASR), speech synthesis (TTS), audio recording and the playing of audio files (see Figure 21).

The AVF consists of different levels. The clients communicate with the upper level, which implies the android specific services. They provide an interface for other programs within the android system software and provide the resource information of engines on an abstract level.

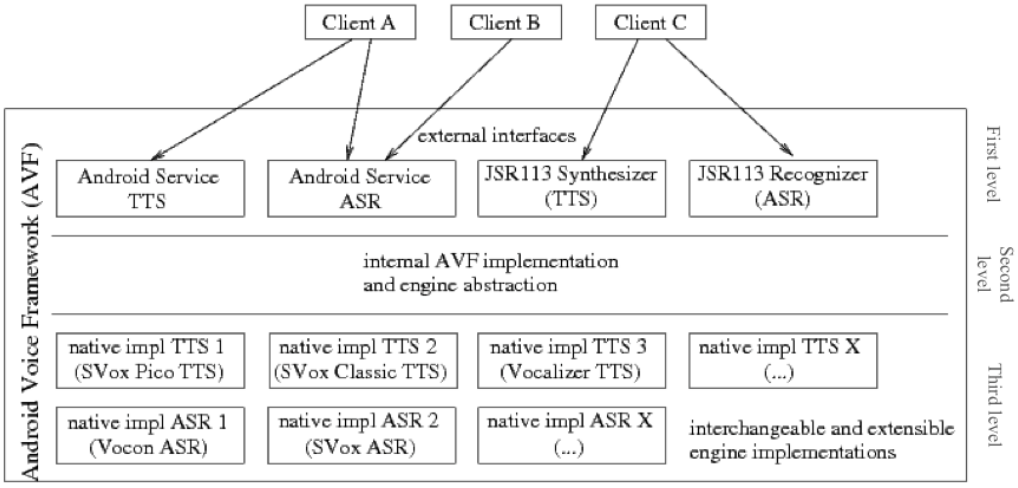


Figure 21. Android Voice Framework (Burkhardt et al., 2010)

As the AVF allows the integration of different external engines on the second level, the Java Speech API 2.0 implementation interprets the information of the engines and delivers a resource-independent code to the android level (first level). The third level consists of the engines mostly implemented in C/C++ and is to be loaded by the Java Native Interface (JNI).

As a standard, SVOX Pico TTS and Google Online-ASR are set by android. For the CLD prototype the ASR Nuance Vocon 3.2 for Automotive as well as the speech synthesis SVOX classic 4.3 Desktop replaced the ASR and TTS. The ASR used in the CLD prototype is a noise-robust, grammar-based, and speaker-independent commercial engine ("suitable for embedded platforms and especially automotive dashboard applications (Burkhardt et al., 2010, p.261) - ASR Nuance Vocon 3.2 for Automotive. The ASR Nuance Vocon 3.2 for Automotive focuses on recognizing short commands but provides also natural language recognition.

The AVF enables integration of speech technology services on Android-based devices. To sum up, the framework enables high-level access to the four function blocks: Automated Speech Recognition, Speech Synthesis, Audio Recording and Audio Playback. Next to voice control and voice output, also touch screen control was integrated.

4.1.2 General Approach for the CLD VUI

Thus, the CLD prototype is a multimodal system. The user is able to interact via the GUI based on a touch screen or via the VUI at any time. The graphical interaction elements are to be operated by soft keys as well as hardware keys. By using a soft or hard key, a key press event is triggered and an action is provoked as well as the execution of audio feedback. The VUI mode is initiated by using the hardware button of the headset, i.e., the so-called 'push-to-activate'. The activation of automated speech recognition was therefore provided through an external hardware button, assembled on the steering wheel (Push-to-Activate Button; Burkhardt et al., 2010). A key press event is detected and the ASR is opened. The CLD ASR grammar consists of valid speech commands and was enriched by the inclusion of semantically same information expressed in different form of words (e.g., valid speech command: "vorlesen" extend with: "abspielen") as well as short cuts. After three seconds of silence, the ASR closes. The ASR also closes if there is one and a half seconds of silence after speech input or after a word is detected but nothing more is traceable in the grammar. Speech recognition in short can be described as "matching the incoming signal with a stored set of patterns to return a sequence of words" (McTear, 2004, p.83).[9] The best match chosen (word or word sequence) is then sent to the dialogue manager, which sets and defines the action to

[9] For further information concerning automatic speech recognition see for example Möller (2005) or McTear (2004).

be performed by the device. Either a dynamic-generated prompt will be initiated or a presynthesized audio file will be played back (both by the TTS SVOX classic 4.3 Desktop - German Petra & Markus Voice). An effort was made to use presynthesized speech prompts as much as possible: Firstly, because they were of higher quality (as they were processed by the tuned TTS), and secondly, because they require less internal memory than generating dynamic prompts. For the presynthesized audio files, the TTS was trained to accurately pronounce previously known words as part of the speech output. Prompts with mixed voices were processed beforehand. Prompts can also be recorded with real voices instead of using speech synthesis. However, it was decided to use speech synthesis only, at least during the development phase, since synthesized prompts are easy to change later. Usually, it is required that professional speakers are paid for recording audio files. As such, it is more time and cost efficient to use TTS for the entire VUI for as long as the prompts are not definite. Additionally, the output is more consistent in terms of quality this way. During the developmental phase an expert listener experiment was conducted. The experts stated that they prefer prompts that "sticks to one voice by using speech synthesis compared to a mix static content by human speakers with dynamic content from speech synthesis." (Burkhardt et al., 2010, p.263).

Fraser (1997) categorizes "computer systems with which humans interact on a turn-by-turn basis" into command systems, menu dialogue systems and spoken dialogue systems (p.567; see also Möller, 2005). The CLD prototype can be referred to as a menu dialogue system since, like menu-oriented systems, the CLD prototype provides different options in every new interaction step. However, menu dialogue systems are usually also characterized by a question-answer interface which is thus system-directed (Möller, 2005). The CLD dialogue is not using question and answers since the driver should not feel forced to answer, for example, by barge-in during highly demanding driving tasks. The initiative to choose the point of time to answer always stays with the user; thus the CLD prototype also used characteristics of command systems. Compared to menu-based systems where the user can choose between options, in spoken dialogue systems, the user is more flexible since natural language understanding (NLU) is possible, which is especially helpful for novices who do not know the valid speech commands (Hockey & Rayner, 2005). However, for system initiative systems, ASR error rate is minimized since less extensive grammar is necessary (Tomoko & Rosenfeld, 2004). Also for the CLD prototype the main scope of the ASR was on recognizing short commands and thus keeping the grammar 'small'. According to Kun et al. (2007), poor recognition negatively influences driving performance. The dialogue in command-based or

menu-based systems is more robust, and therefore is recommended for the automotive context (Hamerich, 2009). Also, Pleschka, Schulz, Ahlers, Weiss and Möller (2009) established that the NLU system is indeed more accepted and liked by users, but in the case of lower recognition rates, this benefit disappears. This is a relevant aspect to be taken into account considering that the noisy car environment in particular results in increased difficulty in word detection. Additionally, Ackermann and Libossek (2006) showed that drivers prefer to be asked for explicit inputs. Similarly, Graham, Aldridge, Carter and Lansdown (1999) observed that users prefer command-based systems over natural language systems. Menus are considered easier to use than systems where a language needs to be learned (Norman, 1986). Hence, in the human-human communication (e.g., work of operational teams) in stressful situations, keywords tend to be used more as opposed to lengthy sentences. An effort is made to keep the dialogue highly efficient by avoiding empty phrases, and by use of more commands and instruction in keyword-based form (Silberstein, & Dietrich, 2003). Based on this literature review, it was decided to use a multimodal menu-based system in the present study. In the following section the infotainment applications of the CLD prototype and their functions will be described.

4.1.3 Infotainment Applications

E-mail. By starting the e-mail app the user accesses the e-mail menu. Here, it is possible to choose between Inbox, News and Search. Via a shortcut (saying “e-mail inbox”), the user can also reach the inbox directly. Shortcuts are available at almost every interaction step and therefore experts do not need to navigate through every interaction step. The newest mail is then presented. Note that it is just one mail per interaction step (not a list of e-mails). The user’s options are to read the mail aloud, or to answer, forward or delete it. With “next” and “previous”, they can navigate through the e-mail list. For this stage of the research “answering” an e-mail was enabled by the function of recording an audio file. Then, three steps were required: input of addresses, recording of audio file, summarization (and confirmation to send the e-mail). In regard to the forwarding option, the user needs to decide to send the mail with or without adding an audio file or message. Since the apps are multimodal, it is also possible to type in a message via a soft keypad. By using the search function, users can search for e-mails from a specific person. Figure 22 shows screen shots for answering an e-mail from the inbox by recording an audio file. In a later version, ‘Speech-to-Text’ was implemented. Here, the user speaks the message and a text is transcribed via Dragon Dictate.

SMS. The SMS app was designed almost identically to the e-mail app. Instead of answering, the option "calling" was provided to directly call the sender. In a later version of the CLD prototype a speech-to-text input was also implemented.

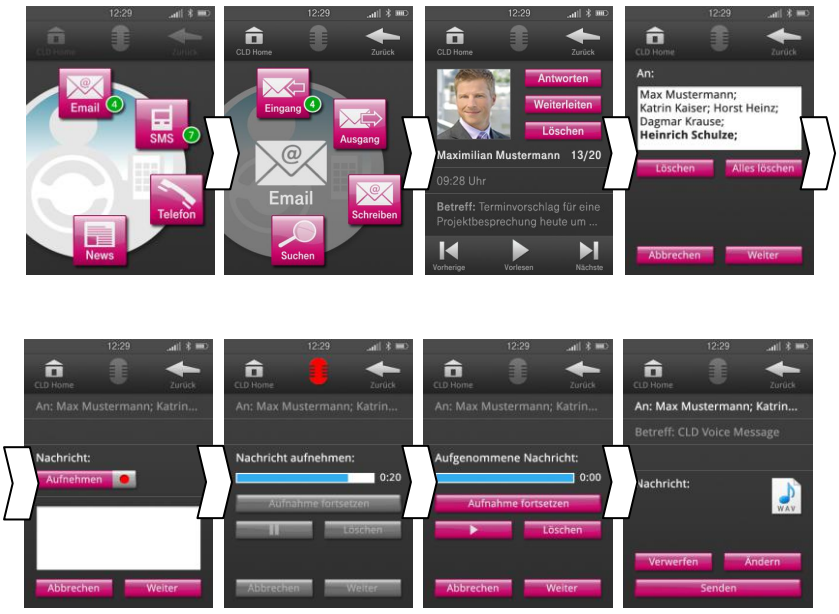


Figure 22. Screen shots for answering an e-mail from the inbox by recording an audio file

Telephone. The telephone app enables the user to call a person from the address book by using speech as the input modality (in addition to the option of typing it).

Infotainment. Infotainment includes news, travel guide and license plate number app.

News. The user is asked to define news categories that he or she is interested in (such as politics, sports, culture, science). According to their choice, various news reports appear as options in the news menu. The news "inbox" resembles the design of the SMS and e-mail inbox functions, and also consists of the same navigation elements as those apps, for purposes of consistency.

Travel Guide. Using the travel guide app, the user can be informed of sights he or she is passing while driving. A notification sound appears. Information of the sight will be read out to the user if he or she opens the app and uses the "read out" command.

License Plate Number App. The user can ask for the location by naming the city code of the license plate number.

4.2 Design Recommendations for Infotainment Applications Based on the SEEV Model

In the first part of this thesis, it was observed that the benefit of visual information for construct controllability can be manipulated via increasing the $effort_{between}$. [10] Additionally, it was shown that the same amount of relevant information content should be presented visually and acoustically. Ideally, the exact same information should be provided, and therefore create maximum conformance (see Chapter 3.3). However, in the first experiment it was also found that the time effort to perceive information for speech output should be taken into account. The trade-off between increasing the relevant information content for speech and increasing the time $effort_{within}$ at the same time needs to be optimized.

In short, design recommendations for (a) increasing controllability (and availability) as well as (b) increasing relevant information content whilst concurrently (c) reducing the time duration of speech prompts were developed and will be described in the following.

4.2.1 Increasing Controllability and Relevant Information Content

> *Sound can be annoying since we cannot close our ears in the same way we can close our eyes.*
> *(Gaver, 1997)*

To increase controllability of speech or acoustical output, an additional function was introduced to the push-to-activate button: while the automatic speech recognition (ASR) was activated with a short-term push, a long-term push initiated the repetition of the speech prompt. In addition to the repetition of the direct speech prompt, more elaborated acoustical information was then also given.

Thus, there were two different output levels:

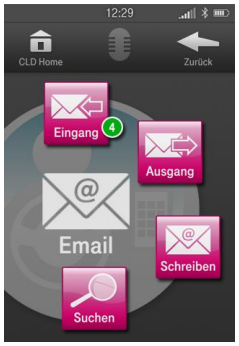
- direct speech prompts for the main interaction path directly presented after an interaction performed (system-initiated speech output)
- long-term push prompts to request all information presented on the screen (driver-initiated speech output)

[10] This chapter reuses text fragments from Niemann et al. (2010a), Niemann et al. (2010b) and Niemann et al. (2010c)

This concept changes the way voice output is generally provided to the driver. Typically, speech output is system-driven, while in the system described, the driver initiated a request for speech output. With the help of the long-term push on the push-to-activate button, the driver can get the desired information whenever he or she intends to (referred to as adaptable information acquisition). Drivers are often interrupted during tasks such as the infotainment task (Baber & Noyse, 2001). Using the driver-initiated speech output, the user is now able to reorientate him- or herself again after being interrupted by the primary driving task. Furthermore, missing a system-initiated direct speech prompt is no longer dreadful, because relevant information can be repeated via the long-term push prompt. This also reduced the cognitive demand of attending to and memorizing the system output in the phonological loop (see Chapter 3.2).

In the present system, the long-term push-to-activate button was initiated after a 500ms push, however this preference can be adjusted by the developer. With this strategy, $effort_{between}$ was observed to increase. In other words, the user needed to perform a conscious motor reaction to request more elaborative speech output next to the direct system-initiated VUI feedback. As the second experiment showed, an increased $effort_{between}$ in this regard did not have negative influences on driving performance, subjectively perceived workload, or glances on the display in low-demand driving situations. It even indicated a decrease in percent dwell time on displays in situations where the secondary task was interrupted and whereby reorientation was necessary. Therefore an increased $effort_{between}$ with an adequate level was recommended. To keep the increased $effort_{between}$ at an adequate level, the use of a button attached to the steering wheel (push-to-activate button) was considered most appropriate. Thus, there is no need for users to perform complex motor actions. Next to the increased controllability provision, the LPA (long-term push to activate = LPA) allows for repeating the speech output and thus increases the availability of speech output (see Chapter 3.3). Note that it has not yet been tested whether an increased frequency of speech prompt presentations leads towards less time glancing at the display. Table 9 shows an example of the speech prompts given as a feedback (direct prompt=DP) and information presented after the long-term push.

Table 9. Examples of speech prompts for the CLD e-mail application.

	VUI	GUI
DP	"Main Menu"	
User	LPA	
LPA	"E-mail, SMS, Telephone, Infotainment"	
User	"E-mail"	
DP	"E-mail"	
User	LPA	
LPA	"Inbox, Outbox, New, Search"	
User	"Inbox"	
DP	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday."	
User	LPA	
LPA	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday. To listen to this mail, say .read out'. To navigate through the list, say ‚previous' or ‚next'. You can also ‚answer', ‚forward', or ‚delete' this mail."	

As can be seen in Table 9, the direct prompt contained menu orientation information to name the actual menu step as well as factual content information such as the e-mail header.

On the LPA-prompt, the menu steps as well as the content information were repeated. Additionally, all possible options were named. As such, almost all information that can be perceived via glancing at the display could be requested by pushing the push-to-activate button for a little longer, at any time the user desires. By providing all information, which is presented on the screen, the second proposed design recommendation (increasing relevant

information content) was attempted to be achieved. Instead of speak-what-you-see, a speak-what-you-hear principle was introduced.

Note that the GUI provides much more indirect information compared to the speech output. Through design provisions related to, for example, colour, font, font size and spatial arrangements amongst others, the development of users' mental models is supported (Hasebrook, 1995). For example, interaction options were marked in magenta colour so they are easier to differentiate from other content information. To represent this meta information for speech, interaction options were announced by using meta information: "you have the following options…" or "for … say…". The difference between the two modalities concerning the amount of relevant information content was endeavoured to be to kept as low as possible (see Table 9). Yet, not quite all of the indirect information content could be presented via speech. Also, note from Table 9 that the home and back buttons were not presented via speech output. This is due to their high probability of occurrence: According to the information theory (see Chapter 2.2.2), a symbol with a high probability of occurrence (i.e., one that's presented nearly all the time) has less information compared to an item with lower probability of occurrence. As the two buttons are presented on every screen (high probability), the information content decreases. However, there will be situations whereby those options are of high relevance (e.g., after the user changes his/her mind and does not want to send an e-mail anymore, but wants to go back to the main menu). Technically speaking, these interaction options must also be read aloud by the system. However, as it was also considered important to keep the time effort low for this special case, it was omitted. Different situations place different demands on information acquisition and therefore different information will be of high relevance.

It is obvious that indeed an increase of relevant information content was achieved. However, again there is a high disadvantage for speech compared to the GUI regarding the factor time effort or $effort_{within}$, which will be focused on next.

4.2.2 Minimizing Time Effort ($Effort_{within}$)

By increasing the relevant information content of in-car infotainment speech output, time effort also increases as a consequence. As the second experiment showed, time effort has a negative influence on attention allocation. In addition to aligning the relevant information content of speech output to the relevant information content of the GUI, keeping the time effort as low as possible and overcoming this trade-off is deemed necessary.

To keep the time effort low for both modalities, firstly it was recommended to reduce information to a minimum and instead offer a deeper menu structure. For example, for the e-mail inbox it was decided to only present one e-mail at one interaction step rather than having a long list of e-mails on one screen, as is familiar on other software (e.g., Microsoft Outlook). Furthermore, the process to record and send an audio file as an answer was divided into three steps (on three different menu levels). According to Rauch, Totzke and Krüger (2004), deeper menu structures are preferred for in-car systems. While broader menu structures in a particular task setting lead to faster performance of the interaction; in the dual task setting the interaction performance for broad and deep menu structures were the same. In fact, a decrease in tracking task performance was also observed for broader menus, as well as longer glance times towards the display (which is congruent with the SEEV Model).

Despite the deep menu structures and therefore a reduction of possible interaction options in every menu step for both modalities (visual and acoustical output), there is still a considerably high benefit for displays regarding the faster detection of information compared to speech outputs (see Chapter 3.1). Complex speech prompts increase mental workload for rehearsing the information in the phonological loop (Salmen, 2002). It has been observed that the amount of items acoustically presented is restricted by the potential of the phonological loop (Balentine & Morgan, 1999; Cohen, Giangola & Balogh, 2004), and it has thus been suggested that there should not be more than four to five items presented at any time.

Next to the presented methods, three design strategies were developed to decrease the time effort ($effort_{within}$) of speech output by continuing to present the same relevant information content:

- Shortening of valid speech commands by using keyword-based speech (enriched by different voices)
- Shortening of land marking by using sounds
- Shortening of content information by speeding up

Shortening of Valid Speech Commands

GUI buttons or interaction elements, which are also valid speech prompts, were magenta coloured. By drawing visual attention to the screen, the user immediately knows which next

interaction options are available and support in differentiation between content information and interaction elements is provided. To present this information acoustically, meta-information is used (e.g., “if you want to answer the mail, say answer”). Meta-information helps the user to differentiate between content information, land marking and system feedback. However, this is very time consuming, and as a solution it was suggested to employ different voices replacing the meta-information. Only valid speech commands would then be read to the user (in the form of a list) by a different voice than all other information. Thus, sentence-based representation of interaction options were replaced by keyword-based speech and enriched by using different voices. By using a different voice for the interaction elements than for other information, the information is implicitly clustered and ‘pops out’. Therefore, it does not require further explicit verbal announcement. This has a positive effect on the time effort because prompts are shortened while the content of information stays the same. Through use of different voices, information was made more salient and differentiable.

Shortening of Land Marking

“Icons can present information in a small amount of space compared to text; nonspeech sounds can present information in a small amount of time as compared to speech.” (Brewster, 2002, p.248)

To avoid disorientation, menu-based systems should give users feedback about their current position in the menu. Chang, Lien, Lathrop and Hees (2009) stated that “users feel more comfortable when they have a mental map of how the system is organized” (p.137). Menu orientations (land marking) help the users to build a conceptual representation of the generic and narrow menu steps and thus not get lost in the menu. If the prior knowledge fits the actual conceptual representation, faster learning can be expected (Totzke, Schmidt & Krüger, 2003). According to Wolf, Koved and Kunzinger (1995), popular contemporary menu-oriented speech-based systems still fail to provide sufficient navigational information. In the CLD project, it was decided to replace speech with land marking information by using sounds, given that they reduce the duration of the acoustical prompt (Brewster, 2002).

There are three main types of nonverbal sounds: auditory icons, earcons and spearcons. Blattner, Sumikawa and Greenberg (1989) invented earcons, which can be defined as abstract, synthetic, short musical motives representing menu items. They differ in the combination of rhythm, pitch, timbre, register and dynamics. Earcons can build up menu hierarchies, which make them especially useful in replacing speech-based navigational information (Brewster,

2002). Blattner et al. (1989) first described a hierarchical structure of earcons. An earcon structured hierarchically under another earcon consists of the same properties as the earcon above and is extended by another motive. Brewster and Cryer (1999) showed that they are also recommendable for in-car navigation systems.

Auditory icons are metaphorical representations of a word or concept (Graver, 1986). In most cases, natural sounds, which possess an intuitive link with the represented item, are used. The most common example is the sound of paper thrown in the garbage by drag and dropping a document in the trash on Macintosh laptops. Important for the use of auditory icons are identifiability and an intuitive link with the intended semantic concept. By meeting these requirements, a high learning rate is given as prior knowledge of the user can be utilized. Garzonis, Jones, Jay and O'Neill (2009) showed that auditory icons compared to earcons in mobile service notification lead to higher intuitiveness, higher learning rate and memorability, as well as higher preference. However, auditory icons are not very suitable for built-up menu hierarchies; and it is usually hard to find natural sounds for all menu steps that are highly semantically linked with the represented item. If auditory icons are not intuitively linked, the learning rate is lower compared to earcons as a consequence (Bonebright & Nees, 2007).

Spearcons were originally invented by Walker, Nance and Lindsay (2006), and are essentially speech audio cues that are sped up to the point that speech is no longer comprehensible. As the sound of the accelerated word is unique as a fingerprint, it is very easy to produce sounds that are different enough from one another. Jeon, Davison, Wilson, Nees and Walker (2009) showed that spearcons are useful in the in-car context for navigating through long song lists (e.g., 150 songs). Text-to-Speech output was enriched by using spearcons. Participants performed better in the primary tasks and navigated more effectively through the song list task when using sounds and TTS instead of only TTS. Spearcons could be used as well as earcons for orientation in the menu and also need to be familiar before being presented on their own. However, Palladino and Walker (2007) compared the two methods in terms of ease of user learning, and found spearcons to be easier to learn than earcons, as less training cycles are necessary. For an overview, see Table 10, which compares characteristics of the three sound types in terms of learnability, effort to develop the sounds, and capability to build up menu hierarchies and aesthetics.

Auditory icons are easy to learn, due to their intuitive link with their target word or concept. Earcons and spearcons need to be learned but are more easily developed. Further, they can build up menu hierarchies. An important aspect of non-speech sounds is the consideration of annoyance and user preference (Brewster, 2002), especially considering that in-car infotainment systems are comfort systems, which add various requirements to the design in terms of acceptance and joy of use. Auditory icons have been proven to be more readily accepted than earcons (Duarte & Carrico, 2008). In review of the literature, no experiment thus far has actually compared earcons with spearcons. However, in the developmental phase of the present research, it emerged in an expert analysis that spearcons were not popular due to their aesthetic appearance. As such, it was decided to use earcons to provide orientation within the menu, as auditory icons do not build up menu hierarchies and for concepts such as e-mail and SMS it is hard to find a natural sound with an intuitive link. Later in the acoustic design phase, auditory icons were added in the menu hierarchy and compared to earcons on their influence on users' mental model (see Chapter 4.3) As aesthetics and learnability were important factors for the interface design, it was decided to use auditory icons only whenever (a) there were no earcons used and (b) there was a sound found with a high association between the item and its target concept. To sum up, to shorten the land marking information earcons were used instead of speech.

Table 10. Evaluation of auditory icons, earcons and spearcons on the concepts of learnability, producibility, possibility of building a menu hierarchy and aesthetics.

	Auditory Icons	**Earcons**	**Spearcons**
Learnability	+	-	0
Producibility	-	0	+
Menu hierarchies	-	+	0
Aesthetics	+	0	-

A sound designer was asked to develop earcons for the CLD menu based on the guidelines (see Table 11) invented by Blattner et al. (1989) and developed further by Brewster and Cryer (1999).

Table 11. Guidelines for designing earcons by Blattern et al. (1989) and Brewster (1999) as cited in Brewster (1999)

GUIDELINES by Blattern et al (1989) and Brewster (1999) as cited in Brewster (1999)
Timbre: Use musical instrument timbres. Where possible use timbres with multiple harmonics. *Pitch*: Do not use pitch on its own unless there are very big differences between those used. Complex intra-earcon pitch structures are effective in differentiating earcons if used along with rhythm. Ranges for pitch are: Max.: 5kHz (four octaves above middle C) and Min.: 125Hz - 150Hz (an octave below middle C). *Register*: If this alone is to be used to differentiate earcons which are otherwise the same, then large differences should be used. Two or three octaves difference give good rates of recognition. *Rhythm*: Make them as different as possible. Putting different numbers of notes in each rhythm is effective. Small note lengths might not be noticed so do not use notes less than eighth notes or quavers. *Intensity*: Some suggested ranges (from Patterson, 1982) are: Max.: 20dB above threshold and Min.: 10dB above threshold. Care must be taken in the use of intensity. The overall sound level will be under the control of the user of the system. Earcons should all be kept within a close range so that if the user changes the volume of the system no sound will be lost. *Combinations*: When playing earcons one after another use a gap between them so that users can tell where one finishes and the other starts. A delay of 0.1 seconds is adequate.

Figure 23 shows the abstract menu hierarchy of the CLD app (note that only for SMS and e-mail a detailed menu structure is presented).

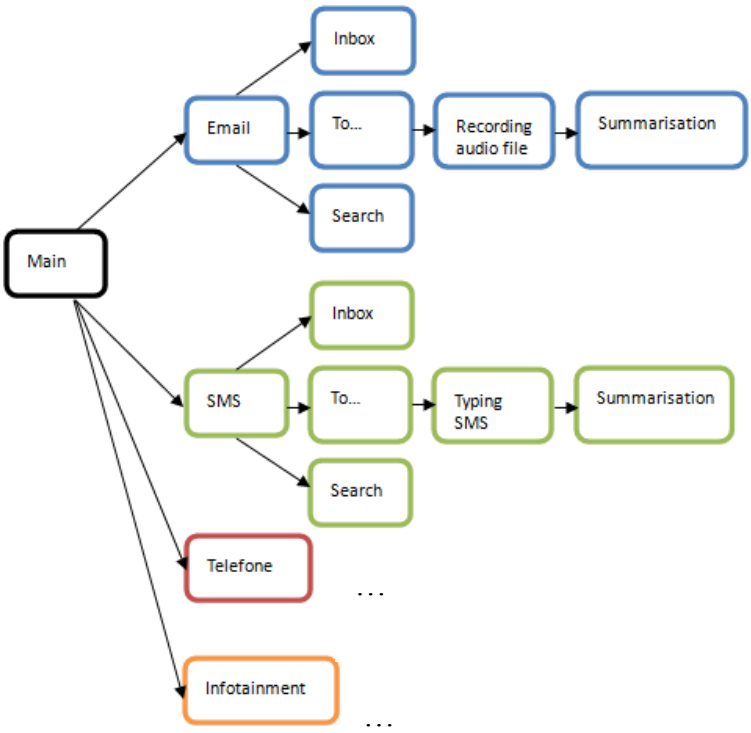


Figure 23. Abstract menu hierarchy of the CLD app

Every menu item on the start screen (e-mail, SMS, telephone and infotainment) has its own sound motive (timbre). Figure 24 shows the sound waves of e-mail menu and inbox as well as SMS menu and inbox. The sound motives of the four apps (e-mail, SMS, telephone and infotainment) exhibit high differentiation – e.g., compare sound waves of e-mail menu part (a) with SMS menu part (a). To represent one lower level in the menu hierarchy of every app, the sound motive of the specific app (a) was extended by another sound (see part (b) in Figure 24). For the particular steps of the main interaction path (e.g., writing an e-mail) within the specific menu items, the menu item sounds of part (b) were varied by changing the pitch. The pitch then rises with every step nearer to the end of the operation. The higher pitched sound is added to the sound of the previous interaction step to allow for differentiation, since rises in pitch without reference are difficult to detect (see Table 11). Thus, this needs to be seen relatively.

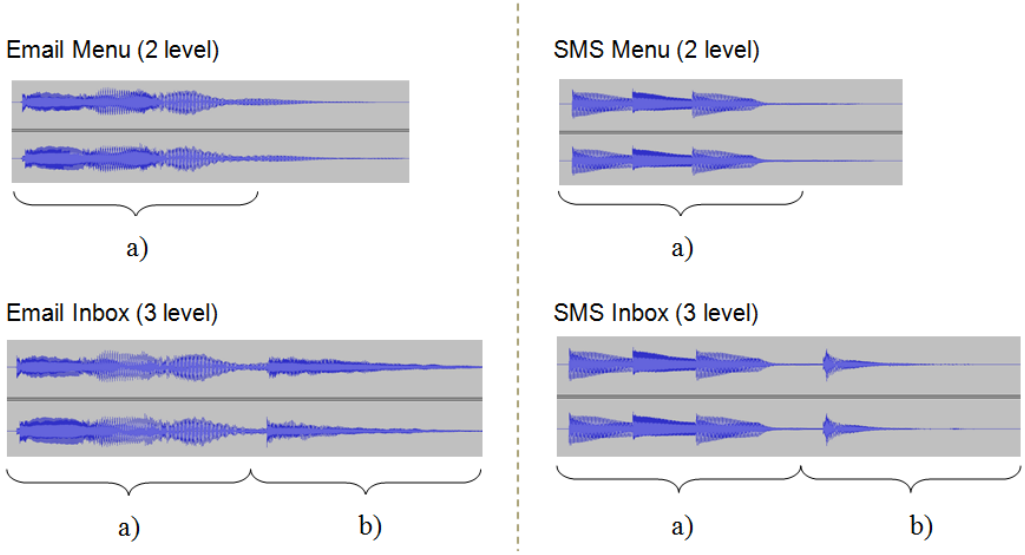


Figure 24. Sound waves of the earcons used for the menu steps e-mail menu and e-mail inbox (left) as well as SMS menu and SMS inbox (right)

Shortening of Content Information

Next to land marking information and interaction options, content must also be presented acoustically. For example, in the e-mail application the e-mail header had to be read to the user as well as names or addresses that he or she has entered. By pressing the LPA the user received a chance to re-read this information that had already been heard. To shorten this, it was suggested to speed up the output. The speech synthesis allows a manual setting of the playback velocity. It was set at 30% (value determined with a pre-test of acceptance based on five experts).

4.2.3 Summary

Based on the SEEV Model by Wickens et al. (2003), it was deducted that displays compared to speech have higher benefits regarding:

a) controllability and availability for allocating attention towards the output and

b) the relation of subjective relevant information content and the time effort to perceive this information.

Offering the LPA increased controllability, the previously described function attached to the steering wheel with which the user is able to request information at any time he or she desires. To overcome this trade-off between relevant information content and time effort, it was suggested to provide the same relevant information content on the GUI as on the VUI and at the same time, to try and reduce the duration of the acoustical prompt. To reduce the duration of speech, three different strategies where integrated into the CLD app:

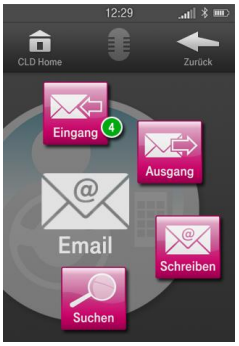
a) Using keyword-based speech that is semantically enriched by using different voices to present interaction options

b) Using sounds (earcons) instead of speech to present landmark information

c) Using uptempo speech to present previously-heard content information

Table 12 shows the new acoustical output. On average (calculated for all prompts), a decrease of nine seconds in the speech prompt time compared to the sentence structure without shortening was achieved.

Table 12. Examples of speech prompts for the e-mail application of the CLD prototype after revision by using the design recommendations described in Chapter 4.2 (red: female voice, blue: male voice).

	VUI	GUI
DP	"Main Menu"	
User	LPA	
LPA	"[Sound Main Menu]. E-mail, SMS, Telephone, Infotainment"	
User	"E-mail"	
DP	"E-mail"	
User	LPA	
LPA	"[Sound Main Menu]. Inbox, Outbox, New, Search"	
User	"Inbox"	
DP	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday."	
User	LPA	
LPA	"[Sound E-Mail Inbox]. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday. Read out, Previous, Next, Answer, Forward, Delete.	

4.2.4 Experiment 3: Evaluation of the Influences of Deducted Design Recommendations on Attention Allocation

A driving simulator study has been conducted to test if the design recommendations to improve the speech output of an in-car infotainment system do indeed reduce eye gazes to the

display because the probability of attention allocation to speech is increased.[11] Therefore, PDT Display was again collected as an indirect measurement for the $P(A_{Speech})$ (see Chapter 2.3). As mentioned before, PDT Display can be reduced by positively influencing the parameters of the speech output (increasing $P(A_{Speech})$) or by negatively influencing the parameters of the visual output (decreasing $P(A_{Display})$). The characteristics of the speech output were varied while the visual output of the different system versions as well as the primary task of driving stayed the same in all experimental conditions.

It was evaluated if enhancing the parameter relevant information content and decreasing $effort_{within}$ in line with design recommendations introduced in Chapter 4.2 would lower attention allocation and therefore percent dwell time towards the display. Three different system versions were implemented. For one of the systems, an LPA function was implemented as previously described. The user can request all information at any time he/she wishes (which results in increased relevant information content and controllability). At the same time, operationalizing the design recommendations reduced the time effort. This system version was compared with a system having LPA function and the same relevant information content but without reducing the time effort. Furthermore, it was compared with a third system version with low controllability and low relevant information content but a low time effort.

Compared to the second experiment (Chapter 3.2.1), a driving task was conducted with which interruption of the secondary task was not initiated, and thus no positive effect for increasing the controllability was postulated. As indicated in the second experiment, the benefit of increasing the controllability will only be observable in highly demanding primary tasks (which was not the case in this experiment). The main focus of this experiment was the effect of overcoming the trade-off between increasing the relevant information content for speech, while at the same time, reducing the time effort.

In Table 13, the three different infotainment systems and the relative characteristics for the SEEV Model parameters as well as the design recommendations are shown. Also the expected $P(A_{Speech})$ is shown. Relative values (high, low) in Table 13 were set in comparison to each

[11] This chapter is a revised and and extended version of Niemann et al. (2010a).

other. Plus or zero written in the parentheses of Table 13 define if this value was assumed to lead to an increase or decrease of attention allocation towards $P(A_{Speech})$.

Table 13. System versions of the parameters of the revised SEEV Model and predicted attention allocation towards the display (without taking $effort_{between}$ into account).

Parameters	Design Recommendations	System 1	System 2	System 3
Relevant information content	same amount of information on VUI as on GUI	Low (0)	High (+)	High (+)
$Effort_{within}$	- non-speech sounds - uptempo speech - commando based speech	Low (+)	High (0)	Low (+)
Expected $P(A_{Speech})$		(+)	(+)	(++)
Expected PDT Display		High	High	Low

Based on the theoretical deductions and results from experiment 1 (Chapter 3.1.1), the following hypotheses were formulated.

4.2.5 Hypotheses

H4: Increasing the relevant information content while at the same time reducing the duration of speech prompts will lead to less percentage dwell time to the display compared to the system versions with decreased relevant information content or an increase of time effort.

H5: The LPA will be used more often if the $effort_{within}$ for the speech output is reduced (compared to not reducing the $effort_{within}$).

4.2.6 Method

In the following chapter, the experimental design, materials used, as well as participants and the test procedure will be described. Figure 25 illustrates the experimental setup.

Experimental Design

As the independent variables of a between-subject design, the variables of the modified SEEV Model were manipulated in three versions of the infotainment system.

The design of the experiment was a 2x3 design. The independent variable infotainment system (systems 1, 2 and 3) was a between factor, whereas factor task (trained versus untrained) was tested across repeated measurement occasions.

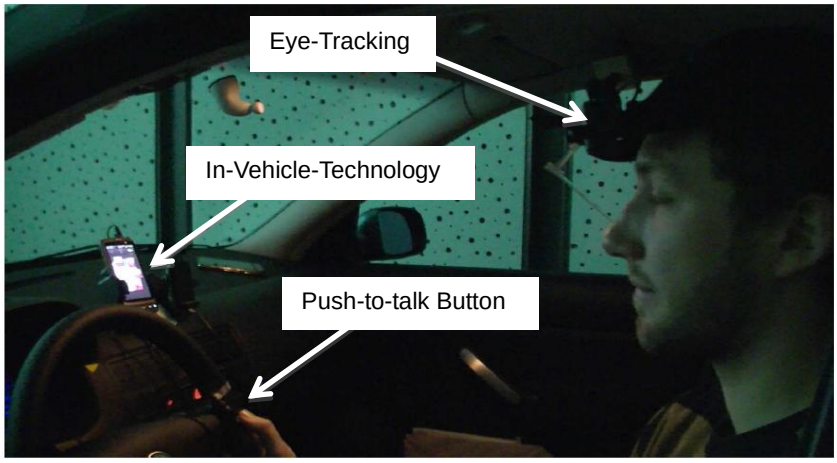


Figure 25. Experimental setup in the third experiment

Independent Variable 'System'. System 1 meets the characteristics of a commonly used 'speak-what-you-see' in-car speech dialogue system. The system gives feedback based on the user's input. The relevant information content and the $effort_{within}$ are all of a low value. As a result of providing all the information given on the screen acoustically by a long-term push on the push-to-activate button (system 2), the relevant information content was increased. In system 3, the LPA (and possibility to request for all information acoustically that were given on the screen) was again implemented. However, in contrast to system 2, the time $effort_{within}$ was reduced by using the design recommendations described in Chapter 4.2.2. On average, a decrease of nine seconds for the speech prompt time was achieved.

Table 14 shows the acoustical output for the e-mail inbox of the three system versions respectively in comparison. The versions only differed in terms of the acoustical output given after pressing the long-term push-to-activate button.

Table 14. Comparison of acoustical output for the e-mail inbox of the three system versions (red = female voice, blue = male voice, differentiation by colour was only made for system 3, since for the other systems version only the female voice was used)

	System 1	**System 2**	**System 3**
DP	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday."	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday."	"E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday."
User	*(no LPA)*	LPA	LPA
LPA	---	"E-mail E-mail Inbox. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday. For listening to this mail, say 'read out'. To navigate in the list, say 'previous' or 'next'. You can also 'answer', 'forward', or 'delete' this mail."	"[Sound E-Mail Inbox]. 4 new mails. [break] E-mail 1 of 13, read. Today, 8:30. Max Mustermann. Meeting for cinema on Friday. Read out, Previous, Next, Answer, Forward, Delete.

Independent Variable 'Task'. Next to the system versions, the influence of the task novelty was tested. An untrained as well as a trained task was performed in the experiment. During the untrained task condition, unfamiliar speech commands (i.e., not already known from the trained interacting task) had to be used, unlike in the trained task condition. Therefore, the relevance of that information (speech command) is increased. The effects of increasing the relevant information content by providing the same number of information items acoustically as is presented visually, compared to the speak-what-you-see principle, is expected to be especially pronounced for the untrained task. However, there were no specific hypotheses formulated (but tested) for the independent variable 'training' (condition 1: untrained, condition 2: trained), as this question was not of central relevance to the current research.

Test Subjects

Subjects were 58 native German speakers, in order to ensure an accurate interaction with the German-based VUI. Another requirement for participation was to hold a current driver's license. Subjects were asked not to wear spectacle glasses (however contact lenses were permitted).

On account of the following issues, 16 subjects' data were excluded from the analysis:

- problems of calibration accuracy (1)
- nausea (simulator sickness) (1)
- low quality of data (high blinking rates, eye tracking coverage below 75%, or technical problems) (14)

The 42 remaining subjects were within an age range of 20 to 50 years (M = 29.64, SD = 6.74; gender ratio: 25:17, M:F). On average, subjects had held their driver's licenses for 11 years. All subjects received 20 Euros remuneration for the two hours spent participating in the experiment.

Materials

Driving Task. Again, the experiment was conducted using the driving simulator at the Centre of Human Machine Systems at the TU Berlin (see Chapter 3.1.4 for further information). The subjects were instructed to drive at 120 km/h on the motorway until they reached a particular car. They were asked to then follow this car at a distance of approximately 50 meters and to keep the distance as constant as possible, while avoiding lane deviations. The velocity of the leading vehicle varied throughout the task (same driving task as in experiment 1).

Secondary Task. For the secondary task, three different system versions of the CLD applications e-mail, SMS, news, travel guide, and telephone were used. After allocating participants randomly to the different conditions of the independent variable, the corresponding system version was installed on a HTC Desire Smartphone. The Smartphone was fixed at the dashboard (at the same location as in the first experiment). The interaction data was logged by using the android standard api. The Smartphone was connected to a Windows laptop to request for the incoming data. The data was saved in a text logfile as well as the time stamp. By using perl[12], the required information was extracted and transformed into .csv data files.

Eye Tracking. Again (as in experiments 1 and 2), the fully mobile, headmounted system iView X HED from SensoMotoric Instruments GmBH was used to record eye gazes. Eye

[12] http://www.perl.org/

tracking was recorded with a sampling rate of 50Hz. For the calibration, a five-point calibration chart was used. The eye tracking system was synchronized with the recorded interaction data by a serial port between the Smartphone connected laptop and eye tracking laptop.

Questionnaires. For measuring mental workload, the NASA-TLX was used (Hart & Staveland, 1988). After the test drive, participants were asked to indicate the level of subjective workload they had experienced. As this was a standardized questionnaire, participants were able to give a rating based on the following six dimensions in terms of both the driving task and the secondary task they had worked on. The rating was indicated on a scale with 21 points (from 0 to 20).

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration

According to instructions, these dimensions should be supplemented by 15 paired comparisons and weighted accordingly. According to Nygren (1991), the weighted total value correlates highly with the non-weighted mean value, and for this reason, paired comparisons were not used. The six dimensions were evaluated separately.

The standardized questionnaire on product acceptance (Davis, 1989) was used for the acquisition of the system acceptability[13]. This questionnaire contained eleven questions, distinguishable into three subscales: usefulness, usability and acceptance. Thus, the questionnaires' data were added up for each dimension across participants.

[13] User acceptance/acceptability is a dimension of Quality of Experience (QoE). Quality of Experience: "The overall acceptability of an application or service, as perceived subjectively by the end user." (ITU-T Rec., 2007, p.10)

Dependent Variables

Table 15 shows the dependent variables collected in the experiment.

Table 15. Dependent variables in the third experiment.

Data category	Measured Variable/Construct	Measuring Tool
Objective Data	Percent dwell time on Display	Eye Tracking
	Lane Deviation	Driving Simulator
	Standard deviations distance to lead car	Driving Simulator
	Task duration	Android.log Smartphone
	Task completion	Android.log Smartphone
	Wrong Speech commands	Android.log Smartphone
Subjective Data (Questionnaires)	Mental Workload	NASA –TLX
	Acceptability	Product acceptance

Procedure

After completing a demographic questionnaire, subjects were trained to use the systems outside the driving simulator in order to confirm that they understood the push-to-activate principle (whereby the user has to push the button before speaking). They performed an e-mail task whereby they were asked to initiate for the third mail in the inbox to be read out, to answer this e-mail by recording a voice message, and finally, to send it. Subsequently, subjects were trained in the driving task of the driving simulator without interacting with the infotainment system. After the training, participants conducted another training phase. They performed the trained e-mail task and the driving task at the same time. The eye tracking system was adjusted and calibrated.

The test phase consisted of two trials. First, participants conducted the trained e-mail task and the driving task described above one final time. Afterwards, they filled out the NASA-TLX questionnaire. Next, a new task with the infotainment system had to be performed. This involved searching for an e-mail using the search function and then forwarding the e-mail without having to record a new message (an untrained task). Speech commands that were unfamiliar had to be used. Again, subjects were requested to complete the NASA-TLX. The eye tracking system was then removed. Finally, participants were invited to complete the product acceptance questionnaire.

4.2.7 Results

H4: Increasing the relevant information content while at the same time reducing the duration of speech prompts will lead to less percentage dwell time to the display compared to the system versions with decreased relevant information content or an increase of time effort.

A two-way mixed methods ANOVA was completed. A significant effect of task ($F(1,39) = 77.01$, $p < 0.001$, $\eta^2=0.664$) and of system ($F(2,39) = 6.03$, $p = 0.005$, $\eta^2= 0.236$) for the percent dwell time on display was observed. Also the interaction effect ($F(2,38) = 6.04$, $p = 0.004$, $\eta^2= 0.247$) was significant.

Table 16. Means and standard deviations of PDT Display for the three system versions for the conditions untrained and trained task.

System	Task	Mean	Standard Deviation
System 1 (low effortwithin, low relevant information content)	Trained	2.29	2.10
	Untrained	15.30	5.72
System 2 (high effortwithin, high relevant information content)	Trained	3.56	2.87
	Untrained	15.42	9.15
System 3 (low effort within, high relevant information content)	Trained	2.13	2.48
	Untrained	6.13	7.29

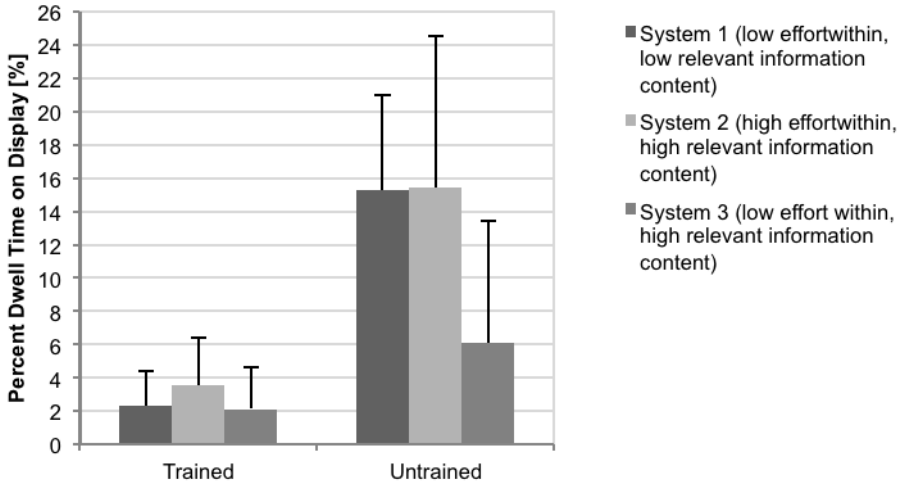


Figure 26. Means and standard deviations of PDT Display for the three system versions for the conditions untrained and trained task

After conducting post-hoc Bonferroni comparisons, significant differences between systems 1 and 3 (p = 0.03) and between systems 2 and 3 (p = 0.01), but not between systems 1 and 2 were observed.

This is in line with hypotheses 4, and with the assumption that the difference between the three systems will be more prominent for an untrained task. For means and standard deviations see Table 16 and Figure 26.

H5: The LPA will be used more often if the time effort is reduced (compared to not reducing the time effort).

A significant effect for the variable task was found for the usage of the LPA (F(1,25) = 30.77, p < 0.001, η^2= 0.552). The interaction effect was not significant, nor was the main effect for 'system'. However, as shown in Figure 27, a tendency in the expected direction could be observed for the untrained ask: Participants of system 2 pressed the LPA on average three times (M=3.0, SD=3.80) while participants of system 3 used the LPA at least five times on average (M=5.08, SD=3.12).

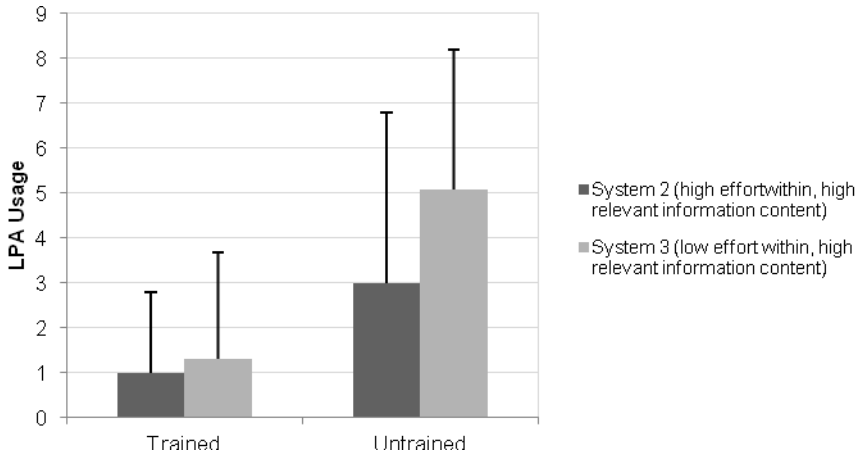


Figure 27. Means and standard deviations for amount of LPA usage for two system versions and for the untrained and trained task

Acceptability

Significant differences between the three system versions were not observed for the standardized questionnaire (all p's > .05).

Driving Performance

For the operationalization of driving performance, lane deviation was measured. After performing a two-way mixed methods ANOVA, no significant effects were observed (all *p*'s > .05). Due to loss of data, the standard deviations of the distance to the lead car could not be analysed.

Subjective Workload

Again, a significant effect for the main effect task was observed ($F(1,39)=13.24$, $p=0.001$, $\eta^2=0.253$). The interaction effect and main effect system was not significant (see Table 17 for means and standard deviation).

Table 17. Means and standard deviations for subjective workload (NASA TLX) for the three system versions and untrained/trained task

System	Task	Mean	Standard Deviation
System 1	Trained	32.79	19.00
	Untrained	40.86	27.82
System 2	Trained	34.80	24.44
	Untrained	43.60	29.92
System 3	Trained	29.69	16.67
	Untrained	42.92	17.93

Task Duration

For task duration, a significant main effect for trained/untrained was observed; $F(1,34) = 14.971$, $p<0.001$. As expected, the trained task was solved faster than the untrained task. The main effect system did not reach significance, however, a tendency was observed; $F(1,34)=3.16$, $p=0.055$. The difference between systems 1 and 3 became not significant (system 1 vs. 3, $p=0.054$). For means, see Table 18.

Table 18. Means and standard deviations for time (in minutes) for solving the infotainment tasks for the three system versions and for the untrained and trained task.

System	Task	Mean	Standard deviation
System 1	Trained	1.57	0.42
	Untrained	1.70	0.55
System 2	Trained	1.69	0.58
	Untrained	2.19	0.68
System 3	Trained	1.84	0.52
	Untrained	2.44	0.71

4.2.8 Discussion

A driving simulator study was conducted to test if overcoming the trade-off between relevant information content and $effort_{within}$ by the presented design recommendations would indeed result in less time spent glancing at the display. Three different system versions were implemented. It was observed that only confirming the acoustical relevant information content with the visual relevant information content would not reduce tendency to glance at the display. This is due to the increased time effort of the speech output caused by the higher amount of information items compared to a system with reduced output information given on the VUI. However, adjusting the relevant acoustical information content to visual information content by also considering the time effort will lead to a reduced PDT Display. In other words, the CLD system with the increased relevant information content on the VUI while at the same time reducing duration of the speech prompts by using the shortening strategies actually decreased the tendency to glance towards the display. On average, the PDT Display could be reduced to 6% when applying the design recommendations (compared to 15%). This was shown for an ‘untrained’ task. Does this consequentially mean that this result would only be found for unfamiliar tasks or for novices? To answer this question, it is noteworthy to consider that even the untrained task was very familiar to all participants, given that e-mail functions and contents (such as ‘inbox’ or ‘forward’) are commonplace. As such, it is unlikely that the results are only relevant in the case of completely new tasks. It is only evident that this effect will disappear when performing a very familiar and very well trained task as it was the case for the email task in the present experiment.

Although the PDT Display could be significantly reduced, no increase in driving performance was observed. However, because driving performance was solely operationalized in terms of lane deviation (due to data loss of measured distance to the lead car), which may not be a

particularly sensitive measurement on a relatively straight highway ride, caution must be executed in interpreting this finding. Participants performed the driving task on a straight highway. Inattention during such a driving situation is difficult to detect. Lane deviations are less likely to occur than on curvy roads. Significant deviations may only be observable for extremely high percentages of dwell time on the display. As mentioned before by using ambient peripheral visual field information, lowly cognitively demanding tasks such as lane keeping can still be well performed (Previc, 1998, as cited in Wickens & McCarley, 2007). Another argument in regard to the non-perceived driving distraction and the lack of differences in driving performance might be that it is not particularly distracting to glance at the display from time to time. Indeed, Kun et al. (2009) could not find a difference in the overall driving data (i.e., lane position, steering wheel angle and velocity) for a decreased PDT Street. However, they did find that in the very moment of eye gazes towards the display, the lane position and steering wheel variances increased. Additionally, Horrey and Wickens (2006) showed in their meta-analysis that use of cellphones does not have a significant influence on driving performance, but in the case of hazards, reduced reaction times were observed. This implies that a more detailed analysis of the data is necessary to properly investigate negative effects in the driving data caused by increased attention allocation towards the display.

No significant differences were found for subjective data concerning acceptability. Further, no differences regarding the perceived mental workload were detected. This lack of an effect could be explained by considering the interaction data (task duration and task completion). No significant differences for task completion and task duration were detected. However, the means showed that systems 1 and 2 were even faster than system 3 (i.e., in terms of faster task competition). Note that participants of systems 1 and 2 used the display more often to get information access compared to system 3 participants. The LPA was used more often for a system with shortened speech output. This was in line with the assumption and with the observation that system 2 participants spent more time glancing at the display. These individuals preferred to glance at the display and to perceive the information visually than to listen to the speech output for a long duration. In other words, using the display is in fact still more efficient (also for shortened speech prompts) and the task can still be solved, however, it is less visually distracting to use the LPA. These effects – both benefits and disadvantages – may have equalized balance for subjective workload and user acceptance in system 3. A further problem could be that the questionnaire used was an inadequate measurement tool. Wechsung and Naumann (2008) recommend for evaluating quality of experience of speech

dialogue system the use of questionnaires adapted to speech recognition systems (such as SASSI, Hone & Graham, 2000 see Chapter 5.1.3) compared to more general questionnaires (such as AttrakDiff, Hassenzahl, 2003, or the product acceptance questionnaire, Davis, 1989). Lastly, examining the ways in which introducing the LPA itself had an influence on the PDT Display was not a primary focus of this study, as it was examined in the second experiment.

In conclusion, the results indicate the recommendation of adjusting speech output in line with relevant information content of visual output, while at the same time reducing the duration of the speech output. In other words: best practice is to present the same information acoustically as is presented visually, and use shortening strategies to overcome the disadvantage of speech that result in slower information acquisition.

4.3 Influence of Shortening Speech Prompts on Users' Mental Model

In the last-described driving simulator study (experiment 3), it was found that reducing the time effort of speech prompts could actually reduce the percentage dwell time on display. These were shortened by using earcons, commando-based (i.e., keyword-based) speech and uptempo speech.[14] The present chapter deals with the problems and benefits of shortening speech prompts. The most apparent benefit is reduced time effort. It is however not yet clear as to whether shortened prompts allow the driver to build an equally good mental model as do extended speech prompts.

4.3.1 Experiment 4: Effects on Users' Mental Models

On average, a decrease of nine seconds in total speech prompt time was achieved employing all three strategies simultaneously which, as mentioned before, lead to lower percent dwell time on display. But since "speech is the most semantically rich acoustic medium" (Garzonis et al., 2009, p.1514), shortening and therefore exclusion of speech could result in information loss. Speech has the characteristics of high identifiability and high symbolic character, which makes it difficult to replace speech with sounds.

[14] This chapter is a slightly modified version of (Niemann et al., 2011b).

The present experiment investigates the extent to which (a) users' knowledge and (b) learnability of the system are decreased through shortening. The users' knowledge of a system is represented in his/her mental model. Mental models contain humans' structural analogies of the world (Johnson-Laird, 1983). They are, for example, representations of complex problems such as technical systems. By use of an internal mental representation, an analogy can be built from the original object, and problems can be solved by using the mental model. In the context of Human Computer Interaction (HCI), the mental model can be considered as the conceptual representation of a device or application. It is identified by a strongly shortened view and a reduction of complexity at a few system elements (and their relations among one other). The user's mental model of a system is developing during the interaction process and will be confirmed and specified along the interaction.

An important characteristic of mental models is that, due to restrictions of human cognitive resources, mental models are often left incomplete (Norman, 1983). Thus, they are simpler and not as detailed as the actual situation or problem they are presenting (Johnson-Laird, 1983). A point of pertinence is also that the device developer should bear in mind what mental model he/she wants the user to build based on the system. According to Norman (1983), the user's mental model will be built upon the system image that communicates the designers' model to the user. The system image contains the user interface. For the designed CLD VUI, providing a mental model resulting in learning about what to say and facilitating understanding of the menu structure (land marking) was hoped to be achieved. Totzke et al. (2003) differentiated between a conceptual, spatial, and motoric representation within the user's mental model. Conceptual representation defines the understanding of which content and functions are provided in the system. In addition to information on where to find various content and functions within the menu, other functions such as the ability to answer to an e-mail will be possible in the inbox whilst the e-mail is presented. Spatial representation describes the location of menu steps within the actual menu, and is comparable with land-marking information. Spatial representations are independent from visual perception (Totzke et al., 2003). The term originates from the mode of visual presentation for hierarchical information (see dual coding, Paivio, 1971). 'Down' means a menu step lower within the menu structure (hierarchy). Motoric representations are not relevant for speech dialogue systems since they describe the motoric or action coding of executing an input. Since the CLD system is operable by speech, it is considered irrelevant in the presented context. However, decreasing the conceptual and spatial representation was avoided by use of shortening strategies. The system, as previously described, is a commando-based, hierarchical, menu-

orientated system including deep menu structures. Therefore, menu orientation (spatial representation) and learning of the speech commands (conceptual representation) are essential components. It was assumed that the shortening of interaction options by use of commando-based speech would not affect learning of the conceptual representation. The shortening of the interaction option via commando-based speech is in fact expected to be more efficient, since all relevant information is still presented, while irrelevant information (e.g., the meta-information) is replaced by different voices. Complex sentences increase the workload for the phonological loop (Chapter 3.2).

For the use of earcons, an improvement (as opposed to no shortening) is not necessarily expected. In comparison to speech, an earcon's semantic link is significantly reduced. However, again it is assumed that the relevant information achieved in creating the mental model is still present in earcons and their hierarchical nature. "A consistent, easy perceivable and clear feedback design can be created through an accurate construction of a semantic system with even non-linguistic sounds" (Vilimek, 2007, p.47). Brewster (1998) found in an experiment with a system hierarchy of "27 nodes and four levels", location recognition of "81.5% accuracy, indicating that earcons were a powerful method of communicating hierarchy information" (p.224). Another strategy to label menu steps with nonverbal sound is with the use of auditory icons. According to Guski (1997), suitability, level of smoothness and identifiability must be considered when designing auditory stimuli. As described in Chapter 4.2.2, auditory icons are very intuitive and therefore provide high identifiability. However, for land marking information, earcons seems to be more adequate since they transfer hierarchical information. Brewster (1998) showed that they are more appropriate for use with land marking than with auditory icons. This is one question that was also aimed to answer: Do earcons provide a better building of spatial representation in the users' mental model over auditory icons? To answer these questions, it was attempted to replace speech output with auditory icons whenever there was a strong and clear associated sound available. For example, instead of the speech output 'Sport' in the News app, a clapping sound from a large soccer stadium was used. Similarly, an auditory icon for cultural news, namely, the sound of an orchestra was found. Earcons were used for e-mail and SMS, while auditory icons were introduced for all infotainment applications and the telephone application. Thus, it was possible to replace all land marking information from the infotainment apps with auditory icons. As the earcons included hierarchical information, it was expected that the positive effects of a strong semantic link (auditory icons) would be adjusted and result in an increase

performance rate for the analogue scale task and naming the menu steps before and after task. For the comparison of speech and shortening via sounds as well as for replacing speech by uptempo speech (30% faster) no pre-assumptions were made.

To sum up, it was predicted that, in comparison to regular speech, commando-based speech would not negatively influence the building of an adequate user mental model. It was further predicted to increase learnability. For earcons, it was not assumed that they would lead to improvements over speech, however, it was expected that hierarchically structured information (like earcons) lead to improved building of a spatial representation than non-hierarchically information (like auditory icons). For uptempo speech, no assumptions were made. To explore these assumptions, it was necessary to measure the abstract mental model of the users.

4.3.2 Measuring Users' Mental Models

[…] The possibility of totally 'capturing' the mental model is rather remote.
(Rouse and Morris, 1986)

One method to measure users' mental models is the thinking aloud technique. Subjects are asked to speak their thoughts aloud while interacting with a system or solving a problem. The goal of using this strategy is to receive qualitative insight regarding cognitive processes, and in the special case of Human-Machine Interactions, to learn about system understanding (Sasse, 1992). However, it has been found to be difficult for users to perform this technique (Norman, 1983). Similarly, the critical incident technique is another measurement tool to measure users' mental models, whereby users are asked to project their thoughts into a critical incident and reflect on their decision making[15].

In the present study it was aimed to measure how fast and easy the users' mental models would adjust to the actual system and the designer's presentation of the system that he/she wants the user to build. The process required for adjusting the mental model can be conceptualized as learnability (Totzke et al., 2003). Learnability is one of the seven design principles of the DIN EN ISO 9241-110 for interface development.

[15] http://www.cog-tech.com/projects/mentalmodels.htm

As previously stated, our goal was to achieve a high quality of conceptual as well as spatial representation. Totzke et al. (2003) measured learnability of spatial representation by using the visual analogue scale (see Figure 28). The users were asked to rate the menu position by adjusting the controller. For a menu step high in the hierarchy, the controller should be positioned at the top.

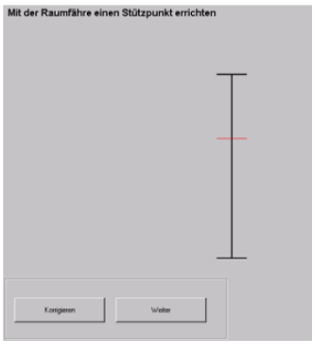


Figure 28. Visual analogue scale by Totzke et al. (2003)

For conceptual representation, Totzke et al. (2003) chose the card sorting paradigm (Miller, 1969) to explore users' knowledge of main categories and the corresponding subitems. The main focus in the present research was to measure the learnability of speech commandos and assign them in correspondence with the relevant content and subitems. It was decided to use retrieval tasks.

Halasz and Moran (1983) showed that another technique to measure the fit between the user's mental model and the actual system characteristics is to ask participants to perform a transfer task. Performance for a transfer task (e.g., reaction time) where the user can apply their knowledge gained in another task with the system can then be measured. The assumption is: the better the performance, the better the mental model.

4.3.3 Method

An experiment was conducted in the Telekom Innovation laboratories (see Figure 29).

Figure 29. Experimental setup of the fourth experiment

Experimental Design

The experiment consisted of four subtests. At first, a transfer task to compare the three shortening strategies with no shortening in an applied context was performed. Afterwards, the strategies were separately compared with no shortening to gain more differential knowledge of the appropriateness of the shortening (see Figure 30). Retrieval and navigation-orientation tasks had to be performed to test specific effects of shortening strategies on the mental models.

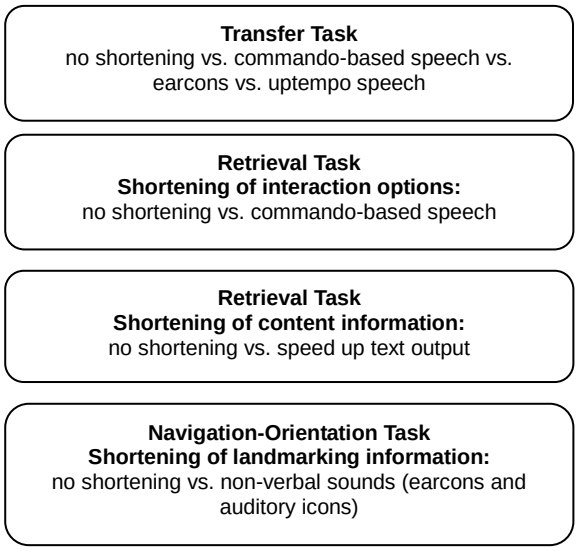


Figure 30. Procedure of the fourth experiment

Test Subjects

Forty-eight individuals were tested: 19 female and 29 male participants (Mean Age = 27, SD = 6). As no driving task was performed during this part of the research, no specific requirements applied (e.g., holding a driver's license, need for spectacle glasses).

For the transfer task, subjects were equally allocated across the four test conditions (no shortening, commando-based speech, earcons and uptempo speech). For the specific comparisons of the individual strategies, subjects were equally and randomly allocated to the two conditions.

Transfer Task

For the transfer task, three system versions were implemented on an Android-based smartphone (HTC Desire), each with different shortening strategies. With the system-initiated direct prompt, a speech-based land marking was presented (e.g., 'inbox' or 'main menu'). By using the LPA, an extensive prompt was given (including land marking, content information and interaction options). The system versions varied in the way the LPA outputs were shortened (see Table 19). For system 1, no shortening strategies were used.

Table 19. Independent Variables of the fourth experiment.

Shortening Strategies	**Independent Variable**			
	System 1	System 2	System 3	System 4
Land marking (non-verbal sounds)	Long	Long	Short	Long
Content (uptempo speech)	Long	Long	Long	Short
Interaction options (commando based speech)	Long	Short	Long	Long
N	12	12	12	12

The display was covered during the whole experimental phase simulating the dynamic driving situation requiring the visual attention allocated to the road. Subjects started in the main menu and first undertook an e-mail task. Subjects had to find the third mail in the inbox, read out the mail and answer by recording an audio file. For every interaction step they were asked to use the LPA to hear the extensive speech prompt and their respective shortening strategies. Subsequently they conducted the transfer task. Subjects again started in the main menu and were asked to search for an SMS in the inbox and have it read out by the system. For the

purposes of this task, using the LPA was not allowed. Usability parameters such as efficiency and effectiveness were measured: the number of speech command errors and the mean reaction time after each interaction step were collected.

Next to performance data, subjects were requested to complete questionnaires regarding mental demands and perceived quality of experience. The Subjective Assessment of Speech System Interfaces (SASSI) was used to measure speech dialogue acceptance. The SASSI developed by Hone and Graham (2000) is a standardized questionnaire designed for speech recognition systems. The level of agreement had to be indicated on a 5-point Likert scale for 34 positively and negatively formulated statements. The propositions had to been assigned to one of the following six dimensions: (1) System Response Accuracy, (2) Likability, (3) Cognitive Demand, (4) Annoyance, (5) Habitability and (6) Speed.

Retrieval and Navigation-Orientation Tasks

Testing Commando-Based Speech. For testing effects of commando-based speech on learnability, participants trained with systems 1 and 2 were asked to indicate the speech commands (interaction options) of the e-mail task in a multiple choice test. Recognition (e.g., in a multiple choice test) is a different process than free recall (Atkison & Juola, 1974). Thus, asking for specific wording was avoided, but instead focused on whether or not the participant retrieved the speech prompts and attributed them to the right menu step (e.g., “Was the option ‘next’ available in the e-mail inbox? Yes/No”). Correct answers and reaction time were measured.

Testing Uptempo Speech. After testing the commando-based speech, all test subjects were allocated into two groups. It was ensured that the same amount of subjects from each previous group was assigned to the new group. The trials for testing uptempo speech and nonverbal sounds were randomized to avoid learning effects. Five e-mail headers were read in sequence to the participants. For 24 participants, headers were read out with a 30% increase in velocity. The other 24 were read the same e-mail headers but with normal speed. To collect if information was correctly understood, a multiple-choice test as well as open questions were presented.

Testing Nonverbal Sounds. First, a training phase was conducted. One group heard the hierarchical order of the menu structure with the speech-based labels of every menu step. The other group heard the sound label in the hierarchical order via earcons and auditory icons. As

mentioned previously, the e-mail and SMS applications were presented via earcons, while infotainment and telephone application were presented using auditory icons. The earcons included hierarchical information, while auditory icons did not. Further, for the speech condition the menu labels for e-mail and SMS were hierarchical-based, as the superior category was always stated (e.g., “e-mail inbox”, “SMS inbox”), while for the infotainment application, only the label itself was named (“sport”).

After listening to the menu order once (either by speech or nonspeech), a sample of the labels was presented and asked to name the superordinated category (see Figure 31). Then, subjects were asked to define the menu depth using an abstract visual scale. The scale consisted of five levels (the menu hierarchy of system had five levels). As mentioned previously and according to Totzke et al. (2003), users’ spatial representation of the menu can be measured by imagining the menu hierarchy on a vertical one-dimensional scale. Subjects were asked to indicate the menu position on this scale. Finally, subjects were asked to recall the menu point after and before, as well as the menu step itself.

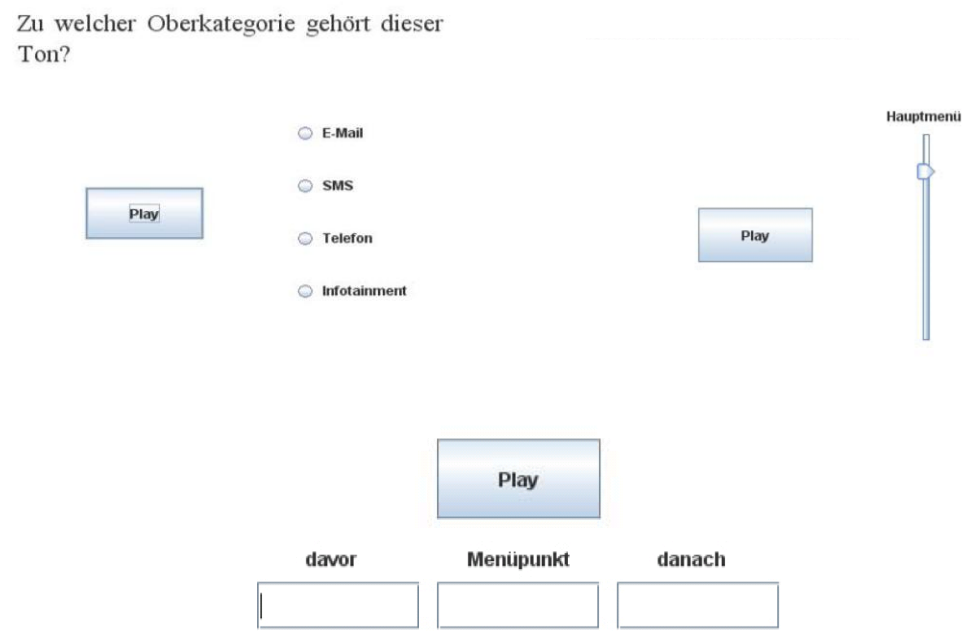


Figure 31. Testing of nonverbal sounds. Naming the subordinated category (up left), analogue scale (up right) and naming the menu steps before and after as well as the actual menu step (down).

4.3.4 Results

Transfer Task

There were no impairments expected to emerge in performance data for the transfer task for any of the three shortening strategies. If anything, it was assumed the commando-based speech could have positive effects. A one-way ANOVA was conducted for the time between two interaction steps. The time between two interaction steps included the duration of the direct prompt and the time until the push-to-talk button was pressed (in order to activate the speech recognition). Since the duration of the direct prompt was equal for all four groups, this was not analysed. Instead, the difference arose from the time participants needed to react to the prompt. A significant effect was found for time between two interaction steps ($F(3,41)=2.73$, $p=0.057$, $\eta^2=0.166$) as well as for number of speech command errors ($F(3,43)=3.25$, $p=0.031$, $\eta^2=0.185$). However, the Bonferroni test showed no significant differences for reaction time, which was likely due to the conservative nature of the adjustment. A tendency ($p<.10$) for duration was shown for nonverbal sounds (M=17.80, SD=4.01) compared to no shortening, M=22.46, SD=5.51, (indicating an improvement for the systems with sounds instead of speech), see Figure 32 for means and standard deviations. A significant difference ($p<.05$) of speech command errors of uptempo speech (M=2.33, SD=1.30) compared to no shortening (M=1.00, SD=1.10) was observed. The uptempo speech group made significantly more speech command errors than the no shortening group (see Figure 33).

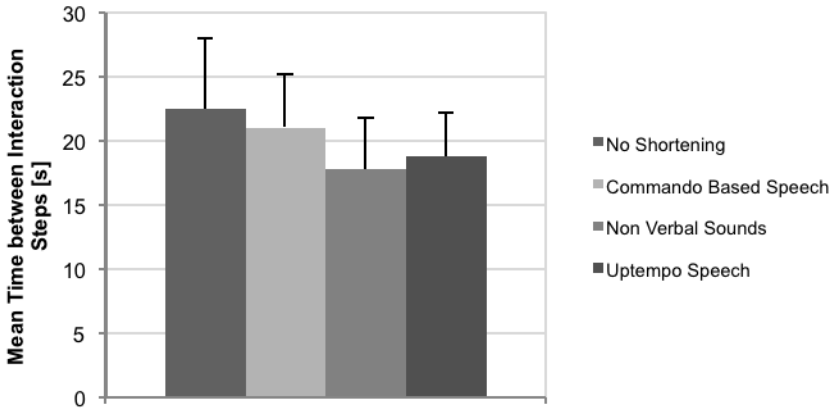


Figure 32. Means and standard deviations of the time between the interaction steps for the three shortening strategies and no shortening

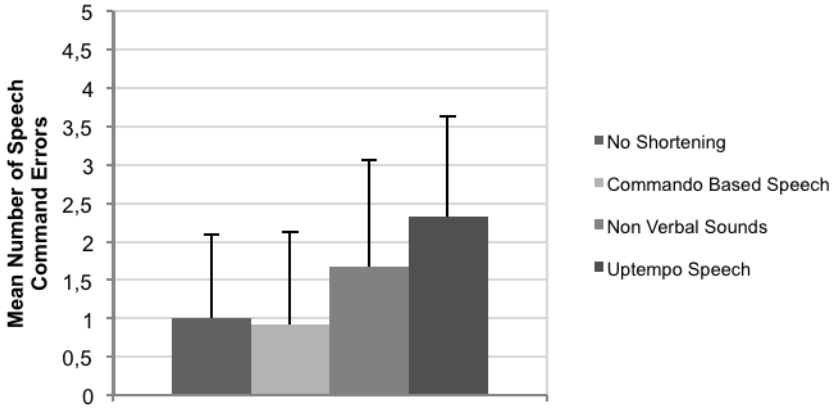


Figure 33. Means and standard deviations of the number of speech command errors for the three shortening strategies and no shortening

No significant effects were found for the Sea-scale or for the SASSI (all *p*'s > .05).

Shortening by Using Commando-Based Speech

No shortening (sentence-based speech) and shortening (commando-based speech) were compared. An independent *t*-test was calculated for correct answers of the multiple-choice test and duration to make an answer. No significant effect for the number of correct answers was observed. However, the group with the commando-based speech was significantly faster (M = 13.40, SD=2.71) compared to sentence-based speech (M = 19.18, SD=8.12, $t(12.20) = 2.24$, $p = .044$, d=0.10), see Figure 34 for means and standard deviation.

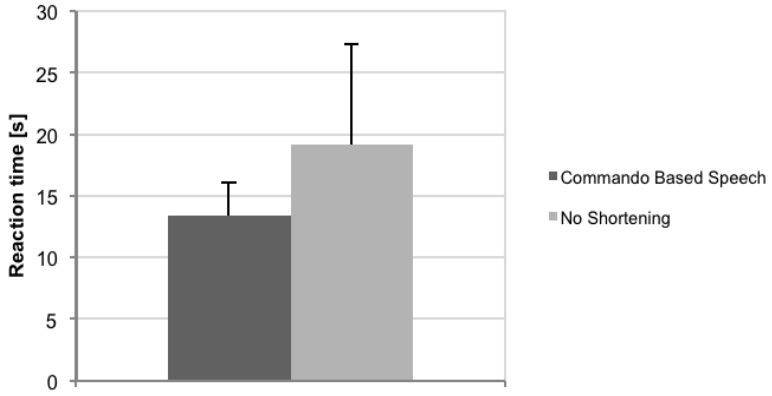


Figure 34. Means and standard deviations of the time between the interaction steps for the commando-based strategy and no shortening

Shortening by Using Uptempo Speech

Shortening by using uptempo speech had no significant effect on the number of correct answers (group no shortening: M = 2.63, SD = 1.66, group uptempo speech: M = 2.43, SD = 1.41) nor for the duration to set the answer (group no shortening: M = 9.44, SD = 3.00, group uptempo speech: M = 9.44, SD = 3.06) (all *p*'s>.05).

Shortening by Using Nonverbal Sounds

To test the effect of using nonverbal sounds instead of speech, the rate of correct answers for the analogue scale and naming the menu steps before and after (generic and narrow menu steps) were compared. The difference between sounds and speech (independent variable 1) as well as the difference between hierarchical information and non-hierarchical (independent variable 2) was investigated. Further, correct answers of naming the superior category for the different sounds (auditory icons = non hierarchical vs. earcons = hierarchical) were analysed. To this effect, a two-way mixed methods ANOVA was conducted with hierarchical information as the repeated measure factor. First, however, it was required to test the presupposition that auditory icons are easier to remember than earcons since the semantic link is stronger. Note that it does not make sense to compare speech with the sounds since the menu labelling would be 100% correct. The number of correct answers was set in relation to the amount of tested sounds (or speech prompts).

Correct Naming. A paired *t*-test was conducted. It was shown that the difference between the two groups was significant: Menu labels presented via auditory icons were significantly better remembered, M=0.64, SD=0.32, ($t(21) = 3.87$, $p=.001$, d=0.135) compared to menu labels presented via earcons (M=0.25, SD=0.27). See Figure 35 for means and standard deviations.

However, it was expected that the benefits of auditory icons compared to earcons for naming the menu step would be minimized if different aspect of the mental model were to be tested (e.g., spatial representation of the menu or order of the steps).

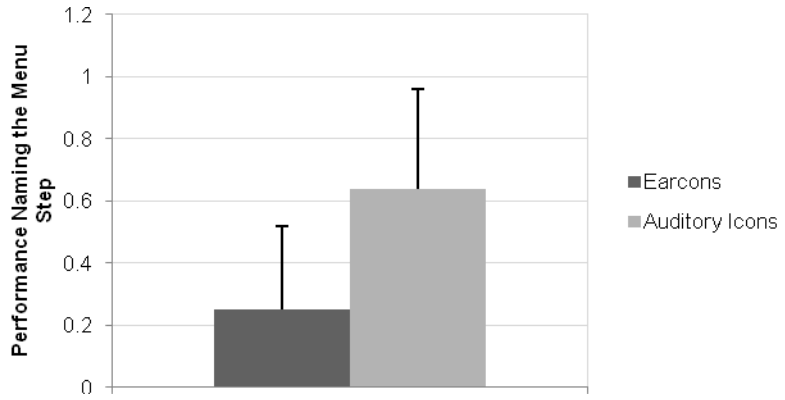


Figure 35. Means and standard deviations correct answer rate (performance) for naming the menu steps for earcons and auditory icons

Naming Superior Category. A paired *t*-test for comparing auditory icons and earcons by naming the superior category was carried out. Neither a significant main effect and nor a significant interaction effect was found ($p > .05$).

Analogue Scale. No main effect was found for hierarchy ($F(1.45= 1.48$, $p=.230$, $\eta^2 =0.032$). Also the main effect for shortening (speech vs. sound), $F(1, 45)= 2.47$, $p=.123$, $\eta^2 =0.052$, and the interaction effect ($F(1, 45)= 1.53$, $p=.469$, $\eta^2 =0.012$) were not significant. Figure 36 shows the means and standard deviations.

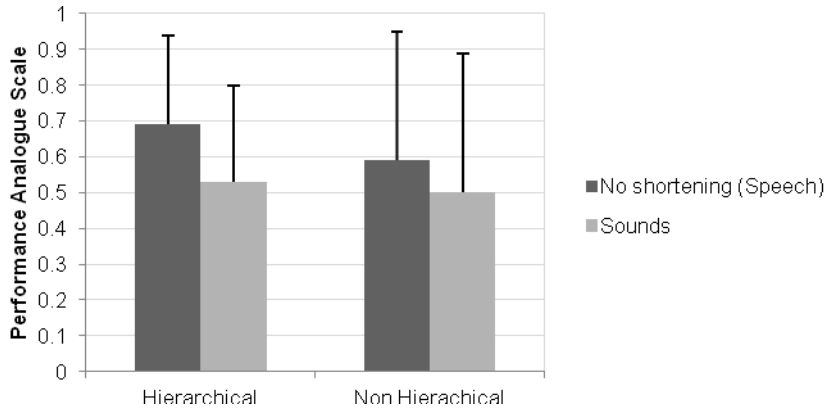


Figure 36. Means and standard deviations correct answer rate (performance) for the analogue scale for no shortening (speech) and sounds (earcons and auditory icons)

Naming Menu Steps Before and After. A significant main effect for shortening ($F(1,45)= 12.06$, $p=.001$, $\eta^2 =0.211$) as well as a significant effect for hierarchy ($F(1,45) = 5.87$, $p=0.019$, $\eta^2 =0.115$) was found. The interaction effect was not significant. Means are given in Figure 37. Performance rate for naming the menu step before and after was significant

decreased for non-hierarchical output (M=0.41, SD=0.35) compared to hierarchical structured output (M=0.51, SD=0.25). Also speech output (M=0.59, SD=0.28) led to significantly better performance than sounds (M=0.34, SD=0.27).

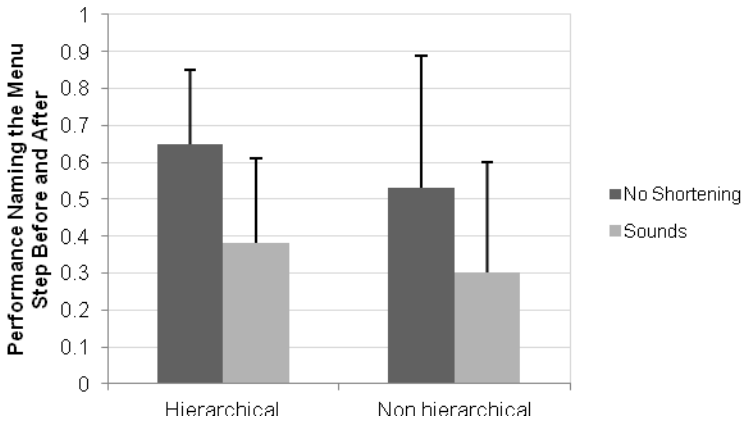


Figure 37. Means and standard deviations correct answer rate (performance) for naming the menu steps before and after for no shortening (speech) and sounds (earcons and auditory icons)

4.3.5 Discussion

A reduction of the time effort (shortening) of speech prompts inhibits the allocation of visual attention away from the road and on to the system (see Chapter 4.2.4). Shortening was achieved by using sounds for land marking information, uptempo speech for content information and commando-based speech for menu options. Since shortening could be linked to information loss, the present experiment tested the effect of shortening on users' mental models. Mental models develop during the interaction process and become more concrete. Ideally, the process would be sped up by building a mental model that corresponds to the developer's image of what users should know about the system; this is somewhat comparable to learnability provided by the system. The mental model was measured via a transfer task as well as multiple choice tests (instead of free recall) and mental spatial representation of the menu structure for comparing the shortening strategies with no shortening. Additionally, subjects were asked to name generic and narrow menu steps as well as the superior category of menu steps.

For the transfer task, only shortening by uptempo speech had significant negative effects on the performance data. Sounds, as opposed to speech, even led to a faster performance (tendency, not a significant effect). It is however noteworthy that it is not clear whether the lack of significant difference is indeed caused by the fact that the shortening strategies do not negatively influence performance in general; the results could also originate from flaws in the

experimental design or low statistical test power. However, it serves to tentatively suggest that replacing speech with commando-based speech and sound will not negatively influence the mental model.

In the more detailed test comparing speech and the particular shortening strategies, it was shown that shortening of interaction options via commando-based speech had a positive effect. Indeed, it did not lead into an increased recognition performance in terms of correct answers, but rather, a faster recognition time was achieved. This positive effect on users' mental models is in line with findings that irrelevant information interferes with recognition (Vilimek, 2007) and the word-length effect (Baddeley et al., 1975).

Earcons led to poorer recognition of the exact label (25%) of the menu step than auditory icons (63%). Asking for the speech-based menu labels was not meaningful as these are 100% semantically linked, and as such, can confidently be assumed to be more effective than sounds. Note that the earcons and auditory icons in this setting were presented without any other information, which makes the recognition or labelling more difficult. In the CLD system the sounds are usually presented with content information or interaction options and only on the LPA prompt. In the direct prompt the label is still presented by speech. Further analyses are necessary to determine the recognition rate with presentation of more context information. Also note that the participants only heard through the earcons set on two occasions. More studies regarding learning effects are essential. However, as land marking information is supposed to build good spatial representation faster, tasks that aimed to gain more information regarding appropriateness of earcons and auditory icons for overall representation of the menu hierarchy and structure were performed. As the earcons included hierarchical information, it was expected that the positive effects of a strong semantic link (non hierarchical speech or auditory icons) would be adjusted in the analogue scale task and in the naming of the menu steps before and after. In the analogue scale task, subjects were asked to rate the menu level either by listening to a speech prompt or alternatively to a sound. Sounds and speech prompts were either of hierarchical character (sounds=earcons, speech=category, menu step, e.g., e-mail Inbox) or of non-hierarchical character (sounds=auditory icon, speech=menu step, e.g., News). It was shown that for the task of 'naming the menu step before and after', hierarchical information had a benefit compared to non-hierarchical land marking. However, speech compared to sounds lead to significant increased performance rates compared to sounds. For the analogue scale no significant effects were observed. This could support the assumption

that advantages of speech over sounds will be adjusted for spatial representations but since no significant effect occurred the results could also be explained by flaws in the experimental design. Overall, in respect to the exact labelling (recognition) of the menu step, earcons have disadvantages over auditory icons, and even higher disadvantages in comparison to speech. For spatial representation (orientation within the menu structure and level), hierarchical land marking information (like earcons) has benefits over non-hierarchical information (like auditory icons).

Table 20 gives an overview of the time reduced by using single shortening strategies compared to each other as well as the effects on the mental model. Recommendations are also given for shortening speech for in-car VUIs. Shortening interaction options by using commando speech (enriched with different voices) as compared to sentence-based structure is highly recommended. This results in reduction of the duration of speech prompts significantly, and also leads to better learnability. Earcons also reduce the time effort compared to speech, but not as much as commando-based speech. Note that in Table 20, a '0' was listed for the earcons and their effects on the mental model (which indicates no negative effects on users' mental model compared to speech). This is only valid for non-hierarchical speech prompts and earcons on the performance of analogue scale (i.e., spatial representation). Sounds always need to be learned beforehand, and the recognition rate is reduced compared to speech. The designer needs to decide whether they wish to provide a good recognition rate for the exact menu label, or if it is sufficient to provide a good spatial representation. This also depends on the relevant information content provided with the earcons.

Table 20. Recommendation based on time effort and effects on the mental model

Effort	**Shortening of…**		
	Interaction options by commando-based speech	Land marking by earcons	Content by uptempo speech
Time reduction	++	+	+
Effects on mental model	+	0	-
Recommendation	++	+	0

5. Beyond Pragmatic Qualities: Influence of Hedonic Aspects on Attention Allocation

So far, it has been shown that the tendency to glance at the display of infotainment systems could be reduced by improving the speech output.[16] Due to this, the relevant information content for speech output was increased and adjusted in accordance with the relevant information content presented via the display. At the same time, the duration of speech prompts was reduced by various shortening strategies. These changes are considered to increase efficiency of the speech output. However, infotainment systems belong to a class of systems that are referred to as 'comfort' systems. Comfort systems involve additional, hedonic quality and satisfaction demands, in order to obtain a positive user attitude towards the system. The design of GUI often implies a nice 'look and feel'. When designing speech and sound output the 'listen and feel' has not, until now, been extensively considered or to the same extent as visual systems. As such, the following experiment was conducted to investigate whether increasing the hedonic quality of speech output will influence attention allocation to the display, under the assumption that quality of experience is not limited to pragmatics, but rather also includes obtaining an impressive 'listen and feel'.

By increasing the relevant information content, expectancy and value of the SEEV Model (Wickens et al., 2003) were increased: higher bandwidth of information and task relevance of information. This can be considered as increasing the pragmatic quality of a system. According to Hassenzahl (2008), pragmatic aspects support achieving so-called do-goals. A do-goal could be, for example, to answer an e-mail by using speech input. Hence, the user needs to know the speech commands to get into the e-mail inbox and to reply to the received mail. By providing all interaction options acoustically (see Chapter 4.2.1), the user knows what to say without glancing at the display. The pragmatic aspect in the presented example is the presentation of possible interaction options via the VUI. In this sense, the value (relevant information content) represents the relevance of the presented information for reaching the task goal, i.e., solving the infotainment task. As previously mentioned, in addition to pragmatic aspects, hedonic aspects also characterize systems (Hassenzahl, 2008). Hedonic aspects support fulfilling users' psychological needs (be-goals). An example for a be-goal is to

[16] This chapter is a slightly modified version of (Niemann et al., 2011a).

be stimulated. This could be achieved via hedonic aspects such as aesthetic design and new interaction paradigms.

Higgins (2006) proposed that hedonic experiences of joy or pain form the basis for value. This leads us to the following question: Can the value of speech output be increased, and thus, can eye gazes away from the road (and towards the display) be reduced by increasing the hedonic quality of the voice user interface (VUI)? Does a factor $value_{(hedonic)}$ exist in the model for predicting attention allocation? In Chapter 2.2.2 it was assumed that there are no differences for the two modalities regarding the overall value or importance within the multitask context since the same task was performed with the two modalities. However, there could be differences for the hedonic quality. The aim of the following study was to examine precisely this question.

5.1 Increasing the Hedonic Value

According to Hassenzahl (2008), hedonic aspects fulfil users' needs. The needs defined by Hassenzahl (2008) are stimulation, autonomy, competence, relatedness, and popularity. Two strategies for increasing the hedonic quality have been chosen for the study:

- Satisfying the need for stimulation by increasing the aesthetics
- Satisfying the need for autonomy through personalization, i.e., providing choices

The object of investigation in the present study is the form of information presentation (i.e., graphical vs. acoustic). Therefore, other product characteristics such as content, functionality, and interaction should be kept stable (see product features according to Hassenzahl, 2003). Also, the other SEEV Model parameters should not be influenced. Given these considerations, the choice of strategies for enhancement of hedonic aspects was limited.

The variation of aesthetics as well as personalization was conducted for both the earcons and auditory icons, which each represent land marking information. For variations in user need for stimulation, sound sets of varying aesthetics representing menu structures by earcons were developed by a sound designer. According to Mahlke (2007), the user is stimulated by aesthetic visual design. Similarly, Hassenzahl (2003) defines stimulation as a "new, interesting, and inspiring […] presentation style" (p.187). For increasing only aesthetic aspects of the sounds presented, it was necessary to keep the functionality of the sounds (in terms of the representation of menu structures) and thus the sound structures constant between

the unaesthetic and aesthetic sound set. Sound structures in this regard refer to the specific tone colour for identifying the main categories of the menu, as well as the addition of sounds of varying tone pitch used for identifying the menu depth within categories (see Chapter 4.2.2). For the unaesthetic sound set, minor intervals were used in contrast to major intervals for the aesthetic sound set. Major intervals used are known to appear more friendly compared to minor intervals, which appear sad (Cook, 2006). Dissonant sounds are perceived as unpleasant (Raffaseder, 2002). Therefore, for the unaesthetic sound set, also dissonant sounds were used which have non-harmonic overtones, not fitting into the natural acoustic spectrum. In addition, a compression and bandwidth limitation (300Hz - 3.4KHz) was conducted, resulting in a subjectively worse perceived quality (Vickers, 2010).

In addition to the aesthetic (modern) and unaesthetic (synthie) sound set, another aesthetic sound set (classic) was developed. Here, the tone structure was also kept constant. The earcons of the third sound set are produced by musical instruments such as a harp or violin, and are thus non-synthetic sounds unlike the other synthetically designed sound sets. The option to select one of the three sound sets was implemented in a Flash animation called ‘Sound Market’. Figure 38 illustrates the Sound Market interface.

Figure 38. Sound market interface

This personalization in terms of offering choices for the sounds being presented is assumed to be in line with the user’s need for autonomy. It has been shown that positive effects on the user’s attitude can be achieved by the mere offering of choices between alternative tasks or systems (autonomy) which in turn enhance the intrinsic motivation for a multitude of tasks (Deci, 1975, 1981; Deci, & Ryan, 1985; Langer, & Rodin, 1976; Taylor, 1989). In addition, the option of personalization results in a more positive attitude towards a system in terms of higher acceptance (Cordova, & Lepper, 1996), more joy of use (Blom & Monk, 2003), and a stronger feeling of relatedness with the product (Mugge, Schifferstein & Schoormans, 2004).

5.1.1 Experiment 5: The Influence of Hedonic Aspects on Attention Allocation

The aim of the final driving simulator study was to examine if increasing the value by enhancing the hedonic quality of sounds results in a reduced number of percentage dwell time towards the display. Can the tendency of attention allocation to the display be lowered by developing aesthetic sounds and offering choices between different sound sets?

5.1.2 Hypotheses

The following predictions were made:

H6: Providing choices (selection options) results in higher intrinsic motivation than offering no choices.

H7: A distortion of sounds (sound set synthie) leads to a lower subjective rating of the sound valence than no distortion.

H8: Providing choices results in a lower percentage dwell time to the display than offering no choices.

H9: A distortion of sounds (sound set synthie) leads to a higher percentage dwell time to the display than no distortion of sounds.

H10: Providing choices results in a higher overall system acceptance (likeability) than not providing choices.

H11: A distortion of sounds (sound set synthie) leads to a lower overall system acceptance (likeability) than no distortion of sounds.

5.1.3 Method

Experimental Design

Three groups were compared in a between-subject design with the independent variable 'degree of hedonic quality enhancement', graduated in:

- System 1: sound set selection option, aesthetic sounds
- System 2: no sound set selection option, aesthetic sounds
- System 3: no sound set selection option, unaesthetic sounds

For offering choices, subjects in Group 1 (system 1) were given the option to select a sound set (classic, modern, or synthie). Subjects in Group 2 (system 2) received the same sounds that the group using system 1 had chosen. Thus, they received aesthetic sounds but played no part in the selection of sounds, in order to ensure that both groups differed only in respect to choice, but not aesthetics. The sound set synthie was assigned to all subjects in Group 3 (system 3). This sound set did not differ from systems 1 or 2 regarding functionality or structure of the sounds (see Chapter 4.2.2). The difference concerned only the spectral limit, the use of dissonant tones and minor instead of major intervals, resulting in a variation of aesthetics. As assumed, no one from Group 1 actually chose the sound set synthie.

Test Subjects

After excluding the invalid data sets due to insufficient data quality (less than 75% eye gaze capture), the data of 42 subjects were analysed. For participation, the minimum age was 18 while the maximum was 50. On average, subjects were 29.6 years old. 23 females and 19 males participated in the driving simulator test. They received 20 Euros for their two hours of participation in the experiment.

Materials

Driving Task. The experiment was again conducted in the driving simulator at the Centre of Human Machine Systems at the TU Berlin (see Chapter 3.1.4 for more details). Again, the test subjects were asked to follow a specific car at a constant distance, just as in experiments 1 and 3.

Secondary Task. The HTC Desire with the implemented information and communication applications E-Mail, SMS, News, Travel Guide, and Phone was used as the secondary task.

Sounds. The three sound sets representing the menu structure were developed with the digital audio workstation software Cubase V.5 (Steinberg Media Technologies GmbH), and with the sound data Symphonic Orchestra (EastWest Communications Inc.). For sound selection, a flash animation running on a Windows PC (see Figure 39) was used. The menu structure was presented with icons of the different menu steps. By clicking on the icons, participants were able to hear the sound for the represented menu step. Using a selection menu in the upper right corner, different sound sets could be chosen.

Eye Tracking. The eye tracking system used to record participants' eye gazes during driving and task execution was the fully-mobile, headmounting system iView X HED from SensoMotoric Instruments GmBH, just as in experiments 1, 2 and 3.

Questionnaires. The Self Assessment Manikin (SAM; Bradley & Lang, 1994) questionnaire was used to collect subjective evaluations of the sounds and the rating of the aesthetics. Good validity for evaluating acoustic stimuli was previously reported by Bradley and Lang (2000) for the SAM. In order to test if providing choices actually fulfils the need for autonomy and thus enhances intrinsic motivation, the Interest subscale from the Intrinsic Motivation Inventory was used (IMI; McAuley, Duncan & Tammen, 1989). After the test was finished, the standardized questionnaire Subjective Assessment of Speech System Interfaces (SASSI, Hone & Graham, 2000) for the evaluation of speech dialogue systems was applied for an overall evaluation of the system.

Procedure

First, subjects were introduced to the functionalities of the applications and to the operation of the speech dialogue system, using the e-mail application as an example. Subsequently, subjects were trained to perform the driving task.

The first group had the opportunity to select one of the three sound sets by using the Flash animation program. After subjects had made their choice, they were instructed to listen again to all sounds of the chosen sound set. Similarly, the second and the third groups were asked to click on every icon of the assigned sound set of the Flash animation in order to familiarize themselves. Afterwards, all subjects filled out the SAM questionnaire. This was followed by

another test drive, which entailed completing the driving task and e-mail task concurrently. The test phase consisted of three trials: one baseline trial (driving without interacting with the system), and two untrained tasks while driving with the CLD system (see Table 21).

The order of the three tasks was randomized to avoid the possibility of learning effects. After completing the task, the Intrinsic Motivation Inventory and the SASSI were completed.

Table 21. Untrained infotainment tasks.

Applications involved	Task
News and Telephone	"Go to the application "News". Get the second article of the category "culture" read out. Remember the headline. Because it is very interesting you want to tell your friend about it. Go to the application "phone" and call your friend "Max Mustermann". Speak the approximate headline on his mailbox and suggest he read the article."
Travel Guide and SMS	"Go to the application "travel guide" and start a new route. Go to the SMS application and let the first SMS be read out. By then you will have passed a site (you will be notified by a sound). Go to the application travel guide and get information read out about the site you've passed."

5.1.4 Results

Subjective Ratings

Questionnaires yielded the following results in relation to intrinsic motivation, sound valence, and overall rating of the system.

Intrinsic Motivation Inventory (IMI).

H6: Providing choices (selection options) results in higher intrinsic motivation than offering no choices.

In order to test whether providing choices would lead to higher intrinsic motivation compared to no choice (by keeping the sound stimuli stable), the means of system 1 and system 2 regarding the independent variable Interest of the Intrinsic Motivation Inventory (IMI) were compared with the help of an independent *t*-test.

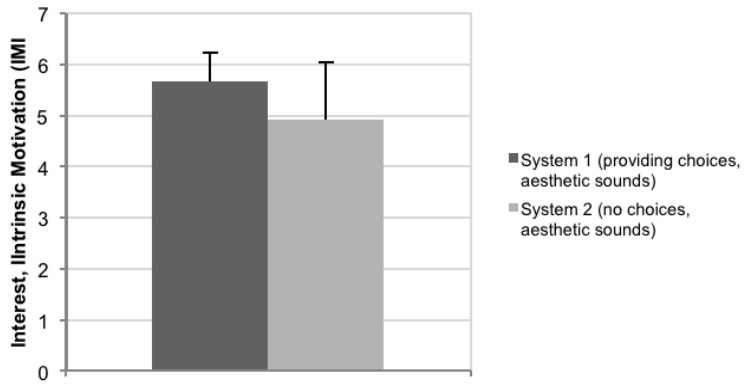


Figure 39. Means and standard deviations for interest (intrinsic motivation, IMI) for systems 1 and 2

It was shown that, as predicted, use of system 1 (M=5.66, SD=0.57) resulted in higher intrinsic motivation ($t(28)=2.26$, $p=0.031$, d=0.94) compared to use of system 2 (M=4.92, SD=1.12), see Figure 39 for means and standard deviations. Thus, hypothesis 6 was confirmed.

Self Assessment Manikin (SAM).

H7: A distortion of sounds (sound set synthie) leads to a lower subjective rating of the sound valence than no distortion.

A one-way ANOVA showed a significant effect for the variable valence ($F(2,39) = 27.96$, $p < .001$, $\eta^2=0.56$). As expected in Hypothesis 7, the Bonferroni post-hoc test showed a significant difference between system 1 (M=7.13, SD=1.25) and 3 (M=3.75, SD=1.49; $p<0.001$) and between 2 (M=6.57, SD=1.09) and 3 ($p<0.001$), see Figure 40 for means and standard deviations. No significant effect between systems 1 and 2 was found. No apriori assumptions were made regarding this comparison.

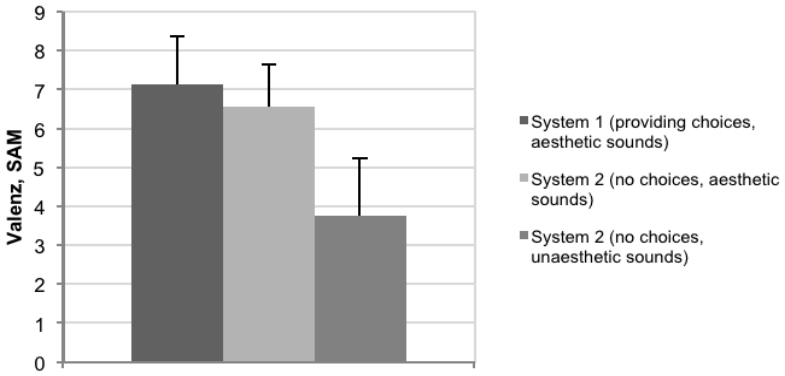


Figure 40. Means and standard deviations for valence (SAM) for the three systems

For the variable arousal a significant effect was observed ($F(2,39)=3.69$, $p=0.034$, $\eta^2=0.163$). System 3 (no choices and unaesthetic sounds, M=3.60, SD=0.74) was rated as more arousing than system 1 (M=5.50, SD=1.01), $p=0.03$. The post-hoc test shows no significant differences between system 2 and 3 as well as between system 1 and 2.

Eye Movements

H8: Providing choices results in a lower percentage dwell time to the display than offering no choices.

H9: A distortion of sounds leads to a higher percentage dwell time to the display than no distortion.

Eye tracking was analysed using percent dwell time (PDT) on Display. A one-way ANOVA was conducted. The main effect system was significant $F(38,2)=4.28$, $p=0.021$, $\eta^2=0.184$. A post-hoc Bonferroni test showed no significant difference between systems 1 and 2. Thus, Hypothesis 9 could not be confirmed. However, a significant difference ($p=0.019$) between system 1 (providing choices and aesthetic sounds, MW=4.17, SD=2.57) and system 3 (no choices and distorted sounds, MW= 15.03, SD=2.84) was observed. The difference between system 2 (no choice but aesthetic sounds, MW= 6.58, SD= 2.62) and system 3 was not significant. However, the means indicate a tendency in the expected direction (system 2 = 6.58%, system 3 = 15.03%), which would lend some support for Hypothesis 9. The PDT Display for the three systems is presented in Figure 41.

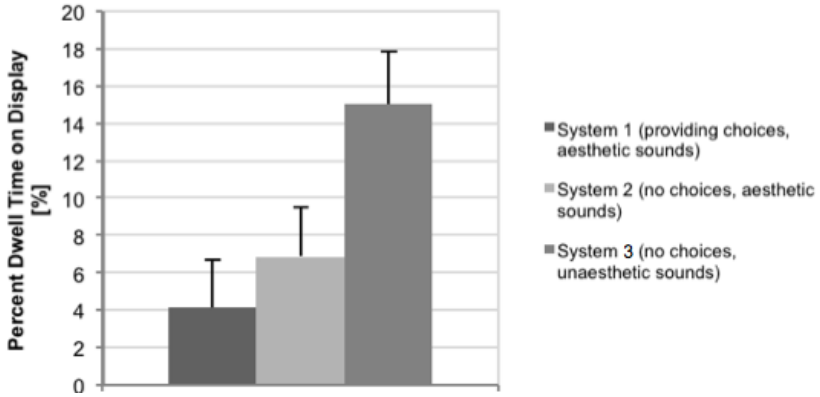


Figure 41. Means and standard deviations percent dwell time on display for the three systems

Subjective Assessment of Speech System Interfaces (SASSI).

H10: Providing choices results in a higher overall system acceptance (likeability) than not providing choices.

H11: A distortion of sounds leads to a lower overall system acceptance (likeability) than no distortion.

Results were analysed using one-way ANOVAs for each of the six dimensions: Habitability, Annoyance, Speed, Cognitive Demand, Likeability, and System Response Accuracy. For subjective ratings of perceived hedonic quality, the critical dimensions were Likeability and Annoyance. Significant effects were shown for the dimensions Likeability ($F(2,39)= 4.13$, $p=0.012$, $\eta^2 =0.173$), Annoyance ($F(2,39)= 4.13$, $p=0.006$, $\eta^2 =0.177$), and Habitability ($F(2,39) = 3.74$, $p=0.016$, $\eta^2 =0.141$). Scheffé's test showed a significant effect of system 1 compared to systems 2 and 3 for the dimensions Likeability and Annoyance (Hypothesis 10). However, no significant differences between systems 2 and 3 were observed (Hypothesis 11). The means of the three SASSI dimensions for the three system versions respectively are presented in Figure 42 and Table 22.

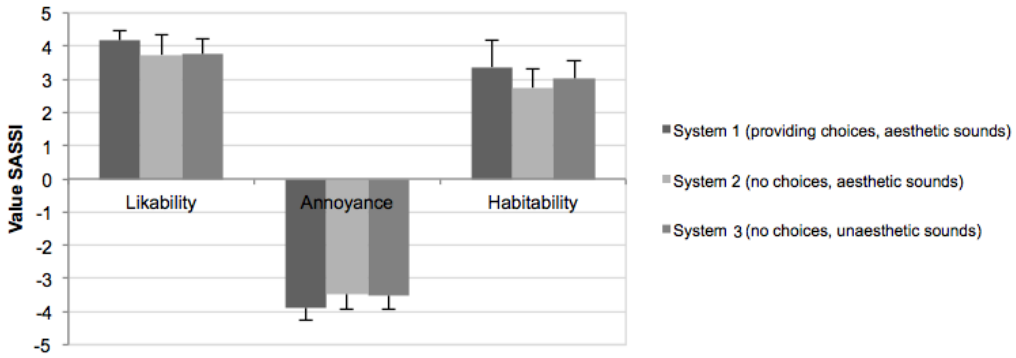


Figure 42. SASSI ratings for the three systems

Table 22. Means and standard deviations for SASSI (Likability, Annoyance and Habitability) for systems 1, 2 and 3

		Mean	Standard Deviation
System 1 (providing choices, aesthetic sounds)	Likability	4.18	0.29
	Annoyance	3.88	0.38
	Habitability	3.37	0.82
System 2 (no choices, aesthetic sounds)	Likability	3.74	0.61
	Annoyance	3.49	0.45
	Habitability	2.75	0.56
System 3 (no choices, unaesthetic sounds)	Likability	3.77	0.44
	Annoyance	3.50	0.43
	Habitability	3.03	0.54

Driving Data

Lane deviation and variance in the distance to the lead car were analysed. No significant effects were observed for driving performance (all p's >0.05).

5.1.5 Discussion

The aim of the present study was the enhancement of hedonic quality and the corresponding increase of the parameter overall value in Wickens et al.'s SEEV Model (2003). The goal was to evaluate if this increase reduces the tendency to allocate attention to the display and, thereby, reduce reception of visual information in place of acoustic reception.

For this purpose, three versions of a speech-based infotainment system containing information and communication services were implemented. The systems varied according to their degree of hedonic quality enhancement. For the system with presumably the highest hedonic value, the hedonic quality was, firstly, enhanced by use of aesthetic sounds. Secondly, the system was enhanced by providing the user with choices (personalization), thus increasing users' intrinsic motivation. The system with lower hedonic quality was characterized by aesthetic sounds but no offer of choice. The third system (lowest hedonic quality) contained unaesthetic sounds but again no choice option.

As the results showed, when sound aesthetics are enhanced compared to an unaesthetic sound set, percent dwell time towards the display was reduced (indicated by a tendency in the expected direction, but no statistical significance). When choices for sound sets are also provided, this further intensifies this reduction in eye gazes, resulting in a significant

difference compared to no choices and unaesthetic sounds. Thus, there is support for our prediction that increasing the parameter value of the SEEV Model by enhancing the hedonic quality indeed reduces users' need to look to the display when speech output is given. However, the difference between the conditions where choosing sound sets was offered versus not offered (while aesthetic aspects were kept constant) was not significant (and was not even close). Also for rating the sound valence, there was no difference between the conditions where choice was offered versus not offered, while aesthetic aspects were kept constant. Nonetheless, statistically significant effects were shown for the rating of perceived intrinsic motivation and the hedonic dimensions of the SASSI (retrospective overall rating of the system), while aesthetic sounds without providing choices had no effect on overall rating of the system (although on the valence scale of the SAM questionnaire, the aesthetic system was rated significant higher than a system with nonaesthetic sounds).

These results indicate that distinctions between different methods of enhancing the hedonic quality must be made. In one respect, there is an affective experience (valence scale, SAM) during the interaction, which also seems to have a stronger influence on attention processes (variables measurable directly online). In another respect, however, there is a long-term enhancement of hedonic quality caused by cognitively assessed emotions, which is reflected in evaluative subject reports (e.g., questionnaires, ratings) subsequent to the interaction. Thüring and Mahlke (2007) also distinguish in their CUE model (component model of user experience, see Figure 43) between actual experienced emotions, which are measured via physiological data, subjective emotions (SAM), motoric expressions, and overall acceptance ratings following an interaction. The model also distinguishes between non-instrumental aspects directly influencing overall acceptance ratings without thereby resulting in different emotional experiences in terms of affect (such as in the present study where choices were offered). Questionable is the lack of increased acceptance through enhanced aesthetics in the overall judgment despite the increase in emotional experience. One explanation could be that sounds compared to graphical information do not have an enduring effect, given that overall judgement was only first surveyed after the experimental task was completed; Paivio Philipchalk and Rowe (1975) illustrated that sounds are not as memorable as visual non-verbal material. This needs to be investigated more fully in light of the process model of user experience.

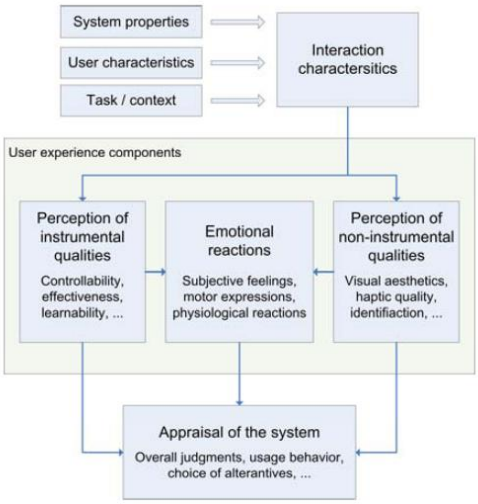


Figure 43. CUE model of user experience (cited from Thüring & Mahlke, 2007, p.262)

However, it is very important to mention that the last discussed effects could also be materialised due to the fact that, for variation of aesthetic, the hedonic aspect was decreased, while for personalization, the hedonic aspect was increased. This should be further investigated.

Next to the more differentiated results regarding the specific influences of different hedonic aspects on user experience parameters, the most important finding of this study was the influence of hedonic quality on attention allocation.

6. Consolidation and Future Work

Drivers have the tendency to retrieve information via the visual output modality of in-car infotainment systems while driving despite the presence of a speech output (Kun et al., 2009). First it was considered which characteristics of visual output compared to speech output affect glancing at the display instead of listening to the speech prompt.

The attention allocation model by Wickens et al. (2003), the so-called SEEV Model, was used to gain more detailed insights. Within the model specific parameters are defined which can be used to predict the probability of attention allocation to an information source $P(A_{IS})$. In Chapter 2.2 of the thesis the parameters of the SEEV Model were individually theoretically analysed. The parameters were more specified and differentiated to enable comparison of the two different modality outputs. A more detailed differentiation for the SEEV Model parameter effort was made: $Effort_{between}$ and $effort_{within}$. While $effort_{between}$ describes the costs to switch from one information source to another (task switch and location switch costs), $effort_{within}$ was defined as the duration to perceive information presented by an information source. It was assumed that $effort_{within}$ best be set in relation to the relevant information content of an information source. Furthermore, the parameters expectancy and value were analysed more closely and the relevant information content was defined (corresponding to the parameters expectancy and value). Afterwards it was theoretically investigated by which of the SEEV Model parameters speech output has disadvantages, where visual output does not, and therefore results in glancing away from the street and allocating attention towards the display of in-car systems. First, with visual output it is possible to perceive information in a shorter time period (lower $effort_{within}$). Presenting all of the information shown on the display acoustically is very time consuming (high $effort_{within}$). Therefore, in common in-car infotainment systems with speech output, less information will be presented via the speech output compared to the visual output. Mostly possible options will not be presented (speak-what-you-see principle). As options are items that are especially relevant for novices, this results in lower relevant information content. Second, it was assumed that graphical information has the benefit of higher controllability: Information acquisition is user-initiated by a conscious motor action (eye movements towards the display). For speech the $effort_{(between)}$ (switching costs) is ‘too low’ since no conscious motor action is necessary to request the speech prompt. As a consequence, the controllability for in-car speech output regarding the point of time the speech output is presented is decreased. In-car infotainment

speech output usually requires attention in the very moment it is presented (system initiated). In two experiments using the secondary task paradigm of driving and performing a secondary task, it was tested whether increasing or decreasing the value for these parameters (i.e., $effort_{within}$, relevant information content and $effort_{between}$ for speech would lower attention allocation towards the display. Tendency to allocate attention towards the display was operationalized by measuring percent dwell time on display. In experiment 1, significant support for the prediction that decreasing the time effort ($effort_{within}$) as well as increasing the relevant information content for speech compared to visual output leads to less time spent glancing at the display. In experiment 2, it was found that lower controllability of speech is not a problem when performing a lowly demanding driving task. However, if highly demanding driving interrupts the speech output of the secondary task (e.g., infotainment), tendency to glance at the display decreases. In further relation to this, it was observed that increasing controllability for speech prompts by increasing $effort_{between}$ (by introducing a button at the steering wheel which needs to be pressed to access the speech output – a conscious motor action) lowers the tendency to allocate attention towards the display. Next to these insights additional knowledge regarding the SEEV Model was gained: It could be proven that the SEEV Model parameters effort, expectancy and value are also applicable to the auditory modality (which, before, remained to be evaluated). Second, it was proven that a specific characteristic of the parameter effort – namely, $effort_{within}$ – influences attention allocation. This research has been, it seems, the first study to experimentally investigate this theoretical assumption.

Based on the insights derived from the first two experiments, in the second part of the thesis design recommendation were deducted for a Deutsche Telekom in-car infotainment prototype (CLD prototype) and tested in an applied context. Speech output for in-car information and communication apps was designed based on the results of study 1 and 2 of this thesis. A primary goal was to keep the visual output of the infotainment apps the same while enhancing the speech output in terms of the factors believed to cause disadvantages in speech (compared to visual) information. Therefore, the following revisions were made for the infotainment apps:

- Providing controllability by implementing an LPA (long-term push-to-activate)
    - User-initiated request for extended speech output

- Providing the same relevant information content acoustically as is available visually
    - all information given on the screen were presented acoustically upon pressing the LPA
- Reducing the time effort of the speech prompts (reducing the duration):
    - Shortening of valid speech commands by using commando-based speech (further supported by use of different voices)
    - Shortening of landmarks by using sounds
    - Shortening of content information by speeding up

The probability of allocating attention towards the display ($P(A_{Display})$) of a multimodal infotainment system was assumed to be minimized when using theses design recommendations to keep visual distraction as low as possible.

Next to these pragmatic aspects, hedonic quality of the VUI was enhanced by using aesthetic sounds and by providing choices. These design recommendations can be referred to as the parameter value of the SEEV Model. Further, it was also investigated whether hedonic aspects also influence attention allocation. See Table 23 for a summary of all design recommendations.

Thus, in addition to developing the VUI concept, the strategies were also evaluated in terms of their influence on percent dwell time to the display of the CLD prototype. Therefore, six different VUI versions were implemented (see experiments 3 and 5). The display design of the CLD prototype, as well as the driving task and all other parameters, were kept constant throughout all experimental conditions to ensure that the reduction of visual distraction was only achieved by enhancing the speech output design. The test subjects would not be instructed to avoid looking at the display. Instead, it was proposed that the tendency to allocate attention to the display may naturally subside by improving the speech output.

Table 23. System versions on the parameters of the revised model

Differentiated SEEV Model	Effort(between)	Effort(within)	Relevant information content (Expectancy* $Value_{(pragmatic)}$)	$Value_{(hedonistic)}$
Design Recommendations	Long-term push-to activate → **$Effort_{(between)}$ medium**	Shortening of valid speech commands by using commando-based speech (support by different voices), shortening of land marking by using sounds, shortening of content information by speeding up → **$Effort_{(within)}$ low**	Presenting the same amount of relevant information visually and acoustically → **Relevant information content (Expectancy* $Value_{(pragmatic)}$) high**	Providing choices for sound selection and enhancing the acoustic aesthetic → **$Value_{(hedonistic)}$ very high**
Goal	Enables to start the speech output user initiated	Faster perception of relevant information presented via speech	Omitting the speak-what-you-see principle. Almost all information, which can be heard, can also be read on the display.	Increases the hedonic qualities of the VUI

Table 23 highlights for which parameters of the SEEV Model a change was made. Based on the results in the first two studies of this thesis (see Chapters 3.1.1 and 3.2.1), and also based on a theoretical analysis, a ranking system was made for the VUIs in respect to their influence on percent dwell time towards the display. Since PDT Display was assumed to be a reverse measurement for $P(A_{Speech})$, a high benefit for the VUI results in a low value of PDT Display. Note that $effort_{between}$ is not reflected in this ranking since the experimental setup of experiments 3 and 5 did not allow to test the positive effects of integrating the long-term push-to-activate to increase controllability (see Chapter 3.2). The values (low, medium and high) in Table 24 were set in comparison to one another. The plus and zero in parentheses again describes if value increases $P(A_{Speech})$ and thus decreases PDT Display.

Table 24. System versions on the revised model parameters, sorted by the theoretically assumed probability to allocate attention towards the display. Codes represent the system versions and the experiment in which the system was tested (Exp=Experiment, Sys=System; Exp_[experiment number]_Sys[system number in the experiment])

Code	$Effort_{within}$	Relevant information content (Expectancy * Value)	$Value_{(hedonic)}$	**$P(A_{Speech})$**	**PDT Overall**
Exp_3_Sys_1	Low (+)	Low (0)	High (+)	**(++)**	**High**
Exp_3_Sys_2	High (0)	High (+)	High (+)	**(++)**	**High**
Exp_5_Sys_3	Low (+)	High (+)	Low (0)	**(++)**	**High**
Exp_3_Sys_3	Low (+)	High (+)	High (+)	**(+++)**	**Medium**
Exp_5_Sys_2	Low (+)	High (+)	High (+)	**(+++)**	**Medium**
Exp_5_Sys_1	Low (+)	High (+)	Very High (++)	**(++++)**	**Low**

Rankings in Table 24 were theoretical assumptions regarding the probability of attention allocation towards the display depending on the design recommendations (see Table 23). Figure 44 and Table 25 show the actual observed PDT Display for the six system versions made in experiments 3 and 5 of this thesis.

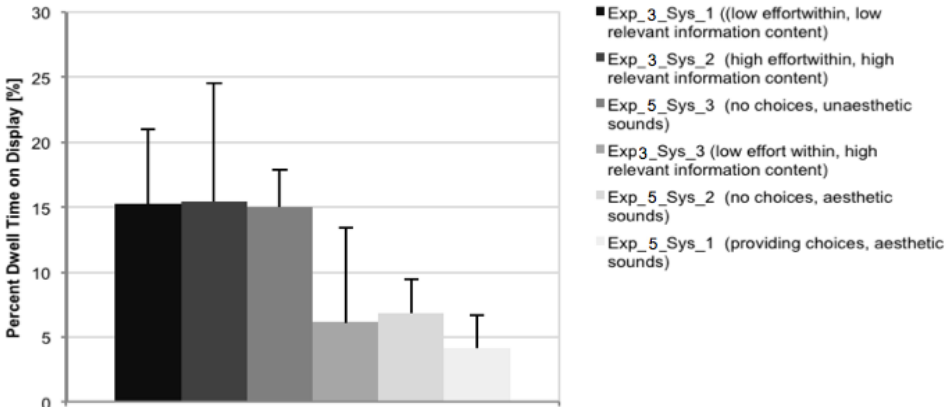


Figure 44. Means and standard deviations of PDT Display for the six system versions (sorted by the theoretically expected PDT Display based on the theoretical assumptions)

Table 25. Means and standard deviations on percent dwell time on display. Sorted by the theoretically assumed probability to allocate attention towards the display. Code represents the system versions and the experiment in which the system was tested (Exp=Experiment, Sys=System)

Rank		Mean	Standard Deviation
1	Exp_3_Sys_1 ((low effortwithin, low relevant information content)	15.30	5.72
2	Exp_3_Sys_2 (high effortwithin, high relevant information content)	15.42	9.15
3	Exp_5_Sys_3 (no choices, unaesthetic sounds)	15.03	2.84
4	Exp_3_Sys_3 (low effort within, high relevant information content)	6.13	7.29
5	Exp_5_Sys_2 (no choices, aesthetic sounds)	6.85	2.63
6	Exp_5_Sys_1 (providing choices, aesthetic sounds)	4.17	2.54

Combining the results of the two studies gives an overview of the percent dwell time on display means for the six system versions. Based on the characteristics in the defined parameters of the revised SEEV Model, the overall tendency to look towards the display was calculated. The system using all design recommendations (see Table 23) achieved the lowest PDT Display. The speech output design recommendations deducted from the elaborated SEEV Model reduced the PDT Display from 15% to 4% (15% for a common speech output design). The experiments conducted indicate that the parameters outlined as predictors of attention allocation offer dynamic possibilities for enhancing speech system output

information compared to the commonly utilized designs of speech prompts for in-car infotainment systems.

To sum up, the following insights were gained in this thesis:

- Theoretical Analysis of SEEV Model parameters and specification of particular parameters to enable comparison of the two-modality outputs (visual and speech).
- Identification of disadvantages for speech on specific parameters that potentially increase attention allocation towards the display of in-car infotainment systems.
- It is proven that improving the characteristics of speech compared to visual output on the identified parameters decreases attention allocation towards the visual output. Additionally, evidence was given that the investigated SEEV Model parameters are also applicable to the auditory modality and influence attention allocation in the predicted direction. For the parameter effort of the SEEV model it was proven that time effort to perceive information inhibits attention allocation. Previously, only location switching costs had been investigated for the parameter effort.
- It is possible to apply the insights gained in the first two experiments to in-car infotainment applications: Deducing design recommendations based on the SEEV Model to improve speech output of in-car infotainment applications to decrease attention allocation to the display.
- It is proven that these design recommendations lead to lower attention allocation to the display and it was evaluated in which way the particular design recommendations to decrease time effort influence the users' mental model of the system.
- It is possible to develop design recommendation which only increase hedonic aspects of speech output.
- It is proven that hedonic aspects also influence attention allocation and therefore lead to lower glancing at the display of infotainment systems.

6.1 Future Work

In addition to the recommendations for further investigations (e.g., investigating the effect of availability on attention allocation) made in the discussion of each experiment, it is suggested to conduct more research on the following topics:

Investigating the Effect of Trustworthiness and Uncertainty of Output Information on Attention Allocation. Kun et al. (2007) found that a poor recognition rate by the system negatively influences driving performance. Systems with low recognition rate could result in more time spent glancing at the display. Next to that, an increased length of time between the user making an input and the system's reply (e.g., due to long system processing times) could lead to more uncertainty. Users expect a system output, and if nothing happens, more checking (i.e., attention allocation) will likely occur. It is suggested that uncertainty corresponds to the parameter expectancy.

Investigating the Effect of a Decreased Salience on Auditory Attention Allocation Due to the Specific In-Car Environment. As was stated in Chapter 3.2, auditory stimuli are more salient than visual stimuli, if the loudness is not intentionally lowered since they are omnipresent. However, the noisy car environment could decrease salience and thus lead towards not understanding the auditory output information. It is assumed that this would result in spending more time glancing at the display.

Investigating Whether Providing More Indirect Information by the Speech Output Could Support the Building of an Adequate Drivers' Mental Model. As mentioned in Chapter 4.2.1, the GUI provides much more indirect information compared to the speech output which supports the development of users' mental models (Hasebrook, 1995). It would be of interest whether spatial presentation of information or how information is grouped in terms of pauses could be an equivalent for the acoustical modality to transport the indirect information without increasing the time effort. In addition, pictures are commonly used for GUIs, while for VUIs mostly speech is utilized. Commercial psychologist Kroeber-Riel (1993) dealt with the research question as to which modalities are more effective at fast communication and concluded that "pictures are fast shots in the brain" (p.53). He found that non-verbal stimuli are considerably faster than words at being transferred. A complex picture can be perceived in one to two seconds, while we could only read five to seven words of simple text in the same period of time. As such, explaining a picture by use of words is indeed complex and time

consuming. Thus, it would be of interest to develop a low time-consuming acoustical output of pictures presented in the GUI. One way would be by the use of auditory icons (see Chapter 4.2.2). Another option would be to define what information is actually relevant in a picture (analogue presentation) for task solving and needs to be presented verbally. For comparing the different modalities (verbal and analogue) such as pictures and speech, symbols can be equated to semantic concepts or propositions. Propositions are fundamental units of meaning composed of elements, arguments and predicates as well as they are linked by relations (Bock, 1982). Propositions allow an abstract representation of verbal and analogue information in a consistent code. Pylyshyn (1981) describes this as a kind of 'interlingua', i.e., non-modal specific symbols which communicate without the use of language (Ballstaedt, Molitor & Mandl, 1989), for example, the spoken word 'okay' is a semantic concept which corresponds to the semantic concept of a picture showing a checkmark. This is a consistent form of representation and in particular has again and again been characterized by opponents of the dual-coding theory (Anderson & Bower, 1973; Pylyshyn, 1981). The concept of propositions or semantic concepts allow the comparison of whether the relevant information content in the picture is also presented by the speech prompt.

Investigating if the Results Gained for Speech Output of Menu-Based/Command-Based Systems is Applicable to Speech Dialogue Systems with Natural Language Recognition. For the menu-based systems described in the present thesis, the user was performing every menu step that he or she also needed to perform to go through the menu via the GUI. So far, only the time effort to perceive and process relevant information for each interaction step (speech prompt) has been investigated. For speech dialogue systems the interaction "can be faster if users immediately can say what they want to achieve without going through the menus or hierarchical pages that are used in GUIs" (Gustafson, 2002, p.7). Short cuts can also be used for menu-based systems, which results in less dialogue turns to achieve the task goal. Therefore also the overall $effort_{within}$ until achieving the task goal and its influence on attention allocation should be investigated in the future. Indeed, the frequency of refreshes as postulated by Wickens et al. (2003) should also be more closely analysed. I assume that only if (a) refreshes of an information source transport high relevance of information for the achievement of specific tasks (relevant information content), (b) more tasks of a specific relevance can be facilitated, and finally, (c) the overall time effort to perceive these tasks is at the same time not increased, then the probability of attention allocation towards that information source will increase.

Another aspect that could be of interest regarding the insights gained for menu-based systems and their transferability to spoken dialogue systems relates to whether using keyword-based speech output to decrease the time $effort_{within}$ for each turn for spoken dialogue systems with natural language detection could result into a decrease of hedonic quality. A natural language input resulting in a telegraphic system output style may negatively influence the perceived naturalness of the conversation. In this thesis, support was found for the theory that hedonic quality also influences attention allocation. Of particular interest would be which parameter has the stronger influence on attention allocation ($effort_{within}$ or $value_{hedonic}$).

7. References

Ackermann, C. & Libossek, M. (2006). System- versus user-initiative dialog strategy for driver information systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP) (pp. 457-460).* Pittsburgh, PA.

Anderson, J.R. & Bower, G.H. (1973). *Human associative memory*. Washington, DC: V.H. Winston & Sons.

Atkinson, R. & Juola, J. (1974). Search and decision processes in recognition memory. In D. Krantz, R. Atkinson, R. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (pp. 242–293). San Francisco, CA: Freeman.

Baber, C. & Noyes, J. M. (1993) Speech control. In K. Baumann und B. Thomas (Eds.) *User Interfaces Design for Electronic Appliances* (pp. 190-208) London: Taylor and Francis.

Baddeley, A. (1986). *Working Memory*. Oxford, UK: Oxford University Press.

Baddeley, A., Thomson, N. & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior, 14*, 575–589.

Balentine, B. & Morgan, D. P. (1999). *How to build a speech recognition application*. San Ramon, CA: Enterprise Integration Group.

Ballard, D.H., Hayhoe, M. M. & Pelz, J.B (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7* (1), 66–80.

Ballstaedt, St.-P., Molitor. S. & Mandl, H. (1989). Wissen aus Text und Bild. In J. Groebel & P. Winterhoff-Spurk (Eds.), *Empirische Medienpsychologie* (pp. 105-133). München: Psychologie Verlags Union.

Bauer, W. (1995). Multimedia in der Schule. In Issing, L. & Klimsa, P. (Eds.*), Information und Lernen mit Multimedia und Internet: Lehrbuch für Studium und Praxis.* Weinheim: Beltz, Psychologie-Verl.-Union.

Bayly, M., Young, K.L. & Regan, M.A. (2008). Sources of Distraction Inside the Vehicle and their Effects on Driving Performance. In M.A. Regan, J.D. Lee & K.L. Young (Eds.*), Driver Distraction: Theory, Effects and Mitigation* (pp.191-213). Boca Raton, FL: CRC Press.

Bengler, K., Geutner, P., Niedermaier, B. & Steffens, F. (2000). "Eyes free - Hands free" oder "Zeit der Stille". Ein Demonstrator zur multimodalen Bedienung im Automobil. *DGLR-Bericht 2000-02. Multimodale Interaktion im Bereich der Fahrzeug- und Prozessführung* (pp.299-307). Deutsche Gesellschaft für Luft- und Raumfahrt e.V. (DGLR), Bonn.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York, NY: Appleton-Century-Crofts.

Blattner, M., Sumikawa, D. & Greenberg, R. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction, 4*, 11–44.

Blom, J.O. & Monk, A.F. (2003). Theory of personalization of appearance: Why users personalize their PCs and mobile phones. *Human-Computer Interaction, 18* (3), 193-228.

Blythe, M., Overbeeke, C., Monk A. & Wright, P.C. (2003): *Funology: From Usability to Enjoyment.* Dordrecht: Kluwer.

Bock, J. (1982): Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*, 1–47.

Bonebright, T. & Nees, M. (2007). Memory for Auditory Icons and Earcons with Localization Cues. In *Proceedings of the International Conference on Auditory Display (ICAD 2007)* (pp. 419–422). Montreal, Canada, June 26-29, 2007.

Bradley, M. & Lang, P. (1994). Measuring emotions: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry, 25* (1), 49–59.

Bradley, M. & Lang, P. (2000): Affective reactions to acoustic stimuli. *Psychophysiology, 37*, 204–215.

Brewster, S.A. (1998). Using non-speech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction, 5* (3), 224-259.

Brewster, S.A. & Cryer, P.G. (1999). Maximising Screen-Space on Mobile Computing Devices. *In Proceedings of ACM CHI'99* (pp. 224-225). New York, NW: ACM Press.

Brewster, S. (2002): Non-speech auditory output. In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 220-239). Mahwah, NJ: Lawrence Erlbaum Associates.

Broadbent, D.E. (1958). *Perception and communication*. London: Pergamon Press.

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 2011 annual conference on Human factors in computing systems* (pp.473-482). New York, NY: ACM Press.

Bubb, H. (2003): Fahrerassistenz – primär ein Beitrag zum Komfort oder für die Sicherheit? *VDI Nachrichten, 1768*, 25–44.

Burkhardt, F., Eckert, M., Niemann, J., Oberle, F., Scheerbarth, T., Seide, S. & Zhou, J. (2010). A mobile office and entertainment system based on android. *Proceedings of the 21st Konferenz zu Elektronischer Sprachsignalverarbeitung (ESSV)*. Berlin, September 8-10, 2010.

Burmester, M., Graf R., Hellbrück J. & Ansgar, M. (2008): Usability – Der Mensch im Fahrzeug. In A. Meroth & B. Tolg (Eds.), *Infotainmentsysteme im Kraftfahrzeug* (pp. 321-351). Wiesbaden: Vieweg.

Canali, J. & Blight, J. (2009): Automotive HMI: Voice Technology and Touch Screens Have Significant Lead. Retrieved from http://www.strategyanalytics.com/default.aspx?mod=reportabstractviewer&a0=4730.

Chang, J., Lien, A., Lathrop, B. & Hees, H. (2009): Usability Evaluation of a Volkswagen Group In-Vehicle Speech System. In *Proceedings of the first International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (137-144). Essen, September 21-22, 2009.

Cohen, M. H., Giangola, J. P. & Balogh, J. (2004). *Voice User Interface Design*. Boston, MA: Addison-Wesley.

Cook, N. (2006). A Psychophysical Explanation for Why Major Chords are "Bright" and Minor Chords are "Dark". *Proceedings of the 1st International Workshop on Kansei.* Kansei University, Japan. Retrieved from ProQuest Learning: Literature on April 24, 2012.

Cordova, D. & Lepper M. (1996): Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88*, 715–730.

Davis, F. D. (1989). "Perceived usefulness, perceived ease of use, and user acceptance of information technology". *MIS Quarterly, 13* (3), 319–340.

Deci, E.L. (1975). *Intrinsic motivation.* New York: Plenum Press.

Deci, E.L. (1981). *The psychology of self-determination.* Lexington, MA: Heaths.

Deci, E.L. Ryan, R. (1985): *Intrinsic motivation and self-determination in human behaviour*. New York: Plenum Press.

Dietrich, R. & Meltzer, T. v. (Eds.) (2003): Communication in High Risk Environments. In *Linguistische Berichte, special issue 12.* Hamburg: Buske.

Dingler, T., Lindsay J. & Walker B. (2008): Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. *Proceedings of the International Conference on Auditory Display 2008* (ICAD08 pp. 1-6). Paris, June 24-27, 2008.

Donges, E. (1982). Aspekte der Aktiven Sicherheit bei der Führung von Personenkraftwagen. *Automobil-Industrie, 27,* 183–190.

Donges, E. (2009). Fahrerverhaltensmodelle. In H. Winner, S. Hakuli & G. Wolf (Eds.), *Handbuch Fahrerassistenzsysteme*, (pp. 15–23). Wiesbaden: Vieweg+Teubner.

Dragon, L. (2007): Fahrzeugdynamik: Wohin fahren wir? In I. Sievers & V. Schindler (Eds.), *Forschung für das Auto von morgen* (pp. 239–260). Berlin: Springer.

Duarte,C. Carrico, L. (2008). Audio Interfaces for Improved Accessibility. In Pinder, S. (ed.) *Advances in Human Computer Interaction* (pp.121-142). I-Tech Education and Publishing KG Vienna.

Eilers, K., Nachreiner, F. & Hänecke, K. (1986): Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung [Development and evaluation of a scale to assess subjectively perceived effort]. *Zeitschrift für Arbeitswissenschaft, 40,* 215–224.

Eimer, M. (1999). Attending to quadrants and ring-shaped regions: ERP effects of visual attention in different spatial selection tasks. *Psychophysiology, 36*, 491–503.

Endsley, M. (1995): Toward a theory of situation awareness in dynamic systems. *Human Factors, 37* (1), 32–64.

Engström, J., Johansson, E. & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F, 8*, 97-120.

Fastenmeier, W. & Gstalter, H. (1998). Ablenkungseffekte durch neuartige Systeme im Fahrzeug. In H.P. Willumeit & H. Kolrep (Eds.), *Wohin führen Unterstützungssysteme? Entscheidungshilfe und Assistenz in Mensch-Maschine-Systemen* (pp. 70-82). Sinzheim: Pro Universitate Verlag.

Farfan, F., Cuayahuitl, H. & Portilla, A. (2003*): Evaluation dialogue strategies in a spoken dialogue system for e-mail.* Universidad Autonoma de Tlaxcala, Mexico. Retrieved from http://www.dfki.de/~hecu01/publications/hc-iasted-aia2003.pdf.

Foyle, D. & Hooey B. (Eds.) (2008). *Human performance modeling in aviation.* Boca Raton, FL: Taylor & Francis.

Fraser,N. (1997). Assessment of Interactive Systems. In D. Gibbon, R. Moore, and R. Winski (Eds.). Handbook on Standards and Resources for Spoken Language Systems, pp. 564–615, Mouton de Gruyter, D–Berlin.

Freed, M. (2000). Reactive prioritization. In *Proceedings of the 2nd NASA International Workshop on Planning and Scheduling for Space*, San Francisco, CA.

Fricke, N. (2006): Semantic auditory icons as warning signals. In *Proceedings of the International Conference on Auditory Display 2006*, London, UK.

Garzonis, S., Jones, S., Jay, T. & O'Neill, E. (2009). Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability & preference. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009* (pp.1513-1522). Boston, MA.

Gaver, W. (1986). Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction, 2* (2), 167–177.

Gaver, W. 1989. The Sonic Finder. Human-Computer Interaction 4 (1). Elsevier Science.

Gaver, W. (1997). Auditory interfaces. In M. G. Helander, T. K. Landauer and P. V. Prabhu (Eds.). *Handbook of Human-Computer Interaction 2nd ed.* (p. 1003 –1041). Elvesier Science.

Gentner, D. & Stevens, A. (Eds.) (1983). *Mental Models*. Hillsdale, N.J.: Lawrence Erlbaum.

Graham, R., Aldridge L., Carter C. & Lansdown T. (1999). The design of in-car speech recognition interfaces for usability and user acceptance. *Engineering psychology and cognitive ergonomics: Job design, product design and human-computer interaction, 4*, 313-320.

Grandt, M. (Ed.) (2003). *Entscheidungsunterstützung für die Fahrzeug- und Prozessführung*. DGLR-Bericht 2003-04. Bonn: DGLR.

Gray, W. D. & Fu, W.-t. (2004). Soft constraints in interactive behaviour: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28* (3), 359-382.

Gray, W., Sims, C., Fu, W.-T & Schoelles, M. (2006): The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behaviour. *Psychological Review, 113* (3), 461–482.

Green, P. (2000). Dealing with potential distractions from driver information systems. *Paper presented at the Convergence 2000 Conference,* Dearborn, Michigan, October 16–18, 2000.

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L. & Grant, P., Kochhar, D., kozak, K. & Blommer, M. (2003). Evaluation of driver distraction using an event detection paradigm. *Journal of the Transportation Research Board, 1843,* 1–9.

Grimes, T. (1990). Encoding TV news messages into memory. *Journalism Quarterly, 67,* 757- 766.

Guski, R. (1997). Psychological methods for evaluating sound quality and assessing acoustic information. *Acustica united with Acta Acustica 83* (5), 765–774.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems Empirical Studies of Spoken Human–Computer Interaction.* Doctoral Dissertation. Stockholm. 2002.

Halasz, F. & Moran, T. (1983). Mental models and problem-solving in using a calculator. In *Proceedings of CHI'83 human factors in computing systems* (pp. 212-216). New York.

Hamerich (2009). *Sprachbedienung im Automobil: Teilautomatisierte Entwicklung benutzerfreundlicher Dialogsysteme*. Berlin: Springer.

Hancock, P. & Meshkati, N. (Eds.) (1988). *Human Mental Workload*. Amsterdam: Elsevier.

Hart, S. & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier.

Hasebrook, J. (1995). *Multimedia-Psychologie.* Heidelberg, Berlin, Oxford: Spektrum: Akademischer Verlag.

Hassenzahl, M. (2003): The thing and I: understanding the relationship between user and product. In M. Blythe, C. Overbeeke, A. Monk & P.C. Wright (Eds.), *Funology: From Usability to Enjoyment* (pp.31-42). Dordrecht: Kluwer.

Hassenzahl, M. (2008). User Experience (UX). Towards an experiential perspective on product quality. In *IHM '08 Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine* (pp. 11-15). New York, NY: ACM.

Hassenzahl, M., Burmester, M. & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187-196).. Stuttgart, Leibzig: B.G. Teubner.

Havaldar, P & Medioni, G. (2009). Multimedia Systems -Algorithms, Standards, and Industry Practices. *Course Technology, 2009.*

Higgins, E. (2006). Value from hedonic experience and engagement. *Psychological Review, 113*, 439–460.

Hockey, B.A. & Rayner, M. (2005). "Comparison of Grammar Based and Statistical Language Models Trained on the Same Data". In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, PA, July 2005.

Hone, K. & Graham, R. (2000). Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI).*Natural Language Engineering, 6,* 287–303.

Horrey, W.J. & Wickens, C.D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors, 48* (1), 196-205.

Horrey, W.J., Wickens, C.D. & Consalus, K.P. (2006). Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied, 12*, 67-78.

Itti, L., Koch, C. & Niebur, E. (1998): A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20* (11), 1254–1259.

International Telecommunication Union, Genf ITU-T Rec. P.10 (2007) *Vocabulary for Performance and Quality of Service.* International Telecommunication Union, Genf

Jacko, J. & Sears, A. (Eds.) (2002). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Janssen, C. & Brumby, D. (2010): Strategic adaptation to performance objectives in a dual-task setting. *Cognitive Science, 34,* 1548–1560.

Jeon, M., Davison, B., Wilson, J., Nees, M.A. & Walker, B.N. (2009). Enhanced Auditory Menu Cues Improve Dual Task Performance and Preference With In-vehicle Technologies. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2009)* (pp. 91-98). Essen, Germany, September 21-22, 2009.

Jersild, A. (1927). Mental set and shift. *Archives of Psychology 14* (89), 81.

Jeschke, B. (Ed.) (2008). Sprachausgaben von Sprachdialogsystemen im Kfz. In *Proceedings ITG-Fachtagung Sprachkommunikation.* Aachen: VDE Verlag.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Juola, J.F. & Botella, J. (2004) Task- and location-switching effects on visual attention. *Percept Psychophys.,* 66 (8),1303-17.

Kalinli, O. &. Narayanan, S. (2007): A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech. In *Proceeding of InterSpeech 2007* (pp.1941-1944). Antwerp, Belgium, August 27-31, 2007.

Kayser, C., Petkov, C., Lippert, M. & Logothetis, N.K. (2005): Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology 15* (21), 1943–1947.

Klatzky, R.L. (1980). *Human memory: Structures and processes*. New York, NY: W.H. Freeman.

Kozma, R. (1991): Learning with media. *Review of Educational Research, 61,* 179–211.

Krantz, D., Atkinson, R., Luce, R. & Suppes, P. (Eds.) (1974). *Contemporary developments in mathematical psychology*. San Francisco: Freeman.

Kroeber-Riel, W. (1993). *Strategie und Technik der Werbung. Verhaltenswissenschaftliche Ansätze* (4th ed.). Stuttgart: Kohlhammer..

Kun, A., Paek, T. & Medenica, Z. (2007). The effect of speech interface accuracy on driving performance. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)* (pp 1326-1329). Antwerp, Belgium 2007.

Kun, A., Paek, T., Medenica, Z., Memarovic, N. & Palinko, O. (2009): Glancing at Personal Navigation Devices Can Affect Driving: Experimental Results and Design Implications. In *Proceedings of the 1st*

International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI). Essen, September 21-22, 2009.

Kurtgözü, A. (Ed.) (2004). *Proceedings of 2004 International Conference on Design and Emotion*. Ankara, Turkey, Juli 12-14, 2007.

Langer, E. & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology, 34*, 191–198.

Latour, P. L. (1962). Visual threshold during eye movements. *Vision Research, 2*, 261-262.

Lee, J., Caven, B., Haake, S. & Brown, T. (2001): Speech-based interaction with invehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors, 43*, 631–640.

Levy, J. & Pashler, H. (2008). Task prioritization in multitasking during driving: Opportunity to abort a concurrent task does not insulate braking responses from dual-task slowing. *Applied Cognitive Psychology, 22*, 507–525.

Levy, J., Pashler, H. & Boer, E. (2006): Central interference in driving: Is there any stopping the Psychological Refractory Period? *Psychological Science 17* (3), 228–235.

Lippmann, R.P., Braida, L.D. & Durlach, N.I. (1981). A study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss. *Journal of Acoustical Society of America, 69*, 524–534.

Mahlke, S. (2007). *User Experience of Interaction with Technical Systems*. (Doctoral Dissertation). Retrieved from http://opus.kobv.de/tuberlin/volltexte/2008/1783/pdf/mahlke_sascha.pdf.

Mahlke, S. (2008). User experience: usability, aesthetics and emotions in human-technology interaction. In *Position paper for the workshop* ""UX Maifesto", HCI2007, Lancaster, GB.

Manzey, D. (1988). *Determinanten der Aufgabeninterferenz bei Doppeltätigkeiten und ressourcentheoretische Modellvorstellungen in der Kognitiven Psychologie*. Köln: Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt.

Mattes, S. (2003). The Lane Change Task as a Tool for driver Distraction Evaluation. *IHRA-ITS Workshop on Driving Simulator Scenarios*. Dearborn, Michigan, October 2003. Retrieved from www.nrd.nhtsa.dot.gov/IHRA/ITS/MATTES.pdf.

Mattes, S. & Hallen, A. (2009). "Surrogate Distraction Measurement Techniques: The Lane Change Test." In M.A. Regan, J.D. Lee, & K.L. Young (Eds.), *Driver Distraction: Theory, Effects, and Mitigation* (pp. 107-121). Boca Raton, FL: CRC Press.

McAuley, E., Duncan, T. & Tammen, V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60*, 48–58.

Mc Tear, M. (2004). *Spoken Dialogue Technology. Towards the conversational user interface.* London, England: Springer.

Medenica Z. & Kun, A. (Eds.) (2007). "Comparing the Influence of Two User Interfaces for Mobile Radios on Driving Performance". In *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2007*. City Iowa, IA: University of Iowa, Public Policy Center.

Meroth, A. & Tolg, B. (Eds.) (2008). *Infotainmentsysteme im Kraftfahrzeug*. Wiesbaden: Vieweg.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Miller, G. (1969): A psychological method to investigate verbal concepts. *Journal of Mathemalical Psychology, 6*,169-191.

Miller, C., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors, 49*, 57-75.

Möller, S. (2005). *Quality of Telephone-Based Spoken Dialogue Systems.* Springer.

Mugge, R., Schifferstein, H. & Schoormans, J. (2004): Personalizing Product Appearance: The Effect on Product Attachment. In A. Kurtgözü (Ed.), In *Proceedings of 2004 International Conference on Design and Emotion.* Ankara, Turkey.

Niemann, J., Naumann A. & Oberle, F. (2010a). Entwicklung eines nutzerzentrierten Sprachdialogsystems im Fahrzeug. In *Proceedings of USEWARE 2010* (pp. 107–119). Düsseldorf: VDI-Verlag.

Niemann, J., Presse, V., Reissland, J. & Naumann, A. (2010b). Developing a user-centered mobile service interface based on a cognitive model of attention allocation. In *Proceedings of Human Computer Interaction Symposium (HCIS 2010).* Brisbane, Australia: Springer.

Niemann, J., Reissland, J. & Neumann, A. (2010). Mobile Dienste im Fahrzeug: Gestaltung von Sprachausgaben zur Reduzierung visueller Ablenkung. In Ziegler, J. & Schmidt, A. (Eds.), *Mensch & Computer 2010: Interaktive Kulturen* (pp. 301-310). München: Oldenbourg Verlag.

Niemann, J., Bongartz, S. & Naumann, A. (2011a). Steigerung der hedonischen Qualität akustischer Ausgaben für Infotainmentsysteme im Fahrzeug. In Schmid, S., Adenauer, J., Elepfandt, M., Lichtenstein, A. (Eds.). *Proceedings of the 9. Berliner Werkstatt Mensch-Maschine-Systeme. Berliner Werkstatt Mensch-*

Maschine-Systeme, Reflexionen und Visionen der Mensch-Maschine-Interaktion (pp.176-177), Oktober 5/7, Berlin, Düsseldorf: VDI Verlag.

Niemann, J., Schulz, K. & Wechsung, I. (2011b). Effects of Shortening Speech Prompts of In-Car Voice User Interfaces on Users Mental Models. In *Proceedings of Interspeech.* Florence Italy.

Norman, D. (1983): Some observations on mental models. In D. Gentner & A. Stevens (Eds.), *Mental Models* (pp.7-14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1986). Cognitive engineering. In Norman, D. A., & Draper, S. W. (Eds.), *User centered system design: New perspectives on human-computer interaction* (32-65). Hillsdale, NJ: Lawrence Erlbaum Associates.

NHTSA (1997). *NHTSA Releases Final Group of 1997 Frontal Crash Test Result*. Washington, DC: U.S. Department of Transportation. Retrieved from http://www.nhtsa.gov/About+NHTSA/Press+Releases/1997/NHTSA+Releases+Final+Group+of+1997+Frontal+Crash+Test+Results.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33* (1), 17-33.

Ortega, L., Guzman-Martinez, E. Grabowecky, M., & Suzuki, S. (2010). "Auditory dominance in time perception". *Journal of vision, 9* (8).

Paivio (1971). *Imagery and verbal processes*. New York: Holt, Rinehart and Winston.

Paivio, A., Philipchalk, R. & Rowe, E. (1975): Free and serial recall of pictures, sounds, and words. *Memory & Cognition, 3* (6), 586–590.

Palladino, D. & Walker, B.N. (2007). Learning rates for auditory menus enhanced with spearcons versus earcons. In *Proceedings of the International Conference on Auditory Display (ICAD 2007)* (pp. 274-279.*)*, Montreal, Canada, June 26-29. 2007.

Parasuraman, R. & Davies, D. (Eds.) (1984). *Varieties of Attention.* New York, NY: Academic Press.

Pashler, H. (Ed.) (1998). *Attention.* Hove, England: Psychology Press/Erlbaum, Taylor & Francis.

Pashler, H. & Johnston, J. (1998). Attentional limitations in dual-task performance. In H. Pashler (Ed.), *Attention* (pp.155-189). Hove, England: Psychology Press/Erlbaum, Taylor & Francis.

Patterson, R.D. (1982). *Guidelines for auditory warning systems on civil aircraft*. London, England: Civil Aviation Authority.

Pleschka, M., Schulz, S., Ahlers, J., Weiss, B. & Möller, S. (2009). Nadia – a natural language spoken dialog system for automotive infotainment applications. In *Proceding of the. 1st Int. Workshop on Spoken Dialogue Systems Technology (IWSDS 09).*

Posner, M., Nissen, M. & Klein, R. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83* (2), 157–171.

Posner, M. I., Snyder, C. R. R. &Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology, General, 109* (2), 160–174.

Proctor, R. & Proctor, J. (2006). *Sensation and perception*. In Salvendy, G. (ed.), Handbook of Human Factors an Ergonomics (3rd ed.) (pp. 53–57). Hoboken, NJ: John Wiley & Sons.

Pylyshyn, Z.W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review, 88*, 16-45.

Radeborg, K., Briem, V.,& Herdman, L. (1999). The effect of concurrent task difficulty on working memory during simulated driving. *Ergonomics, 42,* (5), 767–777.

Raffaseder, H. (2002). *Audiodesign.* München, Wien: Carl Hanser Verlag.

Rasmussen, J. (1983). Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics, 13,* 257–266.

Rauch, N., Totzke, I. & Krüger, H.-P. (2004). *Kompetenzerwerb für Fahrerinformationssysteme: Bedeutung von Bedienkontext und Menüstruktur. VDI-Berichte Nr. 1864. Integrierte Sicherheit und Fahrerassistenzsysteme*. Düsseldorf: VDI-Verlag.

Reese, S. D. (1984). Visual-verbal redundancy effects on television news learning. *Journal of Broadcasting, 28,* 79-87.

Regan, M.A., Lee, J.D. & Young, K.L. (2009). *Driver Distraction: Theory, Effects and Mitigation.* Boca Raton, FL: CRC Press.

Robinson, C. W. & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development, 75*, 1387-1401.

Rogers, R. & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology, General 124* (2), 207–231.

Roetting, M. (2001). A systematic of eye and gaze movement parameters for ergonomic research and application. In *Proceedings o 11th European Conference on Eye Movements*. Turku, Finland, August 22 - 25, 2001.

Salmen, A. (2002). *Multimodale Menüausgabe im Fahrzeug.* München: Herbert Utz Verlag.

Salvucci, D. D. & Beltowska, J. (2008). Effects of memory rehearsal on driver performance: Experiment and theoretical account. *Human Factors, 50*, 834–844.

Sasse, M. A. (1992). User's mental models of computer systems. In Y. Rogers, A. Rutherford & P.A. Bibby (Eds.), Models in the mind. Theory, perspective and application (pp. 225-239). London: Academic Press.

Scharf, B. (1998): Auditory attention: The psychoacoustical approach. In H. Pashler (Ed.), *Attention* (pp. 75–117). Hove, England: Psychology Press/Erlbaum, Taylor & Francis.

Senders, J. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics HFE-1, 1,* 2–6.

Senders, J. (1980). *Visual scanning processes*. (Unpublished doctoral dissertation), University of Tilburg, he Netherlands.

Sievers, I. & Schindler, V. (2007). *Forschung für das Auto von morgen.* Berlin: Springer.

Shannon, C.E. & Weaver, W. (1949) *The Mathematical Theory of Communication*. Urbana, IL: University Press.

Silberstein, D. & Dietrich, S. (2003): Cockpit Communication under High Cognitive Workload. In R. Dietrich & T. v. Meltzer (Eds.), *Communication in High Risk Environments* (pp. 9–56). Hamburg: Buske.

Steltzer, E.M. & Wickens, C.D. (2006). Pilots strategically compensate for display enlargements in surveillance and flight control tasks. *Human Factors, 48*, 166-18I.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652-654.

Spector, A. & Biederman, I. (1976): Mental set and mental shift revisited. *Journal of Psychology, 89*, 669–679.

Taylor, S.E. (1989). *Positive illusions: Creative self-deception and the healthy mind*. New York: Basic Books.

Thomas, B. & Joy, J. (2001): Saying what Comes Naturally. *Speech Technology Magazine*. ww.speechtek.com/st.mag/march01/naturally.shtml

Thüring, M. & Mahlke, S. (2007). Usability, aesthetics, and emotion in human-technology interaction. *International Journal of Psychology, 42*, 253–264.

Tolman, E.C. (1959). Principles of purposive behaviour. In S. Koch (Ed.), *Psychology: A study of science, volume 2* (pp. 92-157). New York, NY: McGraw-HilTomoko & Rosenfeld.

Tomko S. & Rosenfeld, R. (2004). Shaping spoken input in user-initiative systems. In *Proceedings of The 8th International Conference on Spoken Language Processing (Interspeech 2004 - ICSLP) (*pp. 2825-2828). Seoul: Sunjijn Printing Co.

Totzke, I., Schmidt, G. & Krüger, H.-P. (2003). Mentale Modelle von Menüsystemen - Bedeutung kognitiver Repräsentationen für den Kompetenzerwerb. In M. Grandt (Ed.), *Entscheidungsunterstützung für die Fahrzeug- und Prozessführung* (pp. 133–158). Bonn: Deutsche Gesellschaft für Luft- und Raumfahrt e.V.

Treisman, A. & Gelade, G. (1980): A feature integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Treisman, A. (1985). Preattentive processing in vision. *Computer Visions, Graphics & Image Processing, 31*, 156-177.

Vickers, E. (2010). The Loudness War: Background, Speculation and Recommendations. In *Proceedings of the 129th Audio Engineering Society International Conferenc*. San Francisco, CA, USA. Paper number 8175.

Vilimek, R. (2007). *Gestaltungsaspekte multimodaler Interaktion im Fahrzeug. Ein Beitrag aus ingenieurpsychologischer Perspektive*. (Doctoral Dissertation). Universität Regensburg, Regensburg.

Vollrath, M. & Totzke, I. (2000). In-vehicle communication and driving: an attempt to overcome their interference. In *Driver Distraction Internet Forum sponsored by the United States Department of Transportation*. National Highway Traffic Safety Administration (NHTSA). Available: http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/33.PDF.

Walker, B., Nance A. & Lindsay, J. (2006): Spearcons: speech-based Earcons improve navigation performance in auditory menus. In *Proceedings of the International Conference on Auditory Display 2006* (pp.63-68). London, UK.

Weinschenk, S. & Barker D. (2000). *Designing effective speech interfaces.* New York, NY: John Wiley & Sons.

Wechsung, I. & Naumann, A. (2008). Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. In *Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008* (pp.276-284). Heidelberg: Springer.

Wickens, C. (1984). Processing Resources in Attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of Attention* (pp. 63–97). New York: Academic Press.

Wickens C. (1992). Attention, Time-Sharing, and Workload. In C. Wickens & J.G. Hollands (Eds.), *Engineering Psychology and Human Performance (2nd ed.)* (pp. 364–411). New York, NY: Harper-Collins Publishers Inc.

Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*, 159–177.

Wickens, C., Goh, J., Helleburg, J., Horrey, W. & Talleur, D. (2003). Attentional models of multi-task pilot performance using advanced display technology. *Human Factors, 45,* 360–380.

Wickens, C. & Holland J. (2000). *Engineering Psychology and Human Performance.* New Jersey: Prentice-Hall.

Wickens, C. & Horrey, W.J. (2009). Models of attention, distraction and highway hazard avoidance. In M.A. Regan, J.D. Lee & K.L. Young (Eds.), *Driver Distraction: Theory, Effects and Mitigation* (pp. 57-72). Boca Raton, FL: CRC Press.

Wickens, C. & McCarley, J. (2008). *Applied attention theory*. Boca Raton, FL: CRC Press.

Wickens, C., McCarley J., Alexander A., Thomas L., Ambinder, M. & Zheng, S. (2008). Attention-situation awareness (A-SA) model of pilot error. In D. Foyle & Hooey B. (Eds.), *Human performance modeling in aviation* (pp. 213–239). Boca Raton, FL: Taylor & Francis.

Wierwille, W. & Tijerina, L. (1995): Eine Analyse von Unfallberichten als ein Mittel zur Bestimmung von Problemen, die durch die Verteilung der visuellen Aufmerksamkeit und der visuellen Belastung innerhalb des Fahrzeugs verursacht werden. *Zeitschrift für Verkehrssicherheit, 41*, 164–168.

Wolf, C., Koved, L. & Kunzinger, E. (1995). Ubiquitous Mail: Speech and graphical interfaces to an integrated voice/e-mail mailbox. In K. Nordby, P. Helmersen, D. Gilmore & S. Arnesen (Eds.), *Proceedings of IFIP Interact'95*, (pp. 247-252). Lillehammer, Norway: Chapman & Hall.

Wundt, W. (1902*). Outlines of psychology*. New York: The Macmillan Co.

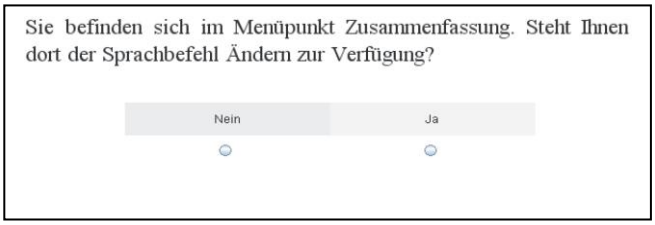
Young, K., Regan M. & Hammer, M. (2003). Driver Distraction: A Review of the Literatur. *Melbourne: Monash University Accident Research Centre*, Report No. 206, p. 2.

Ziegler, J. & Szwillus, G. (Eds.) (2003). *Mensch & Computer 2003: Interaktion in Bewegung*. Stuttgart, Leibzig: B.G. Teubner.

APPENDIX

A.1 Retrieval Task – Shortening of interaction options

Screen Nr. 1



Screen Nr. 2:

Sie befinden sich im Menüpunkt Posteingang. Steht Ihnen dort der Sprachbefehl Suchen zur Verfügung?

Screen Nr. 3:

Sie befinden sich im Menüpunkt Posteingang. Steht Ihnen dort der Sprachbefehl Beantworten zur Verfügung?

Screen Nr. 4:

Sie möchten eine Nachricht aufnehmen. Steht Ihnen dort der Sprachbefehl Aufnahme zur Verfügung?

Screen Nr. 5:

Sie befinden sich im Menüpunkt Posteingang. Stand Ihnen dort der Sprachbefehl Nächste zur Verfügung?

Screen Nr. 6:

Sie befinden sich im Menüpunkt Hauptmenü. Stand Ihnen dort der Sprachbefehl Anrufen zur Verfügung?

Screen Nr. 7:

Sie befinden sich im Menüpunkt Zusammenfassung. Stand Ihnen dort der Sprachbefehl Löschen zur Verfügung?

Screen Nr. 8:

Sie befinden sich im Menüpunkt Posteingang. Stand Ihnen dort der Sprachbefehl Weiter zur Verfügung?

Screen Nr. 9:

Sie befinden sich im Menüpunkt Hauptmenü. Stand Ihnen dort der Sprachbefehl Telefon zur Verfügung?

Screen Nr. 10: Sie befinden sich im Menüpunkt Posteingang. Stand Ihnen dort der Sprachbefehl Abbrechen zur Verfügung?

Screen Nr. 11:

Sie befinden sich im Menüpunkt Zusammenfassung. Stand Ihnen dort der Sprachbefehl Verwerfen zur Verfügung?

A.2 Retrieval Task –

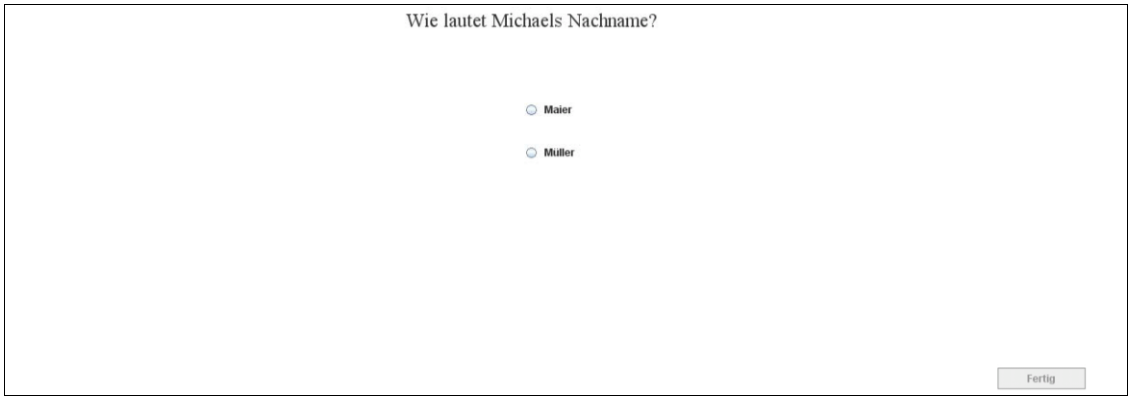
Shortening of content information

A.5.3.1 Presentation of content information

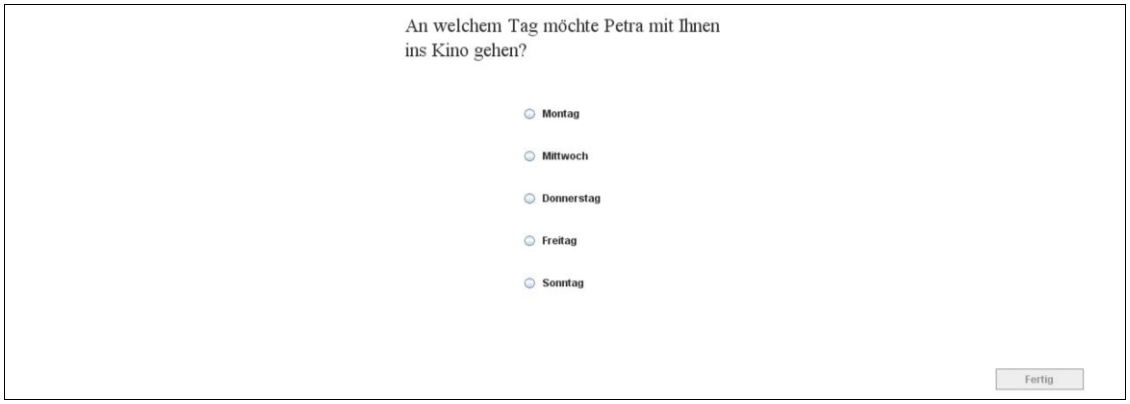
„E-Mail 1 von 5 Ungelesen. Michael Maier. Heute, 9 Uhr 44. Die Vorlesung bei Herrn Müller fällt aus. E-Mail 2 von 5 Ungelesen. Petra Jung. Gestern, 19 Uhr 25. Kino am Freitag. E-Mail 3 von 5 Gelesen. Tim Taler. Gestern, 8 Uhr 12. Treffen um 10 Uhr 30 vor der Sporthalle. E-Mail 4 von 5 Gelesen. Bürgeramt Mitte von Berlin. Fünfter Oktober, 19 Uhr 47. Bestätigung ihrer Anmeldung. E-Mail 5 von 5 Gelesen. Karin Ernst. Erster Oktober, 8 Uhr 12. Urlaubsgrüße von der Nordsee.“

A.5.3.2 Retrieval Tasks

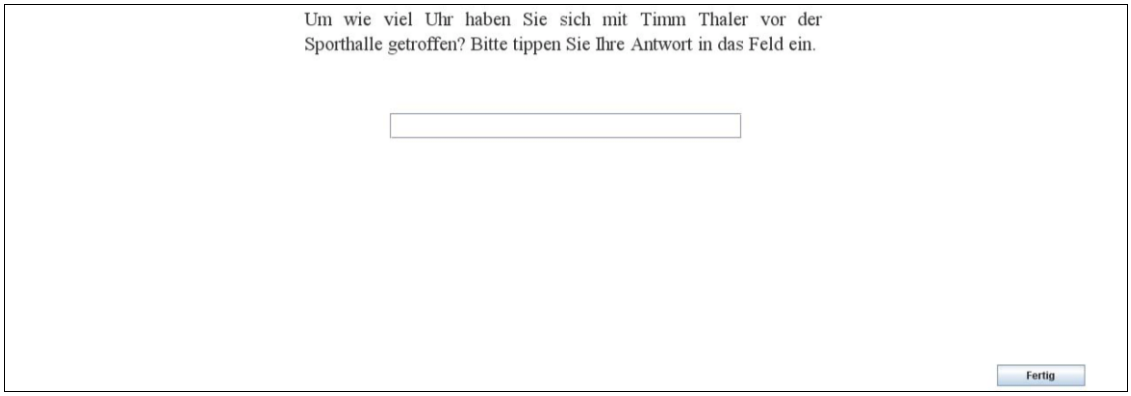
Screen 1:



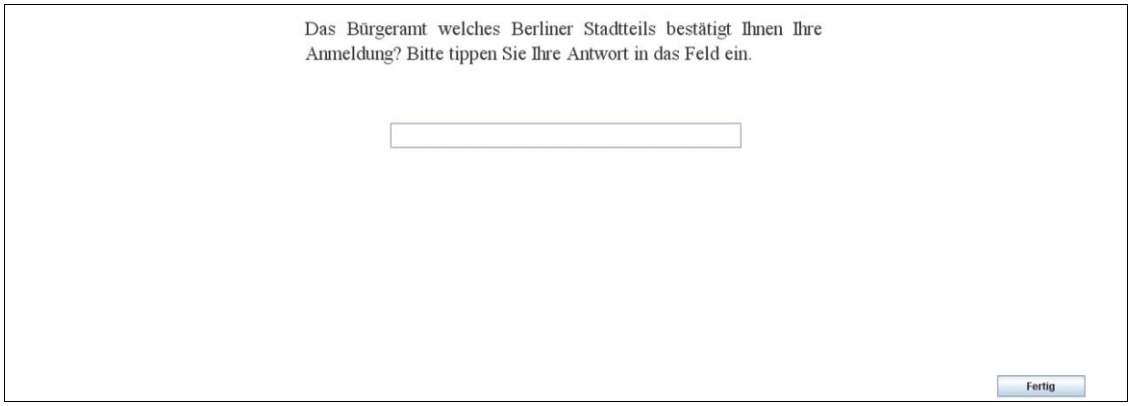
Screen 2:



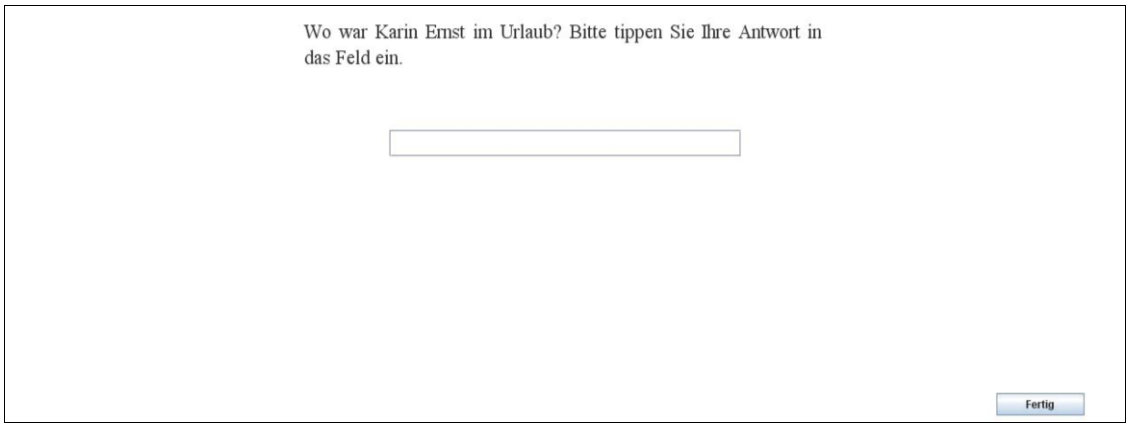
Screen 3:



Screen 4:



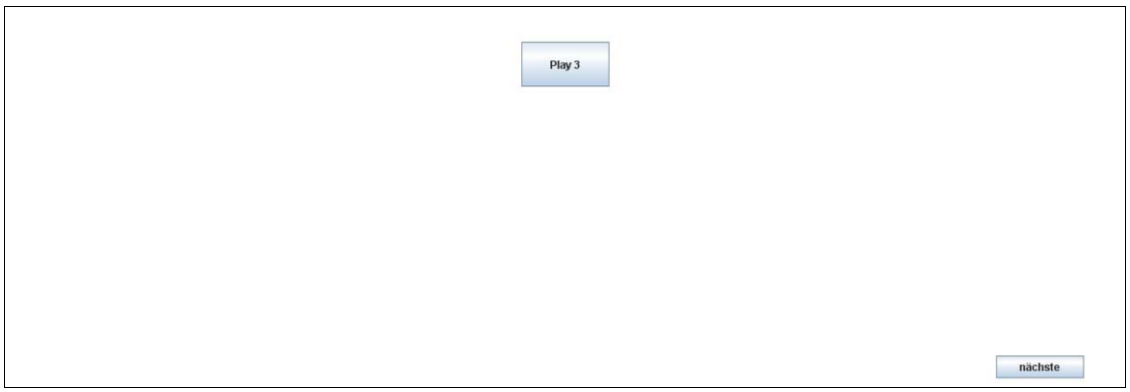
Screen 5:


Wo war Karin Ernst im Urlaub? Bitte tippen Sie Ihre Antwort in das Feld ein.
Fertig

A.3 Navigation and Orientation Task – Shortening of land marking information

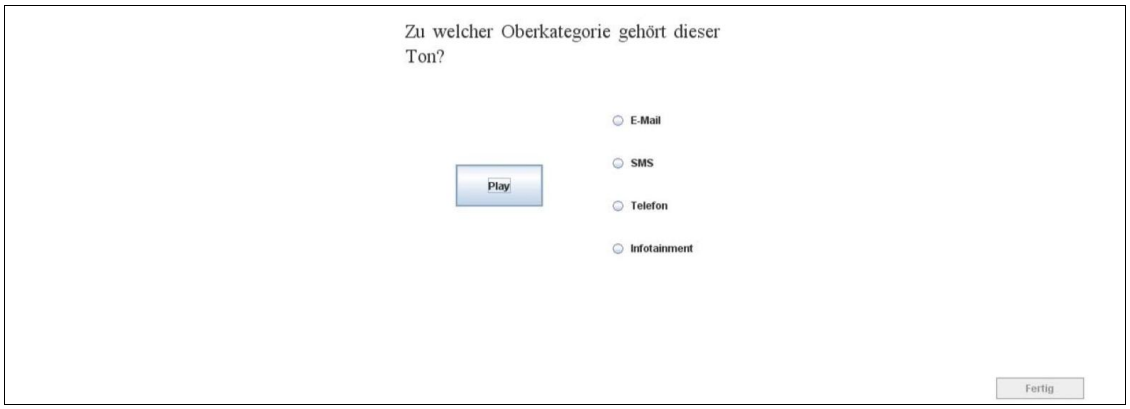
A.3.1 Training phase

Screen Nr. 1

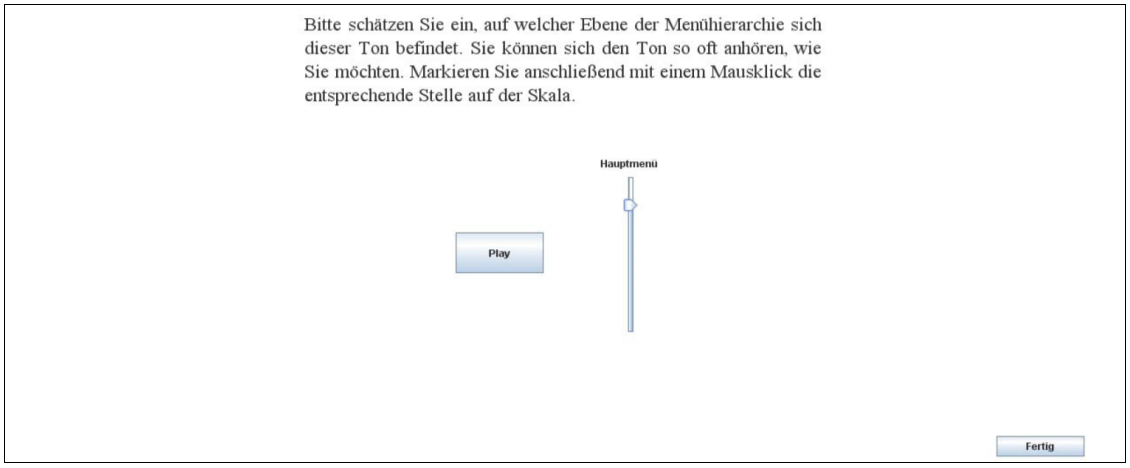


A.3.1 Test phase

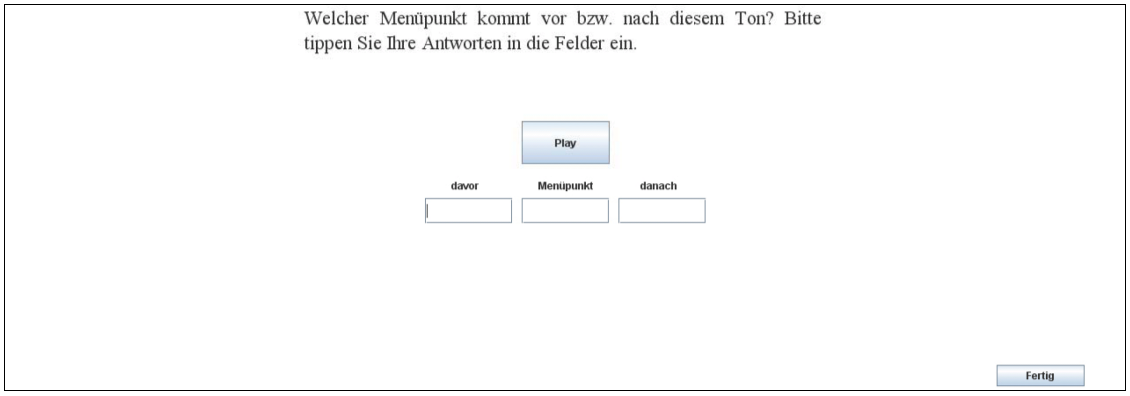
Screen Nr. 2: Defining subordinate category



Screen Nr. 3: Defining menu depth



Screen Nr. 4: Naming the menu steps before and after as well as the actual menu step



Eidesstattliche Versicherung

Hiermit erkläre ich, dass die Dissertation von mir selbstständig angefertigt wurde und alle von mir genutzten Hilfsmittel angegeben wurden.

Ich erkläre, dass die wörtlichen oder dem Sinne nach anderen Veröffentlichungen entnommenen Stellen von mir kenntlich gemacht wurden.