Out-of-Domain Spoken Dialogs in the Car: A WoZ Study

Sven Reichel, Jasmin Sohn,
Ute Ehrlich, André Berton
Speech Dialogue Systems
Daimler AG, Ulm, Germany
sven.reichel@daimler.com

Michael Weber
Institute of Media Informatics
Ulm University
Germany
michael.weber@uni-ulm.de

Abstract

Mobile Internet access via smartphones puts demands on in-car infotainment systems, as more and more drivers like to access the Internet while driving. Spoken dialog systems (SDS) distract drivers less than visual/haptic-based dialog systems. However, in conversational SDSs drivers might speak utterances which are not in the domain of the SDS and thus cannot be understood. In a Wizard of Oz study, we evaluate the effects of out-of-domain utterances on cognitive load, driving performance, and usability. The results show that an SDS which reacts as expected by the driver, is a good approach to control in-car infotainment systems, whereas unexpected SDS reactions might cause severe accidents. We evaluate how a dialog initiative switch, which guides the user and enables him to reach his task goal, performs.

1 Introduction

The acceptance of smartphones is a success story. These devices allow people to access the Internet nearly anywhere at anytime. While driving, using a smartphone is prohibited in many countries as it distracts the driver. Regardless of this prohibition, people use their smartphone and cause severe injuries (National Highway Traffic Safety Administration (NHTSA), 2013). In order to reduce driver distraction, it is necessary to integrate the smartphones functionality safely into in-car infotainment systems. Since hands and eyes are involved in driving, a natural and intuitive speech-based interface increases road safety (Maciej and Vollrath, 2009). There are already infotainment systems with Internet applications like e.g. weather, music streaming, gas prices, news, and restaurant search.

However, conversational spoken dialog systems (SDS) to control all these applications and the car's functionality, are still missing. Current SDSs operate mostly in specific domains and they understand user utterances which are related to these domains. While using natural language, users are not restricted to specific domains. Thus one crucial problem for them is to know which utterances the system is able to understand. People use different approaches to solve this problem, for example by reading the manual, using on-screen help, or relying on their mental model of the SDS. In multi-domain SDSs, utterances can be quite complex and remembering all of them or displaying them on screen would not be possible. As a result, as long as conversational SDSs are not able to operate in much wider domains, sooner or later the user will speak an utterance which is in his mental model of the SDS, but cannot be processed. Such utterances can be divided into out-of-domain and out-of-application-scope (Bohus and Rudnicky, 2005). We induce errors in domain switches and not within one domain, thus only out-of-domain utterances are considered.

In this paper, we present results from a **W**izard **o**f **O**z (WoZ) study on multi-domain interaction with an in-car SDS to evaluate the effects of out-of-domain utterances on driver performance. We considered four different system reactions: successful domain switch, misunderstanding, non-understanding, and a dialog initiative switch. By analyzing them concerning driver distraction and usability, we are able to evaluate whether a dialog initiative switch is an appropriate response to an out-of-domain utterance or not. The results offer valuable clues for the development of multi-domain in-car SDSs.

The remainder is structured as follows: Section 2 provides an overview of studies in this context. Section 3 describes the domain of the study which is shown in Section 4. Data analysis methods are defined in Section 5. We present and discuss the results in Section 6 and conclude in Section 7.

Proceedings of the SIGDIAL 2014 Conference, pages 12–21,
Philadelphia, U.S.A., 18-20 June 2014.

2 Related Work

Driver distractions, due to secondary tasks, are evaluated in many studies (a good overview provides Ei-Wen Lo and Green (2013)). The driver's performance is generally better when using speech interfaces than manual or visual interfaces, however, interacting with an SDS is often worse than just driving (Barón and Green, 2006). Most studies consider specific domains and do not evaluate how to handle domain switches. Kun et al. (2013) evaluated multi-threaded dialogs between humans while driving. By interrupting a dialog, they observed an increase of cognitive load, which affected the driving performance negatively. The participants were prepared that an interruption will be initiated at some time. This means they might be surprised, however, it won't be as unexpected as system reactions in response to out-of-domain utterances. In this work, we evaluate a dialog initiative switch, as a possible reaction to out-of-domain utterances.

In a driving simulator study, Kun et al. (2007) showed that low SDS recognition accuracy affects the steering wheel angle variance negatively. This is first evidence that in-car SDSs need to handle speech recognition or language understanding errors intelligently. In preliminary work to this study, we analyzed a dataset containing dialog errors in relation to driving performance, measured by the lane change task (Mattes, 2003). This showed slight evidence that dialog errors, such as responses to out-of-domain utterances, have an influence on driving performance. However, the lane change task is not the right driving task for such a fine granular analysis, as drivers are only occupied during a lane change and thus not constantly at the same level. Therefore, we analyze driving performance with the **Con**tinuous **T**racking and **Re**action (ConTRe) task (Mahr et al., 2012).

3 User Tasks

In a user experiment it is crucial to set real tasks for users, since artificial tasks will be hard to remember and can reduce their attention. We analyzed current in-car infotainment systems with Internet access and derived eight multi-domain tasks from their functionality (see Table 1). Since only few natural use cases involve more than three domains, every user task is a story of three subtasks. In task number 5 for example, a user has to start a subtask, which navigates him to Berlin. Then he would like to search an Italian restaurant at the destination. Finally, he adds the selected restaurant to his address book.

No	Domain 1	Domain 2	Domain 3
1	POI Search	Restaurant	Call
2	Knowledge	Ski Weather	Navigation
3	Weather	Hotel Search	Address book
4	Play Artist	News Search	Forward by eMail
5	Navigation	Restaurant	Save Address
6	News Search	Play Artist	Share on Facebook
7	News Search	Knowledge	Convert Currency
8	Navigation	Gas Prices	Status Gas Tank

Table 1: Multi-domain user tasks.

At the beginning of a task and during a subtask, the SDS always reacts as it is expected by the users, which means it answers their requests. This increases the stress when the system suddenly starts to react unexpectedly. After presenting the final answer of a subtask, the user has to initiate a domain switch. In response to domain switching utterances four different system reactions were used (see Section 4.2.2).

4 User Experiment

Developing an SDS includes specifying a grammar or training statistical language models for speech recognition. These steps precede any real user test. In system-initiated dialogs, with a few possible utterances, specifying a grammar is feasible. However, in strictly user-initiative dialogs covering multiple domains, this is rather complicated. A WoZ study does not require to develop speech recognition and language understanding as this is performed by a human (Fraser and Gilbert, 1991). In addition, the system reaction is controlled and not influenced by recognition errors. Our study requires such a controlled environment, as an unexpected system reaction, due to a recognition error, would influence the results negatively.

Driver distraction and usability ratings vary among people and depend on age, personality, experience, context, and many more. Therefore, it is essential to conduct a user study with people who might use the SDS later on. A study by the NHTSA (National Highway Traffic Safety Administration (NHTSA), 2013) showed that 73% of the drivers involved in fatal crashes due to cell phone use in 2011, were less than 40 years old. For this reason, our study considers drivers between 18 and 40 years who are technically affine and are likely to buy a car equipped with an infotainment system with Internet access.

4.1 Set-Up of the Experiment

When designing a user interaction experiment, it is important that it takes place in a real environment. As driving on a real road is dangerous, we used a fixed-base driving simulator in a laboratory. A screen in front of the car covers the driver's field of view (see Figure 1). Steering and pedal signals are picked from the car's CAN bus.

It is important that the user assumes he is interacting with a computer as "human-human interactions are not the same as human-computer interactions" (Fraser and Gilbert, 1991). The wizard, a person in charge of the experiment, was located behind the car and mouse clicks or any other interaction of the wizard was not audible in the car. To ensure a consistent behavior of the wizard, we used SUEDE (Klemmer et al., 2000) to define the dialog, which also provides an interface for the wizard. SUEDE defines a dialog in a state machine, in which the system prompts are states and user inputs are edges between them. The content of system prompts was synthesized with NUANCE Vocalizer Expressive[1] version 1.2.1 (Voice: anna.full). During the experiment, the wizard clicks the corresponding edge after each user input and SUEDE plays the next prompt.

Figure 1: Set-up of the experiment

4.2 Design of the Experiment

Driving a car requires the driver to focus on the road and react appropriately to sudden events. However, if drivers are occupied with a secondary task, such as controlling an infotainment system, their attention to the road might suffer. This is due to the fact that the human's performance is reduced when human resources overlap (Wickens, 2008). In this experiment, a dual task scenario is used by driving in a simulator and interacting with an SDS at the same time. There is no visual display in the car, as this would require additional human resources and it would increase the driver distraction (Young and Regan, 2007).

4.2.1 Primary Task: Driving Simulator

One major requirement for the driving simulator is to ensure a controlled and comparable driver distraction measure over all interaction variants and participants. The open-source driving simulator OpenDS provides a driving environment and extensive logging facilities (Math et al., 2012). As explained in Section 2, it is essential to keep the driver occupied at a constant level all the time. Therefore, we used the ConTRe task (Mahr et al., 2012), which consists of a continuous steering task and a reaction task.

Figure 2 shows the ConTRe task with steering cylinders and a traffic light. The yellow steering cylinder moves unpredictably right and left at a constant distance from the driver. The driver has to steer the blue cylinder to superpose it with the middle section of the yellow one. This is similar to driving on a curved road. Sometimes a driver needs to react to sudden events to prevent an accident. A traffic light shows randomly red and green and requires the driver to push the throttle or brake pedal. As the car drives constantly at 50km/h, the pedals are only pushed in response to the traffic light. The movement of the yellow cylinder and the appearance of the traffic light can be controlled by manipulating OpenDS' control variables. We used the "hard driving" condition as described by Mahr et al. (2012).

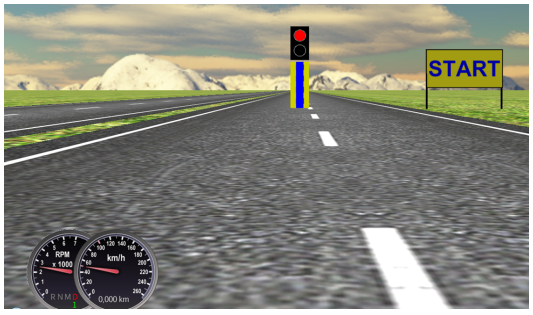

Figure 2: Continuous tracking and reaction task

4.2.2 Secondary Task: Responses to Domain Switching Requests

A task in our experiment consists of three subtasks and each subtask requires two to four semantic concepts. For a user it is possible to insert multiple concepts at once:

U: "Search an Italian restaurant at my destination"

[1] http://www.nuance.com/for-business/mobile-solutions/vocalizer-expressive/index.htm

or as single utterances in a dialog:

U: "Search an Italian restaurant"

S: "Where do you search an Italian restaurant?"

U: "At my destination"

Prompts were created for all possible combinations. SUEDE provides a GUI for the wizard to select which semantic concepts a user input contains. Depending on the selection, either another concept is requested or the answer is provided. Within one subtask, the system always reacts as expected by the user. An answer for the presented example might look like:

S: "There is one Italian restaurant: Pizzeria San Marco."

After this, the user has to initiate a domain switch to save the pizzeria's address into his personal address book. Such user-initiated domain switches challenge current SDSs as language models increase and thus speech recognition as well as language understanding is error prone (Carstensen et al., 2010). Furthermore, the user could request a functionality which is not supported by the system. In case of such a request, SDSs react differently and could apply error recovery strategies if the error is recognized. To analyze the impact of error recovery strategies in the car, we use four different kinds of responses to domain switching requests.

Figure 3 shows the study's conditions. Detailed dialogs that corresponds to them can be found in the Appendix. First of all, we consider the **E**xpected **R**eaction (ER) condition, in which the SDS reacts as expected by the user and switches the domain. As the speech is recognized by a wizard, this is an optimal system without any errors.

Miscommunication can be distinguished between misunderstanding and non-understanding (Skantze, 2007). In the **M**is**U**nderstanding (MU) condition, the SDS does not recognize the domain switch request and it responses in context of the current domain. On the contrary, in the **N**on-**U**nderstanding (NU) condition, it recognizes an out-of-domain utterance and refuses the action by apologizing and encouraging the user to rephrase his utterance (a combination of Bohus and Rudnicky (2005)'s Notify and AskRephrase error handling strategies). The only way to proceed with a MU or NU task in our experiment is to use an explicit domain switching command, such as "start radio application". As we have shown in Reichel et al. (2014), participants do not use such commands naturally in a speech-only infotainment system and only use them after trying numerous unsuccessful utterances. Another approach is a **D**ialog **I**nitiative **S**witch (DIS) to guide the user after recognizing an out-of-domain utterance (Notify and YouCanSay strategy (Bohus and Rudnicky, 2005)). Therefore, the SDS proposes a choice of four possible domains to interact with. Users have to select the first option which was followed by four possible actions within this domain. By selecting the desired action, the SDS reads out four examples of possible utterances. After that, the dialog initiative is given back to the user.

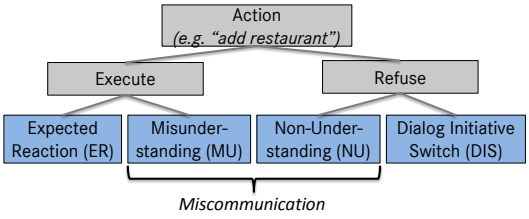

Figure 3: Domain switching response conditions

4.3 Procedure of the experiment

The experiment starts with an initial questionnaire to create a profile of the participant, concerning age, experience with smartphones, infotainment systems and SDSs. Then participants are introduced to the driving task and they have time to practice till being experienced. After completing a baseline drive, they start to use the SDS. For each spoken dialog task users get a story describing in prose what they like to achieve. To minimize priming effects, they have to remember their task and are not allowed to keep the description during the interaction. There is no explanation or example of the SDS, apart from a start command for activation. After the start command, the system plays a beep and the user can say whatever he likes to achieve his task. The exploration phase consists of four tasks, in which the system reacts as it is expected by the user. This enables the user to get used to the SDS while driving. In the second part of the experiment, one task for each condition was completed (ER, MU, NU, and DIS). The conditions were assigned randomly to a task and each one was rated by a **S**ubjective **A**ssessment of **S**peech **S**ystem **I**nterfaces (SASSI) (Hone and Graham, 2000) and **D**river **A**ctivity **L**oad **I**ndex (DALI) (Pauzié et al., 2007) questionnaire. At end of the experiment, each participant completed a second baseline drive without using the SDS to analyze whether the driving performance changed to the first baseline drive or not.

After that, the four conditions were compared in a questionnaire.

5 Evaluation Metrics and Hypotheses

The goal of this study is to evaluate four SDS response conditions concerning driver distraction and usability. Therefore, we used four kinds of measurements (see Table 2): objective driving performance logged by OpenDS, subjective driver distraction with DALI questionnaires, usability scores measured by SASSI questionnaires, and dialog performance. The steering deviation value measures the driver's performance to keep the blue cylinder superposed to the yellow one in the ConTRe task. Reaction times between the appearance of a traffic light and the pedal press are logged as well as wrong and missed pedal presses. The DALI questionnaire consists of 7 questions which are assigned to 7 domains to evaluate the driver's cognitive load. We did not ask for visual or haptic demand, as the system does not have visual output or haptic input. A 7-point Likert scale was used: *low* cognitive load (-3) to *high* cognitive load (+3). SASSI is widely used to measure the usability of an SDS covering 6 dimensions with 34 questions. We used a 7-point Likert scale from *strong disagree* (-3) to *strong agree* (+3). High values mean good usability, except for annoyance and cognitive demand ratings, which are opposed.

objective driving performance (OpenDS)	steering deviation reaction time missed reaction wrong reaction
cognitive load (DALI)	global attention auditory demand interference temporal demand
usability (SASSI)	system response accuracy (SRA) likeability (Like) cognitive demand (Cog Dem) annoyance (Ann) habitability (Hab) speed
dialog performance	task success user response delay system turn duration user turn duration

Table 2: Evaluation metrics

Obviously, we expect that drivers perform best during the baseline drives without controlling the SDS. As ER does not stress or frustrate drivers and they do not need much cognitive power to think what to say, there won't be huge differences between ER and baseline drives. On the contrary, if the system does not react as expected (MU and NU), we expect a worse driving performance and poor usability ratings. NU should be rated better than MU, as the SDS explains the problem. The interesting part is how a DIS will perform as an error handling strategy to out-of-domain utterances. We assume that it is rated better than MU and NU and worse than ER. As the help dialogs in DIS are long, DIS might tend towards MU and NU in terms of driver distraction. However, it will be rated better in terms of usability because the task success is expected to be higher.

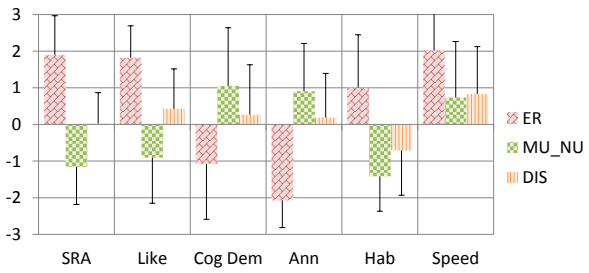

Figure 4: Usability ratings, all of them are significant ($p<.001$) except of: speed between DIS and MU_NU

6 Results

In the following, evaluation results of the four domain switching responses are shown. We analyzed data from 30 participants (16m/14f), with average age of 26.65 (SD: 3.32). Their experience with SDS is little (6-Likert Scale, avg: 3.06, SD: 1.48) as well as the usage of SDSs (5-Likert Scale, avg: 2.04, SD: 1.16). We asked them how they usually approach a new system to learn its interaction schema and scope of operation. All 30 of them try a new application on their smartphone without informing themselves how it is used. Concerning infotainment systems, trying is also the most used learning approach, even while driving (26 people). This means, people do not read a manual, but the system has to be naturally usable. In terms of driving experience, all participants have a driver license for average 8.6 (SD: 3.5) years and most of them use their car daily. Considering the objective driving performances of the two baseline drives, there are no significant differences, which means the participants performed at a constant level over the entire experiment. Figure 4, 5 and 6 show a detailed overview of the evaluation results, which will be explained in this Section.

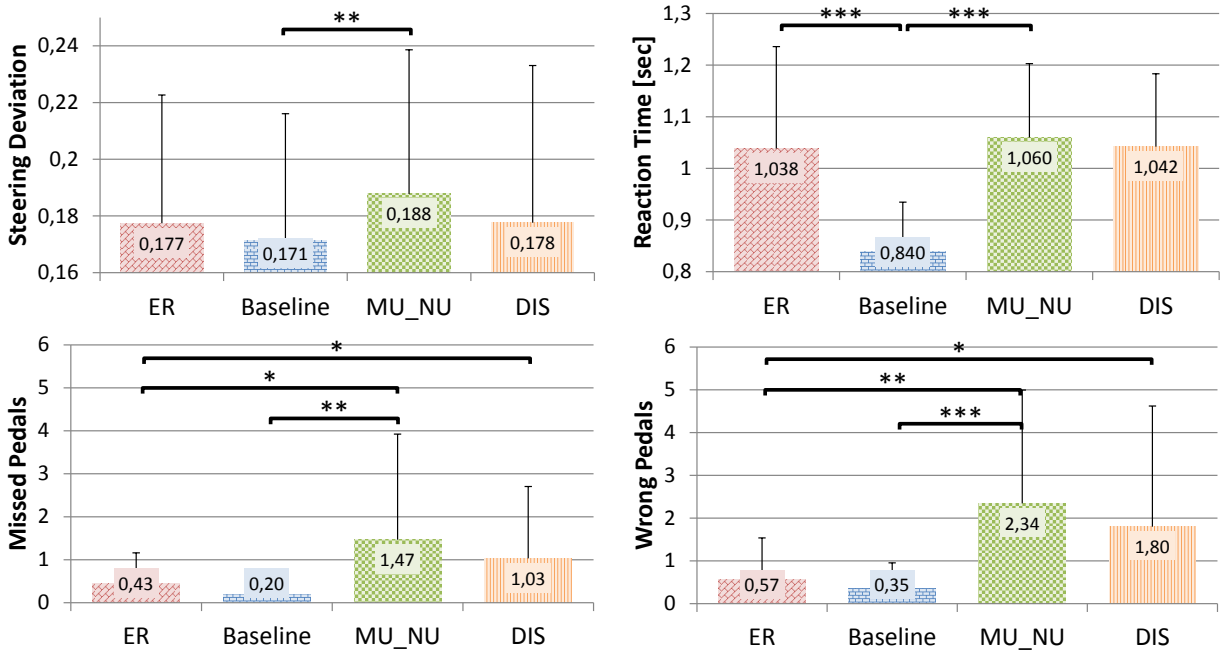

Figure 5: Objective driving performance (OpenDS), significance levels: p<.05(*), p<.01(**), p<.001(***)

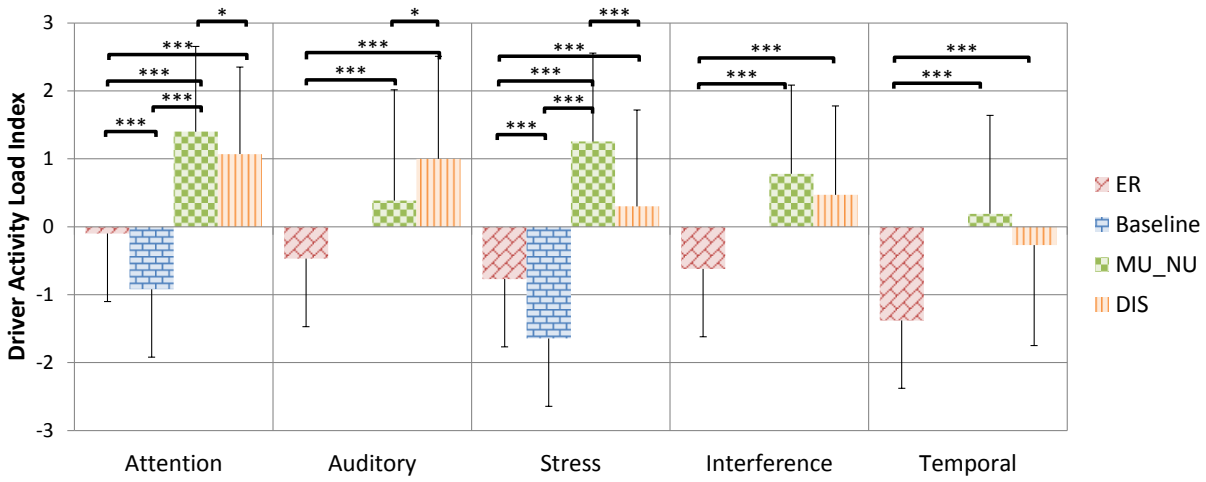

Figure 6: Cognitive load: driver activity load index (DALI), significance levels: p<.05(*), p<.01(**), p<.001(***)

6.1 SDS which Reacts as Expected (ER)

First of all, results of an optimal SDS (ER), which reacts as expected and does not make any mistakes, are presented. The objective driver performance (see Figure 5) is slightly worse than the baseline drives in terms of steering and pressing the right pedals, but not significantly. However, reaction times are worse than without interacting with an SDS. This corresponds to the results from Patten et al. (2004), who observed an increase in reaction times when drivers talk to someone on the phone. The cognitive load (see Figure 6) caused by an optimal SDS is negative in all dimensions, which means an optimal SDS does not put high demands on the driver. In general, ER was rated very good in terms of usability (see Figure 4) and would most likely be accepted by young drivers.

6.2 Mis- and Non-Understanding (MU, NU)

The results of MU and NU do not show significant differences in any dimension. Therefore, the mean value of MU and NU is used. As shown in Figure 6, the driver's cognitive load is high in all dimensions for MU_NU. In terms of stress and attention, it is significantly higher than during baseline drives (other DALI dimensions are not assessed for baseline drives). Due to the increased cognitive load, the driver's performance (see Figure 5) concerning steering, reaction times, and pedal presses decreases significantly compared to baseline drives. Especially the number of times

drivers do not react to external events at all (missed pedal), or they do not react appropriately (wrong pedal), increases strongly. The usability ratings provide evidence how users rate an SDS which is not usable.

As expected, ER performs better than MU_NU. An unexpected system reaction causes higher cognitive load in all dimensions. However, in contrast to what one might expect, the driver's steering performance and reaction times are not better than for ER ($p_{steering}$=.083 and $p_{reaction}$=.215).

6.3 Dialog Initiative Switch as an Out-Of-Domain Handling Strategy (DIS)

Previous Sections have shown that it is important to minimize misunderstandings and non-understandings in a safe and usable in-car infotainment system. Comparing DIS with an optimal and a worst-case SDS shows whether it is a reasonable approach to handle out-of-domain utterances or not. We use a single factor variance analysis (ANOVA) with repeated measurements to identify the best (Helmert contrast) and worst (difference contrast) condition out of ER, DIS, and MU_NU. If DIS lays between ER and MU_NU, we analyze whether DIS tends towards ER or MU_NU. Therefore, we compare the differences of ER-DIS with MU_NU-DIS and use a one sample t-test.

6.3.1 Driving Performance

The ANOVA did not show any significant differences in drivers' steering performances or reaction times (see Figure 5). Using a Helmert contrast to determine the best response, the ANOVA identified ER as the condition with significantly fewest missed and wrong reactions. There is no difference between DIS and MU_NU, thus DIS tends in terms of objective driver distraction more towards MU_NU than to ER.

6.3.2 Cognitive Load

Analyzing the cognitive load of ER, DIS, and MU_NU (see Figure 6), the ANOVA identifies ER as the significant best condition ($p<.002$). The significant worst one in terms of attention, stress, and interference is MU_NU, which means DIS lays in between for these dimensions. However, no evidence is found for stress or interference whether DIS tends towards ER or MU_NU. In global attention, DIS tends slightly ($p<.031$) towards MU_NU. Furthermore, the long prompts in DIS put high auditive demands on the driver.

6.3.3 Usability

As task success of MU_NU dialogs is poor (see Section 6.4), it is obvious that ER is the best ($p<.001$) and MU_NU is the worst condition ($p<.001$) in terms of usability (see Figure 4). All DIS ratings, except of speed, are between ER and MU_NU ($p<.001$). Speed is basically identical to the MU_NU rating, which is due to the long prompts. There is a slight tendency of DIS towards ER in system response accuracy ($p<.051$) and in habitability ($p<.077$), however, this is not significant. In annoyance DIS tends towards MU_NU ($p<.002$), which might be due to the three step help dialog. For cognitive demand and likability, DIS lays exactly between ER and MU_NU.

6.4 Dialog Performance

The task success is pretty low in MU (29.03%) and NU (19.35%) as the task was aborted by the wizard, if drivers did not use explicit domain switching commands after multiple attempts. On the contrary, the task success for ER (96.8%) and DIS (93.6%) is good, however, 3 tasks were aborted by users. Figure 7 shows the average user response delay, system turn duration, and user turn duration. The rectangular bars drawn in line patterns show successful interactions during a subtask and the ones drawn in checked pattern dialogs between two subtasks.

When the system responds as expected, users need between 2 and 3 seconds to respond. If the system does not react as expected (between two subtasks), drivers need significantly more time to respond, as they need to think what to say. In DIS, they only need to repeat the proposed term, thus they respond faster. In MU_NU, the system turns in dialogs between subtasks are shorter, whereby the user turns are longer (user turn duration does not include the user's response time). So either drivers speak slower or provide longer sentences, if the SDS does not react as expected. Due to the four proposed utterances in DIS, system turn durations are longer in dialogs between subtasks.

6.5 Summary and Discussion

In general, if an SDS reacts as expected by the user, it will be a good approach to control the in-car infotainment system. Except for the driver's reaction time, an optimal SDS does not influence the driving performance. However, a delayed reaction of 200ms might be better than glancing at a

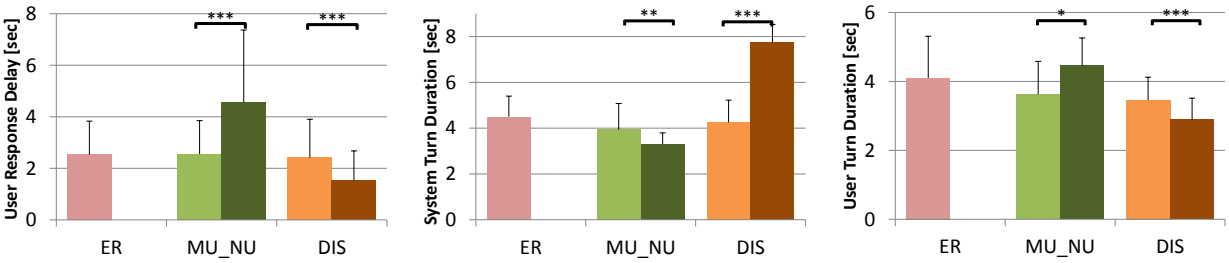

Figure 7: Dialog performance (light color: interaction during subtasks, dark color: dialog between two subtasks), significance levels: p<.05(*), p<.01(**), p<.001(***)

display. For example, the Driver Focus-Telematics Working Group (2006) states in their guidelines to visual distraction: "single glance durations generally should not exceed 2 seconds".

As long as conversational SDSs are not able to operate in much wider domains, sooner or later the user will provide an utterance the system is not able to respond to. Comparing the MU and NU conditions shows that an out-of-domain recognition with a simple rephrase error recovery strategy does not work. This is understandable, as both conditions increase the cognitive load, which influences the driving performance negatively. Especially the reaction to external events, such as traffic lights, suffers. In our experiment, the traffic light was in the middle of the screen. According to Victor et al. (2005), drivers concentrate their gaze on the road center at the expense of peripheral glances during auditory or complex driving tasks. Thus we would expect even worse results if the traffic light occurs in the driver's peripheral vision. This means an intelligent handling strategy for out-of-domain utterances needs to be established, which informs drivers of the system's capabilities.

We evaluated a dialog initiative switch as a response to out-of-domain utterances. Mostly, this strategy performed somewhere between the optimal and worst-case SDS. Due to long narrative system prompts, the auditive demand is rated high by drivers and thus the driving performance tends towards the worst-case SDS. The dialog initiative switch was rated as usable, but different variants need to be developed and evaluated in the future.

After the experiment, the participants rated the four conditions with two questions from ITU-T P.851 (ITU, 2003) on a 7-point Likert scale from *strong disagree* (-3) to *strong agree* (+3):

Q1: "Would you have expected more help from the system?"

Q2: "You feel adequately informed about the system's possibilities?"

	ER (SD)	MU (SD)	NU (SD)	DIS (SD)
Q1	-1.73(1.78)	1.47(1.81)	2.1(1.32)	-1.1(1.58)
Q2	0.43(2.13)	-1.53(1.36)	-1.7(1.49)	0.73(1.66)

Table 3: Adequate system help

Table 3 shows the results, whereby DIS tends towards ER in Q1 (p<.004) and is even better than ER in Q2. This means the drivers felt informed adequately of the SDS, however, further research is necessary to evaluate how to present this information. Shorter helping prompts might be better. Furthermore, multimodal aspects needs to be considered. For example, head-up displays are able to present information, such as possible utterances, right in the driver's view. This might reduce the auditive demand.

7 Conclusions

In this paper, we showed results from a WoZ study on user-initiated multi-domain SDSs in the car. If an in-car SDS cannot fulfill a user's request due to, for example, missing functionality, the driver's cognitive load and distraction will increase. Therefore, out-of-domain utterances need to be identified and handled adequately by in-car SDSs. Switching the dialog initiative is a good approach to guide users to the task goal and reduce their cognitive load. However, if drivers need to process any information, some mental activity will be required. Therefore, the design and implementation of a dialog initiative switch strategy need further efforts to minimize the driver's distraction and to make it enjoyable for the user. Other modalities than speech-only SDSs, such as head-up displays, need to be evaluated in future studies.

Acknowledgments

The work presented here was funded by GetHomeSafe (EU 7th Framework STREP 288667).

References

Adriana Barón and Paul Green. 2006. Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review. Technical report, University of Michigan Transportation Research Institute.

Dan Bohus and Alexander I. Rudnicky. 2005. Sorry, i didnt catch that! an investigation of non-understanding errors and recovery strategies. In *Proceedings of SIGdial*, Lisbon, Portugal.

Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, and Hagen Langer. 2010. *Computerlinguistik und Sprachtechnologie*. Spektrum, Akad. Verl.

Driver Focus-Telematics Working Group. 2006. Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems.

Victor Ei-Wen Lo and Paul A. Green. 2013. Development and evaluation of automotive speech interfaces: Useful information from the human factors and the related literature. *Int. Journal of Vehicular Technology*, 2013:13.

Norman M. Fraser and G.Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language*, 5(1):81 – 99.

Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3&4):287–303.

International Telecommunication Union (ITU). 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems (itu-t rec. p.851).

Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: a wizard of oz prototyping tool for speech user interfaces. In *Proc. of the 13th annual ACM symposium on User interface software and technology*, New York. ACM.

Andrew L. Kun, Tim Paek, and Zeljko Medenica. 2007. The effect of speech interface accuracy on driving performance. In *INTERSPEECH*, pages 1326–1329, Antwerp, Belgium.

Andrew L. Kun, Alexander Shyrokov, and PeterA. Heeman. 2013. Interactions between humanhuman multi-threaded dialogues and driving. *Personal and Ubiquitous Computing*, 17(5):825–834.

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

Angela Mahr, Michael Feld, Mohammad Mehdi Moniri, and Rafael Math. 2012. The contre (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In *Adjunct Proceedings of the 4th AutomotiveUI*, Portsmouth. ACM.

Rafael Math, Angela Mahr, Mohammad M Moniri, and Christian Müller. 2012. Opends: A new open-source driving simulator for research. *Adjunct Proceedings of the 4th AutomotiveUI*.

Stefan Mattes. 2003. The lane-change-task as a tool for driver distraction. In *Proceedings of IGfA*, Dearborn.

National Highway Traffic Safety Administration (NHTSA). 2013. Distracted driving 2011. Technical report.

Christopher J.D Patten, Albert Kircher, Joakim stlund, and Lena Nilsson. 2004. Using mobile telephones: cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, 36(3):341 – 350.

Annie Pauzié, J Manzan, and Nicolas Dapzol. 2007. Driver's behavior and workload assessment for new in-vehicle technologies design. In *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.*, Stevenson, Washington.

Sven Reichel, Ute Ehrlich, André Berton, and Michael Weber. 2014. In-car multi-domain spoken dialogs: A wizard of oz study. In *EACL Workshop Dialog in Motion*, Gothenburg, Sweden.

Gabriel Skantze. 2007. *Error Handling in Spoken Dialogue Systems*. Ph.D. thesis, KTH Computer Science and Communication.

Trent W. Victor, Joanne L. Harbluk, and Johan A. Engstrm. 2005. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):167 – 190. The relationship between distraction and driving performance: towards a test regime for in-vehicle information systems In-vehicle information systems.

Christopher D Wickens. 2008. Multiple resources and mental workload. In *Human factors*, volume 50, pages 449–55. USA.

Kristie Young and Michael Regan. 2007. Driver distraction: A review of the literature. *Distracted Driving*.

Appendix

ER:
U: Navigation
S: Navigation, what do you want?
U: City Berlin
S: To what address or place of interest do you want?
U: Brandenburger Tor
S: Brandenburger Tor in Berlin was set as your destination
U: Italian restaurant near Brandenburger Tor?
S: There is one Italian restaurant in Berlin: Pizzeria San Marco
U: Save the address
S: Alright, the address was added

DIS:
U: I would like to go to the Brandenburger Tor in Berlin
S: Brandenburger Tor in Berlin was set as your destination.
U: Italian restaurant nearby?
S: Action unknown, you can say: restaurant, navigation, hotel, or radio
U: Restaurant
S: You can: search a restaurant, order a table, rate a restaurant, request opening times
U: Search a restaurant
S: Say for example: Search a restaurant at the destination, I search a Chinese restaurant, "Die Krone" in Stuttgart, or Mexican restaurant
U: I search an Italian restaurant
S: Where do you search an Italian restaurant
U: Near Brandenburger Tor in Berlin
S: There is one Italian restaurant in Berlin: Pizzeria San Marco
U: Save the address
S: Action unknown, you can say: address book, restaurant, navigation, or hotel
U: address book
S: You can: add a contact, remove a contact, access contact details, or delete address book
U: Add a contact
S: Say for example: I'd like to add a contact, add an address, add Paul to my address book, save the number 0165332
U: Add San Marco to my address book
S: Alright, the address was added

MU:
U: Navigation to Berlin, Brandenburger Tor
S: Brandenburger Tor in Berlin was set as your destination.
U: Italian restaurant at the destination
S: Italian restaurant could not be set as a destination
U: Navigate to a restaurant in Berlin
S: The address restaurant in Berlin was not found
U: I would like to search a restaurant
S: Sorry?
U: Start restaurant application
S: Restaurant search, what do you want?
U: An Italian restaurant in Berlin
S: There is one Italian restaurant in Berlin: Pizzeria San Marco
U: Add the address to my address book
S: The address of Pizzeria San Marco is Mainstreet 1, Berlin
U: Save that to my address book
S: Sorry?
U: Save Pizzeria San Marco
S: Pizzeria San Marco is famous for its delicious Italian dishes
U: Address book
S: Address book, what do you want?
U: Save the address of San Marco
S: Alright, the address was added

NU:
U: Drive me to the Brandenburger Tor in Berlin
S: Brandenburger Tor in Berlin was set as your destination.
U: Is there an Italian restaurant near Brandenburger Tor?
S: Action unknown, please say something else
U: Give me an Italian restaurant along the route
S: Action unknown, please say something else
U: Italian restaurant
S: Action unknown, please say something else
U: Restaurant in Berlin
S: Action unknown, please say something else
U: I'd like to search a restaurant
S: Action unknown, please say something else
<aborted by the wizard>

Appendix 1: Detailed dialog examples of the four response conditions. Translated from German and based on real experiment data. Red rectangles identify the domain switching dialogs between two subtasks.