Spring 2010

Human-human multi-threaded spoken dialogs in the presence of driving

Oleksandr Shyrokov
University of New Hampshire, Durham

HUMAN-HUMAN MULTI-THREADED SPOKEN DIALOGS IN THE PRESENCE OF DRIVING

BY

OLEKSANDR SHYROKOV

BS, Electrical and Computer Engineering, Odessa State Polytechnic University, 2000

MS, Electrical and Computer Engineering, University of New Hampshire, 2002

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

May, 2010

This thesis has been examined and approved.

Thesis Director, Andrew Kun, Associate Professor, Department of Electrical and Computer Engineering, University of New Hampshire

Peter Heeman, Associate Research Professor, Department of Science and Engineering, Oregon Health and Science University

Rebecca Warner

Rebecca Warner, Professor, Department of Psychology, University of New Hampshire

Richard Messner, Associate Professor, Department of Electrical and Computer Engineering, University of New Hampshire

Thomas Miller, Professor, Department of Electrical and Computer Engineering, University of New Hampshire

William Lenharth, CHFP, Associate Research Professor, Department of Electrical and Computer Engineering, University of New Hampshire

03/01/2010

Date

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Andrew Kun for the time, knowledge, and guidance he provided for this research.

I also would like to thank Dr. Peter Heeman, Dr. Rebecca Warner, Dr. Richard Messner, Dr. Thomas Miller, and Dr. William Lenharth for serving on my dissertation committee and providing me with valuable comments.

I would like to thank everyone at Project54 for helping me in many different ways with all aspects of this research.

Special thanks goes to my loving family, my mother, Elvira, and my sister, Alina, for supporting my decision of starting a PhD program.

Finally, I would like to thank my loving wife, Polina, and her family for their care and support.

This work was funded by the National Science Foundation under grant IIS-0326496 and the US Department of Justice under grant 2006DDBXK099.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

LLG	Last letter word game (interrupting task)
TQG	Twenty questions game (ongoing task)

ABSTRACT

HUMAN-HUMAN MULTI-THREADED SPOKEN DIALOGS IN THE PRESENCE OF DRIVING

by

OLEKSANDR SHYROKOV

University of New Hampshire, May, 2010

The problem addressed in this research is that engineers looking for interface designs do not have enough data about the interaction between multi-threaded dialogs and manual-visual tasks. Our goal was to investigate this interaction. We proposed to analyze how humans handle multi-threaded dialogs while engaged in a manual-visual task. More specifically, we looked at the interaction between performance on two spoken tasks and driving. The novelty of this dissertation is in its focus on the intersection between a manual-visual task and a multi-threaded speech communication between two humans.

We proposed an experiment setup that is suitable for investigating multi-threaded spoken dialogs while subjects are involved in a manual-visual task. In our experiments one participant drove a simulated vehicle while talking with another participant located in a different room. The participants communicated using headphones and microphones. Both participants performed an ongoing task, which was interrupted by an interrupting task. Both tasks, the ongoing task and the interrupting task, were done using speech. We collected corpora of annotated data from our experiments and analyzed the data to verify the suitability of the proposed experiment setup. We found that, as expected, driving and our spoken tasks influenced each other. We also found that the timing of interruption influenced the spoken tasks. Unexpectedly, the data indicate that the ongoing task was more influenced by driving than the interrupting task. On the other hand, the interrupting task influenced driving more than the ongoing task. This suggests that the multiple resource model [1] does not capture the complexity of the interactions between the manual-visual and spoken tasks. We proposed that the perceived urgency or the perceived task difficulty plays a role in how the tasks influence each other.

CHAPTER 1

INTRODUCTION

Driving has a significant social importance. The U.S. Census Bureau reports that Americans spend more than 100 hours a year on the road [2]. At the same time, the number of in-vehicle devices is increasing. As the computational capabilities of in-vehicle devices continue to increase, more and more services and functionalities will be available to drivers. For example, location-based technologies, such as GPS navigation, are gaining widespread popularity with consumers, even though the interaction with these devices may interfere with driving performance [3-5]. For example, setting the destination on the navigational device using a touch screen while driving takes the driver's eyes away from the road and hands from the steering wheel [6]. Dialing a cell phone also takes the driver's attention away from the road [3]. An increasing concern for safety resulted in the acceptance of laws concerning the usage of cell phones while driving [7]. For instance, some states prohibit using a cell phone while driving.

As an attempt to find a better way to control in-car devices while driving, the interaction with the devices is shifting to speech interactions [8]. Progress of spoken language research has already been applied with commercial success to enable hands-free interaction with devices in cars [9]. As a result, for instance, newer models of GPS

navigation systems come equipped with speech input and speech output. Examples of such devices are Garmin Nuvi 855, TomTom GO 920, and Pioneer AVIC-F500BT to name a few. Unfortunately, it is well known that spoken tasks can interfere with driving. Green [3] showed that interactions with cell phones increase the risk of a crash for drivers. Medenica and Kun [10] showed that interaction with police mobile radio negatively influences driving performance. McCarley [4] found that drivers engaged in a conversation do not scan the scene for potential dangers as much as drivers who are not engaged in a conversation. In general, the question of how these new technologies affect drivers, as well as the question of how to integrate these technologies so as to reduce the threat of accidents has not been adequately addressed.

The presence of multiple voice controlled devices in a vehicle gives rise to multi-threaded dialogs. We define a dialog thread as an exchange of information on one particular topic between two parties, either a human and a device, or two humans. If more than one topic or more than two parties are involved in the exchange of information, then multiple dialog threads are present, forming a multi-threaded dialog. Multi-threaded dialogs are natural for humans: we have all been in conversations in which we had to bring up a new topic before finishing the current one, and then go back to the original topic, or in a conversation in which we were interrupted by another person before we could return to discussing the original topic of our conversation.

People are capable of being involved in such dialogs while performing a manual-visual task. Car drivers can talk to passengers or on a cell phone, but engaging in a spoken task could influence the manual-visual task performance. For instance, conversing on a cell phone while driving might increase the risk of a crash [3]. This

interaction between manual-visual tasks and multi-threaded dialogs is a two way interaction. On one hand, the manual-visual tasks could influence the dialog. For example, in our previous work [11], we found that people driving a vehicle answered questions slower as compared to people not engaged in a manual-visual task. On the other hand, different parts of the spoken dialog could influence the manual-visual task performance. For example, conversations might decrease the visual scanning range of a driver [4], which, in turn, may lead to an accident. A better understanding of the processes involved in the interaction between humans and computers in eyes-hands-busy environments is required. This knowledge can help build devices which can efficiently accommodate users engaged in a manual-visual task.

1.1 Problem

The problem that motivates our work is that engineers designing human-computer interfaces do not have enough data about the interaction between multi-threaded dialogs and manual-visual tasks. In order to build a human-computer speech interface that supports multi-threaded dialogs there needs to be a set of conventions for the interface to follow. Human-human conversations can provide us with such a set of conventions. Nass and Brave [12] showed that oftentimes people utilize similar behaviors when interacting with a person and a computer. The authors also showed that human-human interactions may not be the best model for human-computer interactions, because of the differences between human cognition and current computer organization. For example, modern computers can preserve and retrieve information exactly as it was received, but most humans have difficulty remembering exact information, such as long numeric values. Nevertheless, human-human interactions as a model for human-computer

interactions have the advantage of being natural to people. This is a very important factor when the technology must be utilized by a broad range of consumers.

Figure 1.1 illustrates a multi-threaded dialog between two people who are discussing driving directions to a restaurant (thread 1) while driving to that restaurant (manual-visual task). One person is the driver, and the other person is the passenger. At some point in time (point A) they start talking about the food choices in the restaurant (thread 2). Before finishing the discussion about the food choices they switch back to the driving directions (point B), due to a complex intersection ahead. After the intersection is cleared the participants discuss the directions again, in order to make sure that they are still on the right path. This leads them to discuss if they have enough gas to reach the destination (thread 3, point C). They return to discussing directions (point D), because now they need to stop by the gas station. When the passenger attempts to resume the discussion of the food choices (thread 2, point E), the driver asks a few more questions about the directions (thread 1), and thus, the return to thread 2 is not successful. Once the driver is sure about the driving directions, the food discussion continues (thread 2, point F).

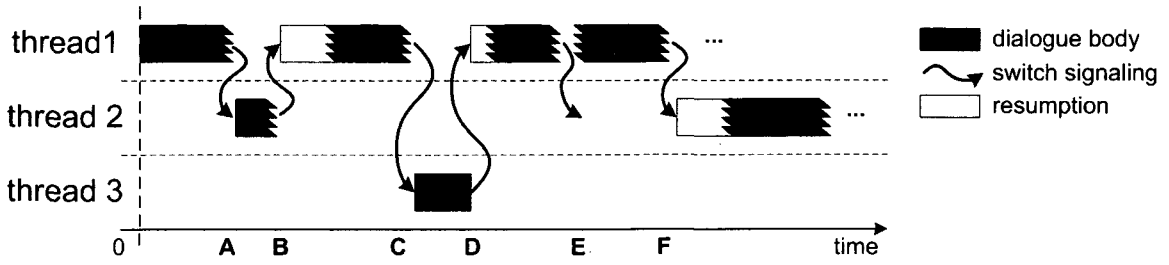


Figure 1.1: Switching between threads.

Participants in the above dialog are changing topics. Hence, they must manage switching topics and resuming previously discussed topics. Switching and resumptions

play a major role in achieving a successful and an efficient multi-threaded communication. Switching and resumptions facilitate maintaining the common ground, which enables the conversation to proceed. Common ground is the knowledge shared between the participants of a dialog. Clark and Brennan [13] show that all collective actions are built on common ground and its accumulation. Switching is the process of signaling a thread change and establishing a new common ground. Resumption is the process of restoring the common ground from a previous dialog thread. There is a substantial body of research on how people signal topic shifts in monologs and dialogs [14,15], such as using prosodic cues and discourse markers [16]. Grosz and Sidner [17] explored the switches in single threaded dialogs. Recently, some research has been done on exploring task switching in multi-tasking dialogs [18]. The novelty of this dissertation is in the focus on the intersection between a manual-visual task and a multi-threaded speech communication as shown as a black area in Figure 1.2.

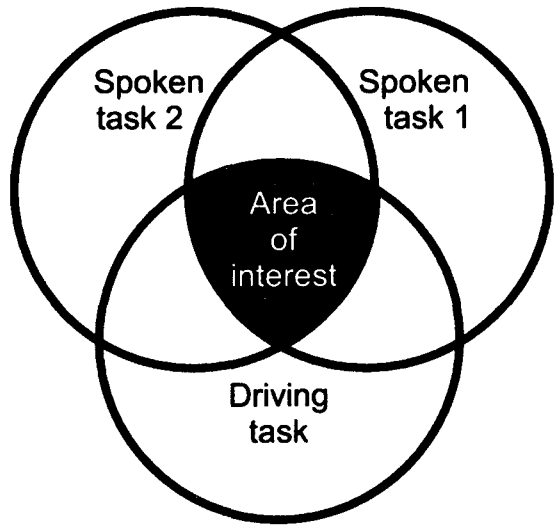


Figure 1.2 Area of interest for this dissertation.

1.2 Goal

Our first goal is to investigate the interaction between multi-threaded dialogs and manual-visual tasks. More specifically, we look at the interaction between the performance on two spoken tasks and driving. Our second goal is to investigate how people manage multi-threaded dialogs when one participant is driving a vehicle. The first goal focuses on the performance on the spoken and manual-visual tasks, while the second goal focuses on the behavioral strategies employed by humans.

Driving is our choice of a manual-visual task for the reason that driving has an important role in our society [2]. It is also common for people to be engaged in a spoken task while driving. There are tools for measuring the driving performance during controlled experiments, such as driving simulators made by DriveSafety [19]. In addition, there is a range of driving tasks, which allow us to control the difficulty of the manual-visual task. For example, it is known that driving on a straight highway with no traffic is easier than driving through complex intersections in a city during rush hours [20]. Finally, it is relatively easy to find competent subjects for the experiments.

We focus on a multi-threaded dialog consisting of one ongoing task and one interrupting task. To achieve our goals we chose spoken tasks which allow us to measure task performance and switching behavior. The ongoing task is organized in the form of question/answer or statement/confirmation pairs. Such discourse structure is common in command and control applications [21,22]. We use definition of an adjacency pair proposed by Schegloff and Sacks [23]. The authors defined question/answer or statement/confirmation as an adjacency pair. One benefit of using question/answer pairs is that there is little ambiguity with the annotation and classification of the dialog

utterances. A single adjacency pair consists of a question or statement by one participant, and an answer or confirmation from the other participant. Multiple adjacency pairs aimed to achieve a particular goal form a dialog thread. The purpose of the interrupting task is to take attention away from the ongoing task. This allows us to observe the behavior subjects exhibit when they switch from the ongoing task to the interrupting task and back.

1.3 Hypotheses

To achieve our first goal we focus on effects of driving on the spoken tasks, effects of the spoken tasks on driving performance, and how the timing of a switch between the tasks affects the spoken tasks. To achieve our second goal we focus on methods people utilize to switch between the spoken tasks and how urgency affects these methods. The following sub-sections describe hypotheses we aim to test in this dissertation. The first three hypotheses address our first goal, and the last two hypotheses address our second goal.

1.3.1 Spoken task performance while driving (hypothesis 1)

We predict that spoken task performance degrades in the presence of driving. Models that are used to estimate response times and memory recalls for single or dual task setups show that task performance degrades with decrease in attention [24]. We expect similar results to be present when attention is captured by driving. It is plausible to see longer response times for drivers than non-drivers. Our hypothesis states that in relation to performance measures in the multi-threaded dialog, the person driving a vehicle will be worse than the person not engaged in a manual-visual task.

We also predict that more demanding driving conditions will negatively influence spoken tasks. More attention must be diverted to the driving in a difficult situation, and, therefore, less attention will be available for the spoken tasks. This might result in a degraded performance on the spoken tasks.

1.3.2 Spoken tasks affect driving performance (hypothesis 2)

We hypothesize that spoken tasks will affect driving performance. Driving and managing a multi-threaded dialog could be too challenging for the driver, which, in turn, could result in degraded driving performance. This hypothesis states that there is a difference in driving performance when comparing the driving performance while the driver is engaged in the primary spoken task with the driving performance while the driver is engaged in the interrupting spoken task. The driver knows that the primary task must be resumed and thus not only an interrupting task must be completed, but the state of the primary task must be remembered. This increased cognitive demand might be noticeable in the driving performance.

1.3.3 Timing of a switch influences spoken tasks (hypothesis 3)

We predict that there is an interaction between the time when a second dialog thread interrupts the first dialog thread and the performance associated with the dialog threads (such as number of utterances, length of pauses, etc.). Yang and Heeman [25] identified two types of context restoration techniques employed by participants in their experiment: utterance restatement and information review. Utterance restatement resumes

the interrupted conversation from the point where it was interrupted by repetition of the last utterance. Information review, on the other hand, provides the critical information that the other speaker might have forgotten. Given that a dialog involves building up a context, we surmise that it will take longer for the participants to restore the context when the switch happens later in the dialog. For example, the dialog shown in Table 1.1 illustrates the build-up of a context.

Code	Speaker	Utterance	Details
U1	Person A	I would like to order an appetizer.	Fact 1
U2	Person B	Okay.	
U3	Person A	I do not want a salad, though.	Fact 2
U4	Person B	No salad then.	
U5	Person A	Fish for an entrée would be nice.	Fact 3
U6	Person B	I see.	
U7	Person A	And I am not sure about the dessert, yet.	Fact 4
U8	Person B	Sounds good.	

Table 1.1 Building up the dialog context.

Table 1.1 shows a dialog of two people discussing a dinner. Person A contributes multiple facts during this dialog in utterances U1, U3, U5, and U7. If this dialog was interrupted after the very first utterance, the participants would only have to remember one fact to continue their conversation. In this case they might utilize utterance restatement. If the dialog was interrupted after the last utterance, then participants would have to remember four facts and the information review could be more appropriate.

We expect to see the change in performance measures for spoken tasks depending on the timing of an interruption. For instance, interruptions introduced later during the ongoing task could decrease performance measures for both tasks.

1.3.4 Switching behavior (hypothesis 4)

Before switching to a different task the participants must agree to switch from the current task to the other task and then resume this other task if it has been already started [25]. How people engage in these behaviors might be influenced by the presence of a manual-visual task. We predict that people will utilize a number of switching behaviors. For example, people might mark the switch from one task to another [18]. The marking can be done using special cue words or prosody [18]. This has a potential to simplify the communication for the participants. Presence of a manual-visual task might cause people to utilize different behavior as compared to people not engaged in a manual-visual task. For instance, we might see that people who are not driving use cue words, while drivers do not, because added workload might cause drivers to simplify their switching behaviors. When switching back to the ongoing task drivers might not provide a summary of the task because they have to deal with driving. On the other hand, the person who is not engaged in a manual-visual task might choose to help the driver by keeping track of the task status for the driver.

1.3.5 Urgency of the interrupting task (hypothesis 5)

We hypothesize that more urgent interrupting tasks will be dealt with more quickly. This implies that subjects might choose different methods when introducing the interrupting task into the ongoing task depending on how quickly the interrupting task must be resolved. For example, if the interrupting task is urgent, subjects might choose to interrupt immediately, independently of who is currently speaking. On the other hand, if

the interrupting task is not urgent, and someone is currently speaking, subjects might wait until the person stopped speaking before introducing the interruption.

1.4 Approach

In order to achieve our goals we created multiple experiments to test our hypotheses (two experiments are described in this document). We experimented with different spoken tasks in order to find ones that proved suitable for our purposes. We chose to use a driving simulator, because it allowed us to have a controlled environment for the experiment. The driving simulator provided measures for the driving performance that are representative of real-life performance [26]. After that we ran the experiments and collected data. Finally we analyzed the data, and presented the results in this document.

1.5 Dissertation organization

Chapter 2 describes the previous research relevant to the stated problem. Chapter 3 describes our first experiment setup with the analysis of the data obtained from this experiment. Chapter 4 describes our final experiment setup. Chapter 5 discusses the results of our final experiment. The conclusion remarks are given in Chapter 6, and Chapter 7 describes the direction for further research.

CHAPTER 2

BACKGROUND

Exploring multi-threaded dialogs during manual-visual tasks presents a new research problem. Psychology, computer science, and human factors researchers address areas related to manual-visual or multi-threaded dialogs. However, most of the research setups in these areas cannot be directly adapted for use in experiments that combine multi-threaded dialogs and a manual-visual task. Nevertheless, the previous research provided us with guidelines to follow.

2.1 Multi-threaded dialogs

Research on multi-threaded dialogs suggests that people keep track of multiple threads. Rosé et al. [27] showed that incorporation of information about multiple threads of the conversation into the discourse structure is more beneficial as opposed to a stack structure of the discourse. The authors proposed an approach which allows having a stack with multiple top elements, corresponding to different dialog threads. In a multi-lingual speech-to-speech computer system, the discourse processor that used this extension performed slightly better than the simple stack discourse processor when analyzing negotiation dialogs. The authors used dialog threads that related to the same topic of

conversation, for example, in discussing which day suits better for a meeting, discussion about Monday is considered one thread, while discussion about Tuesday is another thread. In this dissertation the threads relate to different tasks.

Some work was also done in the area of conversational multi-threading in dialog management by Lemon et al. [28]. The authors used tree-like structures to describe dialog moves and activities, where different branches correspond to different threads. In their later work Lemon et al. [29], extended this concept to improve the robustness of their interfaces. They used thread information for context-sensitive speech recognition and interpretation of corrective fragments. The results suggest that multi-threaded dialogs should not be treated the same way as single threaded dialogs. This serves as a motivation for this research.

Heeman and Fan [30] experimented with an ongoing task in which two participants had to work together to form a poker hand. Participants communicated via headsets with microphones using speech to share information about their cards (the participants could not see each other, which made the communication unimodal). Periodically, one of the participants was prompted to determine whether the other conversant has a certain picture displayed on the screen (interrupting task). The urgency of the interrupting task was an experimental variable and varied between 10, 25, or 40 seconds given to complete the interrupting task. The authors found that this setup elicited both rich collaboration for the card game [25] and interesting task management. Unfortunately, the card playing task cannot be used as an ongoing task for our research, because it requires subjects to see their cards, and for a person involved in a manual-visual task it would create interference with the driving (section 4.1, pg. 49).

2.2 Conversations during meetings

Research on "meetings" mostly focuses on facilitation or retrieving and processing data collected during meetings. The setting of such research is very different from ours, due to the multimodal nature of interactions between participants. In real life if given an option, people use multiple modalities to facilitate multi-threaded dialogs. Given that we would like the driver to keep his eyes on the road, we decided that having a passenger in a car might give subjects an opportunity to use modalities other than speech. The following research indicates that, indeed we need to control what modalities subjects utilize for communication.

Oh et al. [31] showed that gaze direction can be used to determine the intended recipient for an utterance. With their Wizard-of-Oz experiment (subjects were thinking that they interact with a computer system, but it was another person who controlled the responses of the computer system) they showed that "look-to-talk" is a natural alternative to speech indication of the target listener. McCowan et al. [32] presented a framework for computer observation and understanding of interacting people in the meeting context. The authors used a multi-sensor meeting room to collect the data. The processing of the collected data allowed the authors to locate, track, and identify participants, as well as recognize participants' individual actions, such as monologues, discussions, and presentations, to name a few. The research suggested that we need to control the modalities of interactions between participants. Thus we decided to place subjects in different rooms and allow them to communicate using headphones and microphones. This guarantees that speech is the only modality of interaction.

2.3 Switching between non-verbal tasks

According to the dual coding theory [33] humans process visual information and spoken information differently. Therefore, we cannot easily transfer conclusions from experiments with visual tasks into the domain of speech interactions. Still it is possible to utilize techniques, methods, and performance measures from these experiments.

Arroyo et al. [34] used modalities such as heat, smell, sound, vibration and light to signal interruptions. The authors conclude that individual differences control the effect of interrupting stimuli. They argue that it is possible to build an interface that would dynamically select the proper modality for an interruption, based on its effectiveness for a particular person. This research indicates that we might expect to find individual differences between the subjects.

Gillie and Broadbent [35] studied what makes an interruption the most disruptive in the domain of visual tasks. The authors conclude that the time when interruption happened and the length of interruption are less important than the complexity and similarity of the tasks. Hence, in our research we controlled the complexity and similarity of the tasks.

Miyata and Norman [36] gave an overview of psychological theory of human behavior when involved in multiple activities and related it to the design of windows in graphic user interfaces. The authors discussed task-driven and interruption-driven processing. People utilize the interruption-driven processing when they are engaged in one task while expecting to be interrupted at any time. Their behavior in this condition is

different from behavior when there is no expectation of an interruption. In our research we focus on interruption-driven behavior.

Bailey et al. [37] proposed and evaluated a technique for notifying users about new information while they are browsing the World Wide Web. The authors showed that their technique of notification, called "Adjusting Windows" provided the best (of tested techniques) balance of information awareness with intrusion in comparison with background window and a dialog window. Their method was preferred by many of the users over other methods of notification. In subsequent work Adamczyk and Bailey [38] performed experiments to measure the effects of interrupting users at different moments (beginning, middle, end) of task execution. The tasks were document editing and summary writing after watching a video clip. The authors showed that different interruption moments have different impacts on user emotional state. This is an indication that timing of interruptions might affect performance measures of the subjects. This serves as a motivation for our hypothesis 3, which focuses on the timing of interruptions and performance on spoken tasks.

McFarlane [39] discussed the major dimensions of interruption taxonomy. The taxonomy identified the four ways of coordinating user-interruption: immediate, negotiated, mediated, and scheduled. In our domain, an example of an immediate interruption is a blown up tire. The driver must respond immediately to this event. An example of a negotiated interruption is when a passenger asks a driver: "Can I ask you a question?" The driver has an option of choosing the time when and how to answer. An example of a mediated interruption is when a passenger from a back seat asks the front seat passenger to ask the driver something, when it seems that the driver can respond. An

example of a scheduled interruption is a scheduled phone call, so the driver knows that at 1:00pm there will be a phone call for him. This taxonomy can be used to classify how people engaged in manual-visual task chose to coordinate their interruptions.

2.4 Cognitive load

Cognitive load or mental workload is defined as the relationship between the cognitive demands placed on a user by a task and the cognitive resources of that user [1,40]. Higher cognitive load implies that the user has a higher chance of making an error. There are three commonly used ways of estimating cognitive load: physiological (pupil dilation [41,42], heart-rate variability [43], galvanic skin response [44], etc.), subjective (NASA-TLX questionnaire [45,40]), and performance measures. Physiological measures depend on other factors, for example, environmental conditions (temperature, noise), the user's cognitive state (stress [46]), and the user's physical activity. Subjective measures show subjective assessment of the amount of cognitive load experienced by a user. These measures, however, cannot assess rapid changes in cognitive load that might be the result of changes in experimental conditions. Performance measures show how well the user performs a given task. For driving, this can include measures such as variance in lane position and amount of visual attention to the outside world. On the other hand, performance measures might not linearly correspond to the cognitive load, but might only signal when the cognitive load is too high for the user to successfully complete the task. We decided to use performance measures to capture cognitive load, because subjective measures cannot capture changes in experimental conditions over the course of the experiment. We also collect some physiological data, such as pupil dilation, however, the analysis of such data is left for future work.

For the tasks of driving, a number of specialized physiological measures have been used. Recarte and Nunes [47] investigated effects of verbal and spatial-imagery tasks on eye fixations while driving. They found that during a verbal task the visual inspection window shrinks, which means that the driver does not pay as much attention to the road. Spatial-imagery task shrinks that window even more. Horrey et al. [48] examined the impact of in-vehicle task on driver performance and visual scanning. Their experiments accounted for 95% of the variance in scanning using a computational model of visual attention, which indicates increased cognitive load on the driver. This could be used for an indirect measure of the cognitive load of drivers.

Wickens [49] used multiple resource theory to show that it is possible for one task performance to be negatively influenced by other tasks done in parallel. The 4-dimensional multiple resource model described by Wickens [1] gives guidelines for the design of the spoken interaction tasks. The four dimensions of the model are: sensory modalities, codes, channels of visual information, and stages. We would like to separate the manual-visual task from the multi-threaded dialog as much as possible along these dimensions, to localize the interference to particular dimensions. This allows us to better understand the relationship between the manual-visual task and multi-threaded dialogs. This model provided us with the starting point for development of our driving and spoken tasks as described in section 4.1 (pg. 49).

2.5 Task interference

Understanding of how different task interfere with each other will allow engineers to design human-computer interfaces in a way that would minimize this

interference. Modeling how tasks affect cognitive load is a step in this direction. This dissertation aims to provide more information which can be used to improve existing models.

Horrey and Wickens [50] used a computational version of the multiple-resource model to quantify how much demand different in-car tasks have for different resources and how different tasks interfere when using common resources. In their validation study subjects drove a simulator on urban and rural routes of varying complexity while engaging in secondary phone number read-back tasks presented by displays positioned in different locations in the cabin. The secondary task was presented on screens or auditorily. The study showed that the model was able to predict 85% of the variance in performance decrements in secondary task latency and 98% of the variance in response times to critical road hazards. Still, shortcomings of the computational model are that expertise is required to establish conflict values and demand vectors, and the model provides only a relative assessment of task interference between various task combinations. Our research can be used to provide data for establishing conflict values and demand vectors, which are explained in section 4.1, pg. 49.

Strayer et al. [51] showed that listening to radio broadcast or a book on tape did not affect the driving performance as much as a conversation on a cell phone did. They argue that cell phone conversations disrupt performance by diverting attention to an engaging cognitive context other than the one associated with driving. In their later work, Strayer and Johnston [52] showed that performance of a manual visual task was affected by a task that required word generation. The authors suggested that disrupted performance on manual visual task is due to the diverted attention to an engaging

cognitive context other than the one immediately associated with driving, which is not consistent with the multiple-resource model [1]. Our research provides more data about this issue. We also utilize the interrupting task used by Strayer and Johnston in their experiments [52].

2.6 Driving performance

To test our hypotheses 1 and 2, which focus on the interaction between the spoken tasks and driving, we need to track the driving performance. Many researchers have worked on evaluating the visual and cognitive load of driving as well as that of participating in other in-car activities concurrently, such as talking on a cell phone. There is a strong evidence for the interaction between driving task and in-car activities. Driving performance measures can also be used to estimate cognitive load (section 2.4, pg. 17).

In order to help the development of crash countermeasures, Neale et al. [53] collected data about the driving habits, performance, and other factors of 100 drivers over a period of one year. Their study provides useful data on the causes of crashes and near-crashes. For example, the most common cause of crashes was a lead vehicle braking. Green [54] analyzed a large number of studies related to brake reaction times. He pointed out that it was difficult to reconcile results from various sources, since individual studies used different setups, but Green's work has a thorough research overview of the field.

Jamson and Merat [55] used processed steering angle data to measure the driver's fatigue. Their work was based on the research done by MacDonald and Hoffman [56] who investigated a relationship between steering wheel angle and driving task demand. The authors argue that whether the relationship is positive or negative depends

on the level of task difficulty relative to the driver's capacity to cope with it. In short, the driver's capacity must be accounted for in order to use steering wheel angle data. For example, the driver's experience influences the steering wheel angle measurement. This means that standalone steering wheel angle measures may not be directly translated into the driving performance measures. Therefore, we utilize the steering wheel angle measurements along with other driving performance measurements (lane position, distance to leading vehicle, etc.).

Tsimhoni and Green [57] used the visual occlusion method to estimate the visual demand of different road types. They found that visual demand increases significantly with the increase of curve radius. This research suggests that driving on curvy roads should be more difficult than driving on straight roads. We use this information to create two road types with different driving difficulty.

2.6.1 In-car devices

Driving is the choice of manual-visual task for our experiments. Research on in-car devices is tightly coupled with the research of driving performance. The following research confirms that, indeed, multi-tasking in a vehicle can lead to a crash, if multi-tasking is not organized properly.

Green [58] reviewed research concerning effects of in-car devices on driving performance or visual attention. He found that interacting with visual navigational devices causes more frequent lane departures, which is a potential for a hazardous situation. Strayer et al. [59] examined the effects of hands-free cell phone conversations on simulated driving. The authors found that conversations using hands-free cell phone

impaired driver's reaction time to vehicles braking in front of them. This supports our hypothesis 2, which focuses on the effects of spoken tasks on driving performance

Barón and Green [8] summarized the human factors literature on the use of speech interfaces for different in-car tasks, such as music selection, email processing, etc. They conclude that generally driving performance was better when using speech interfaces in comparison with manual interfaces, but using speech interface was often worse than just driving. In a driving simulator experiment, Chisholm et al. [60] looked at manual-visual interactions with mp3 players while driving. They found that complicated interactions with the mp3 player increased reaction time to road hazards. Using an eye gaze tracker, the study also concluded that the complicated interactions redirected driver attention from the road to the mp3 player, increasing the chance of crashes.

Lamble et al. [61] concluded that ability to detect the approach of a decelerating car ahead diminishes as the eccentricity of the visually demanding in car task increases. The eccentricity was defined as the angle subtended at the drivers eye by the arc between the task indicator and the line of sight of the driver straight ahead. The authors found a strong inverse relationship between time-to-collision and the distance from the normal line of sight to the location of a secondary task stimulus. Experiments done by Tsimhoni et al. [62] showed that messages shown on head up display in the locations within five degrees of straight ahead gave the best performance results on the reading task. The latter research tells us that, in order to minimize influence on driving, any visual information presented to the driver must be as close to the center of the screen as possible without obstructing the view of the road.

2.6.2 Simulator

As shown below, high fidelity simulators offer good transfer of training from simulated environments to the real world environments. Slick et al. [26] tested multiple training scenarios on a DS-600c driving simulator. The data indicate that there is no significant difference between training using the simulator and real car for high-risk scenarios. High-risk scenarios used in the experiments were right turn at a stop sign, left turn at a stop sign, right turn at a traffic light with a lane change just prior to the turn, and left turn at a traffic light. These experiments indicate that simulator can be used as a substitute for on-road experiments. Therefore, we decided to utilize the driving simulator in our experiments.

Lew et al. [63] explored how well simulator performance can predict driving performance among participants recovering from traumatic brain injury. In their study, they used driving performance measures from the simulator, such as lane position variance and steering wheel angle variance, in conjunction with human observation data, to predict driving performance at a future date (when participants have hopefully recovered some of their abilities lost due to the injury). They found that driving performance measures were good predictors of future performance, thus justifying the use of driving simulator studies to predict performance in the real-world.

Kemeny and Panerai [64] evaluated perception in driving simulation experiments and concluded that driving simulators can lead to a more thorough understanding of human perception and control of self-motion, especially when speeds and accelerations are higher than in natural locomotion. Mourant and Thattacherry [65] examined whether the severity and type of simulator sickness differs due to the type of

driving environment and driver's gender. They indicate that vehicle velocity might be a factor in driving simulator sickness. Hence, it might be desirable to limit the experiment scenarios to those which do not require high speed of a simulated vehicle. Together, these studies indicate that it is possible to extend the conclusions obtained from the experiments involving a simulator to real life scenarios.

2.7 Dialog management in vehicles

Vollrath [66] investigated the influence of spoken tasks on driving performance by examining a number of different studies. He used the multiple resource model [1] (explained in section 4.1, pg. 49) as the framework to process the data from the studies. He concluded that in order to minimize effects of verbal tasks on driving, the verbal tasks must be simple and short; the quality of the speech and recognition rates must be high; non-verbal aspects of the speech, such as speech volume and rate should be chosen to produce positive evaluation by the drivers. We followed these recommendations during the design of our experiments.

Villing at el. [67] performed human-human multi-threaded dialog experiments in a real car on city roads. The driver and a passenger were given a navigation task and a memory task. The subjects were not restricted on how these tasks had to be accomplished. Video recording of the subjects and the road was taken. The authors found specific Swedish cue phrases that were used for marking topic shifts, similar to "oops", "alright", "let's see". Drivers used these cue phrases only in 17% of the marked topic shifts, while passengers used them only in 12% of the marked topic shifts.

In further research Lindstorm et al. [68] looked at speech disfluency rates as a function of cognitive load. The authors found that under high cognitive load for the driver, the passenger's disfluency rate decreases. This indicates that the passenger makes an attempt to be extra clear and concise when he perceives that the driver is in a difficult situation. The research, by design, utilized multiple modalities for driver-passenger communication, and was focused on natural language features. In contrast, our research is focused on a single modality of interaction between the participants and has more structured tasks.

CHAPTER 3

NAVIGATION EXPERIMENT

This chapter describes our first experiment [11] that used driving in a simulated vehicle as a manual visual task. This experiment setup was inspired by the Map Task experiments [69]. We investigated in which dialog state participants choose to initiate a switch to the interruption dialog thread. This was done to test hypothesis 4, which stated that we expect to see different switching behavior in different situations. We also analyzed how the urgency of the interrupting task affects how subjects initiate interruptions. This was done to test hypothesis 5, which stated that more urgent interruptions should elicit a quicker response.

In this experiment, one conversant was a driver and operated a simulated vehicle, while the other conversant was a dispatcher and helped the driver navigate city streets in order to reach a sequence of destination points. The subjects communicated using headsets with microphones and could not see each other, which made the communication unimodal. The dispatcher knew the required destination points and had a map of the streets. However, the dispatcher did not know that some of the city streets were blocked by construction barrels and, therefore, the driver could not use those streets. This forced the subjects to collaborate and find an alternative route. Periodically, the

driver had to prompt the dispatcher about a message shown on the screen (interrupting task). The prompt for the interrupting task included information about the urgency of this task.

3.1 Preliminary experiments

In experiments we conducted prior to the navigation experiment, subjects interacted with an actual spoken dialog system [70] to complete simple tasks. The tasks included addition problems, circular rotation of number sequences, discovery of short letter sequences, and category-matching word detection. These tasks, however, were not engaging and the resulting dialogs did not exhibit complexity of behaviors. Motivating subjects by telling them they were playing a game and their goal was to solve as many tasks as possible did not help to create an engaging behavior. The navigation experiment used a more engaging and realistic task.

3.2 Hardware setup

This section describes hardware used in the experiment, such as driving simulator, eye-tracker, and audio equipment.

3.2.1 Driving simulator

The experiment involved driving a high fidelity DriveSafety DS-600c simulator [19] shown in Figure 3.1.

Figure 3.1 Driving simulator DriveSafety DS-600c.

The key features of the simulator are:

- Wide field of view (180°);
- Realistic vehicle dynamics (motion, vibration, and sound);
- Simulation system with support of ambient traffic;
- Audio/visual channel computers;
- Scenario creation tools.

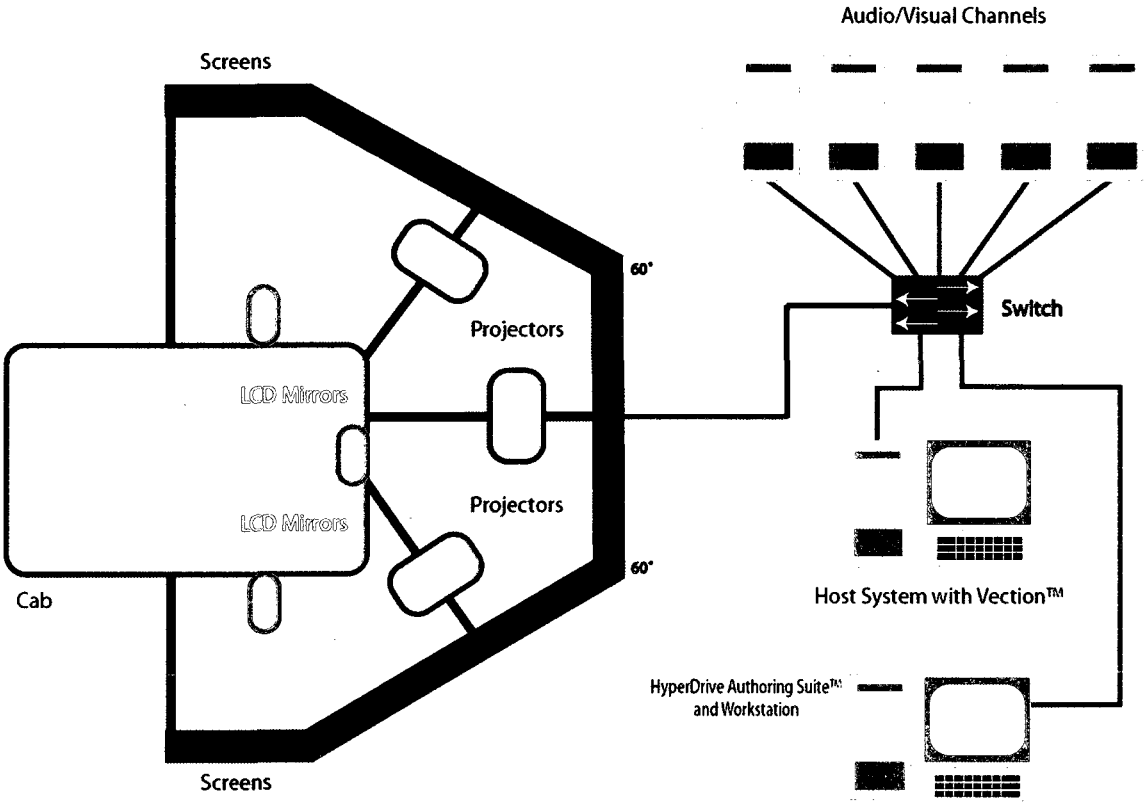


Figure 3.2 DriveSafety DS-600c system overview [19].

The simulation system has three aspheric mirror projectors that produce the 180° field of view. Figure 3.2 shows that the projectors cast the simulation onto three screens. The Ford Focus cabin has a fully functional dashboard with a speedometer and a tachometer. Gas and brake pedals provide haptic feedback. The steering wheel has an electric motor which provides force feedback. A motion platform, sound effects from the simulated environment, and vibrations add to the realism of the simulation. The motion platform simulates pitching movement of the car. Four speakers, located in the front part of the cabin, and two transducers, one under the driver's seat and one in the steering column, simulate car engine vibrations. The same four speakers produce environmental sounds.

The scenario tools allow the design and programming of driving environment scenarios. The scenarios support residential, rural, urban, sub-urban, commercial and industrial environments. Vehicles can be added to be a part of the ambient traffic or they can be programmed to traverse a specific path. Tcl programming language enables developers to add more control to their scenarios.

The DS-600c driving simulator produces standard driving performance measures at 60 Hz frequency. These measures include:

- Lane position, which constitutes the position of the center of the simulated car (measured in meters);
- Steering wheel angle (measured in degrees);
- Vehicle's velocity (measured in meters/second).

These measures will be explained in more detail in section 3.8 (pg.39).

3.2.2 Audio communication and recording

Two people participated in each experiment. Figure 3.3 shows a driver in the driving simulator with headphones and microphone used to communicate with a dispatcher. Figure 3.4 shows the dispatcher wearing headphones. The drivers and the dispatchers were located in separate rooms and could only communicate using headphones and microphones. All communication was recorded synchronously at 44100Hz as mono signals in two separate channels (one channel for the dispatcher and another channel for the driver).

Figure 3.3 Driver in the simulator cabin.

Figure 3.4 Dispatcher in the dispatcher's room.

3.3 Ongoing task

All dispatchers had a map (shown in Figure 3.5) with four marked locations that the drivers had to visit (shown by arrows in Figure 3.6). All drivers started at point 1 and the dispatchers were instructed (Appendix C) to follow the fixed order of points: from 1 to 2, from 2 to 3, from 3 to 4, and from 4 to 1. In order to ensure that the drivers and the dispatchers engaged in a dialog with each other, some city streets were blocked with construction barrels, as shown in Figure 3.7. The barrel locations changed dynamically depending on the driver's location. The drivers had to explain to the dispatchers if a street was closed, so the dispatchers could make corrections to their instructions. The dispatchers had names of points of interest located in the city on their map, for example, gas station and fire station. This allowed the dispatchers to understand where the drivers were on their map. The subjects were instructed to communicate naturally and there was no restriction on how the communication should proceed.

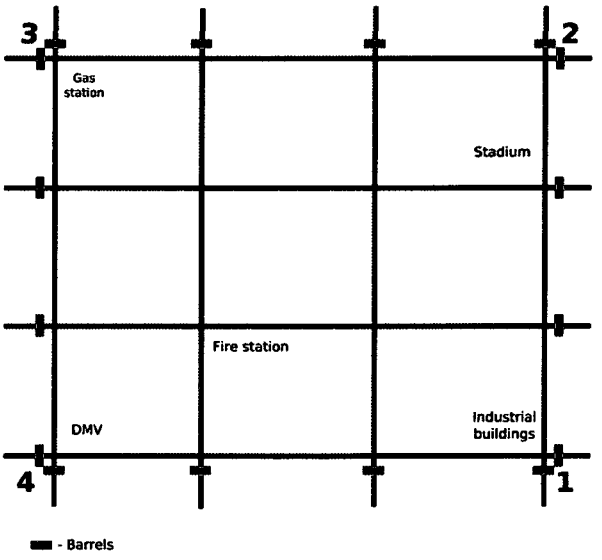


Figure 3.5 Map given to the dispatcher during the experiment.

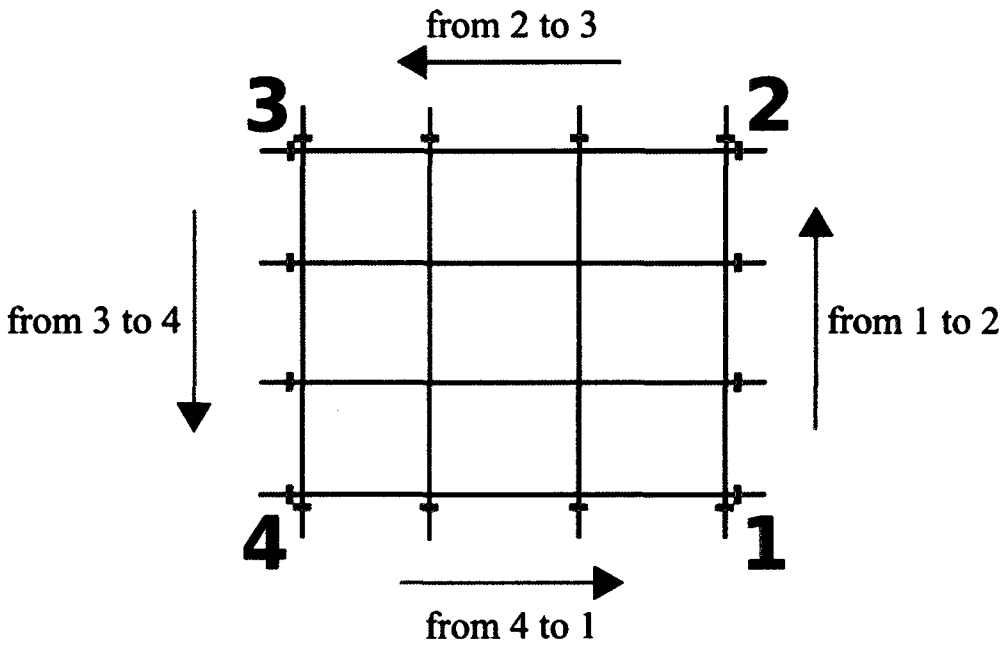


Figure 3.6 Intended travel directions for the driver.

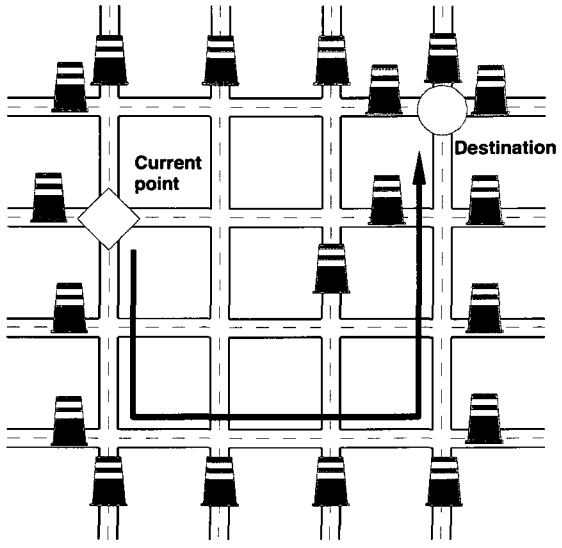


Figure 3.7 Blocked streets and possible path.

3.4 Interrupting task

Periodically the drivers were presented with a visual stimulus. The drivers then had to tell the dispatchers about the visual stimulus. Visual stimuli consisted of a text message with a progress bar, shown in Figure 3.8 and Figure 3.9. We used two different

text messages for the interrupting task: "check engine" and "check link". Each message required a different response from a dispatcher. If a driver told the dispatcher that "check engine" is shown, then the dispatcher had to ask about the speed of the vehicle. When "check link" was shown, the dispatcher had to ask about the distance between the car and the next intersection. Having two different messages ensured that the participants shift their attention from the ongoing task. The drivers had to notice an interruption, shift their attention to the visual stimulus to read what the message states, and then chose the appropriate response. In contrast, if only one kind of a message would be used, then the drivers only had to notice the visual stimulus to initiate the interruption. Time between presentations of visual stimuli was randomly generated and varied from 5 seconds to 40 seconds. The randomly generated sequence was the same for all experiments.

Figure 3.8 Interruption shown to the driver (view from the cabin).

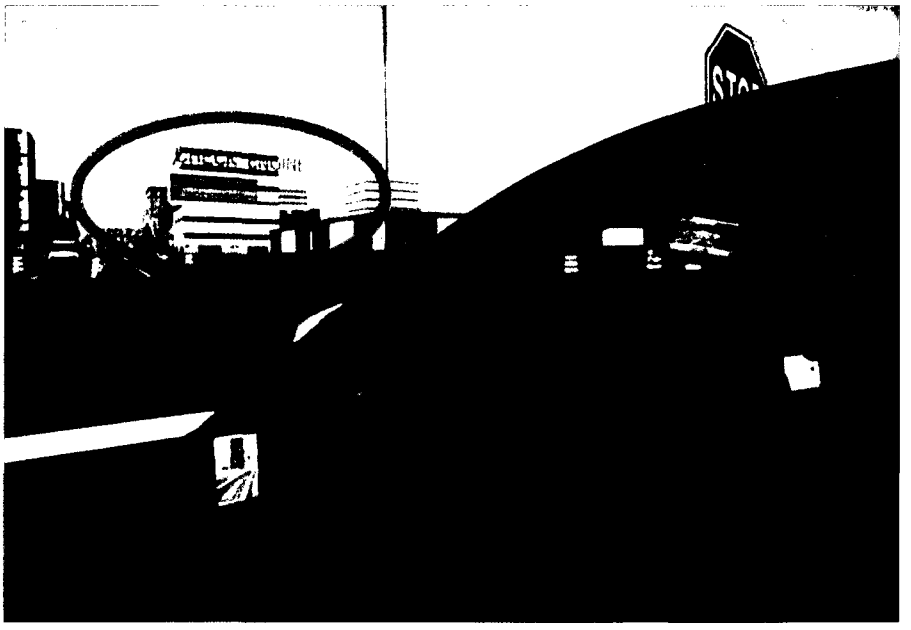


Figure 3.9 Interruption shown to the driver (view from the side).

A progress bar was used to inform the drivers about the urgency of the stimulus. Visual stimuli had one of two urgency levels. The drivers had to respond to *urgent* visual stimuli (47% of all visual stimuli in all experiments) within 10 seconds. For *non-urgent* visual stimuli drivers had 20 seconds to respond. If a driver failed to inform a dispatcher about a visual stimulus within these time limits, the car would stop moving for 10 seconds. These car break-downs were controlled by the experimenter. Participants were told to complete the ongoing task as fast as possible, and car break-downs provided an additional incentive to inform the dispatcher about visual stimuli quickly. Car break-downs slowed down the drivers, which was annoying and most importantly interfered with the instructions to complete the tasks quickly.

3.5 Driving

The driving task was to follow the dispatcher's instructions and drive to four destinations. The simulator presented a city scenario with two-lane roads (a single lane

3.6m wide for each direction). The city consisted of sixteen intersections organized in a four-by-four grid, as shown in Figure 3.7. The limits of the area were marked with construction barrels. The drivers were instructed not to drive past the barrels. Participants were not allowed to travel faster than 30mph (the car would not go faster than 30mph), and they were required to stop at every stop sign, in order to lower the possibility of motion sickness [65]. Every intersection had four-way stop signs. The streets had medium traffic conditions (controlled automatically by the simulation software) and pedestrians walking on the sidewalks and sometimes crossing streets. Traffic and pedestrians were introduced to create a realistic environment for the drivers.

3.6 Independent variables

The ongoing task did not have any independent variables and stayed the same for all subjects. All subjects had to navigate to the same points in the same order. The interrupting task had one independent variable, the urgency of the task, with two levels: urgent or non-urgent. The urgency of the interrupting task was presented in a fixed order for all subjects. The time between presentations of visual stimuli was randomly generated and varied from 5 seconds to 40 seconds. The randomly generated sequence was the same for all the experiments. Due to the difference in driving habits of the drivers and different directions from the dispatchers, the interruptions happened on different streets at different speeds for every driver. This is the reason why we decided not to counterbalance the possible ordering effects.

3.7 Dependent variables for spoken tasks

Figure 3.10 shows a model of the local dialog state of the ongoing task, based on sequences of adjacency pairs [23]. In the first part of an adjacency pair, either the dispatcher or the driver speaks (e.g. poses a question). We denote the first part with "a" when the dispatcher speaks and with "e" when the driver speaks. After a pause (denoted with "b" after the dispatcher speaks and "f" after the driver speaks), the dialog continues with the second part of the adjacency pair. The second part is denoted with "c" when the driver speaks and with "g" when the dispatcher speaks. Finally, when the second part ends, and before the next first part begins, we have a pause in the dialog, denoted with "d."

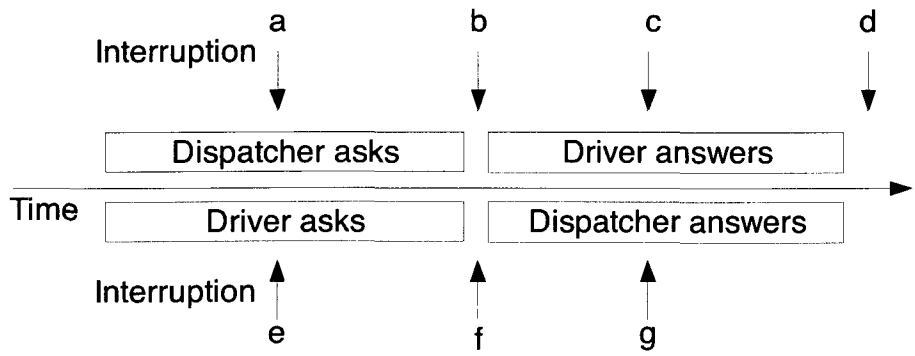


Figure 3.10. Interruption timing.

We coded each presentation of a visual stimulus with "a" through "g" based on where it happened with respect to the model in Figure 3.10. Each presentation resulted in the eventual initiation of an interruption (switch to the interrupting task). We also coded the interruption initiated by the drivers based on where it happened with respect to the model in Figure 3.10.

Figure 3.11 shows an example of how timing is assigned to a segment of speech. Before the dispatcher gives an instruction, there is no communication and it is part "d" of the adjacency pair. When the dispatcher gives the instruction "Take right at the next intersection" it is part "a" of the adjacency pair. Pause before the driver provides response is marked as "b", and the driver's response itself is "c". Now the first adjacency pair is done and in between the adjacency pairs we have pause "d". When the driver makes a statement "I just passed subway on the left" it is part "e". Part "e" is followed by the pause "f" before the dispatcher provides the response "Ok", which is part "g", which ends the second adjacency pair.

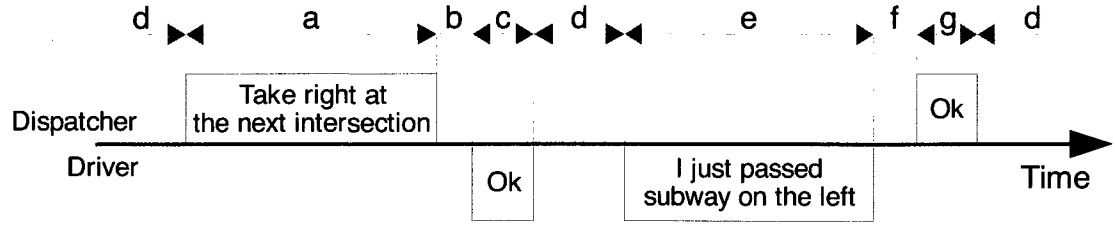


Figure 3.11 Example of codes assigned to adjacency pairs.

It is possible for an adjacency pair to be incomplete, for example, if the driver makes statement after statement without any response from the dispatcher the adjacency pairs are marked as shown in Figure 3.12. The first part of an adjacency pair "I am approaching an intersection" does not have a response from the dispatcher. When after a pause the driver starts the next statement "I am proceeding to take that right" it is again the first part of the adjacency pair. If statements were separated by 750 milliseconds they were considered different utterances belonging to different adjacency pairs. This duration was used by Nakajima and Allen [71] in their research on discourse structure.

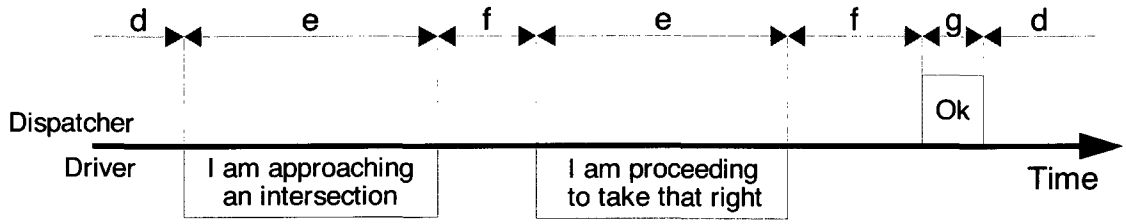


Figure 3.12 Example of incomplete adjacency pairs.

The codes for dialog states allowed us to see what behavior subjects utilize when switching between tasks, which is the subject of hypothesis 4 (switching behavior). We used the response time to see the effects of urgency on the ongoing task, which is the subject of hypothesis 5 (effects of urgency of the interrupting task). The time between visual stimulus presentation and introduction of the interruption by the driver is considered the response time to the interruption stimulus. Figure 3.13 shows how the response time to the visual stimulus was calculated.

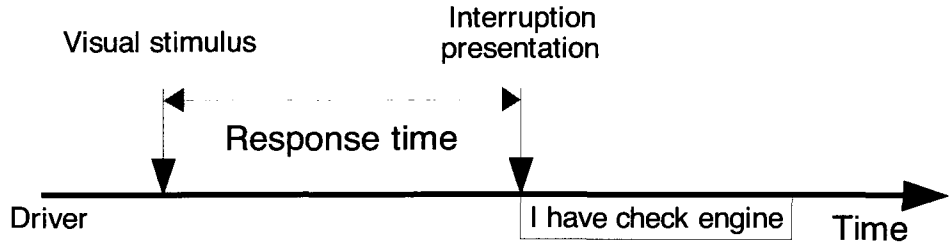


Figure 3.13 Response time to visual stimulus.

3.8 Dependent variables for driving

The DriveSafety DS-600c driving simulator allows the recording of standard driving measures, such as lane position, vehicle velocity, and steering wheel angle. All the values within a 10 meter radius from the center of an intersection were assigned to a difficult road condition, while the other values (straight segments between the

intersections) were assigned to easy conditions. Intersections and straight roads formed separate road segments. We calculated variances for each measure for every segment. The variances were averaged for each segment to obtain a single value per segment. These values were averaged for each subject to obtain a single value per subject.

Lane position is the position of the center of the simulated vehicle and is measured in meters. Higher variance characterizes poor driving performance, since it indicates that the participant weaved in the lane, and perhaps even departed from the lane, which has potential to cause an accident if there is a car in the adjacent lane.

The vehicle's velocity is measured in meters per second. Higher velocity variance does not necessarily mean poor driving performance. Nevertheless, drivers tend to reduce the speed [56] when they are concerned about their safety, for instance, when driving on a narrow road, or when they are distracted, for example, when talking to a passenger. This implies that a slower velocity for a portion of the road could indicate that the driver was concerned about safety or otherwise distracted.

Steering wheel angle is measured in degrees. Higher steering wheel angle variance does not necessarily show poor driving performance, for instance, when driving on a curvy road the variance is higher because following a curvy road requires varying the steering wheel angle constantly. In spite of this, comparing the performance of multiple participants on the same road can be used as a relative measure of driving performance. A higher variance could be an indication of increased effort of a driver to remain in his lane.

3.9 Experiment procedure

The following steps were taken during the experiment:

1. Subject preparation: consent forms, questionnaires, and introductions;
2. Training for the ongoing task;
3. Training for the interrupting task;
4. Training for the ongoing task with interruptions;
5. Experiment;
6. Subject release: questionnaires, debriefing, and reward.

All participants were given an overview of the simulator, and were trained to perform the ongoing task, interrupting task, and then both tasks at the same time. Training took about 10 minutes during which the dispatchers were given a map shown in Figure 3.14. Participants then performed the actual experiment which lasted about 40 minutes. At the end, the participants completed questionnaires and were debriefed. The subjects were presented with printed questionnaires which are shown in Appendix B. The text of the game instructions as given to the participants can be found in Appendix C.

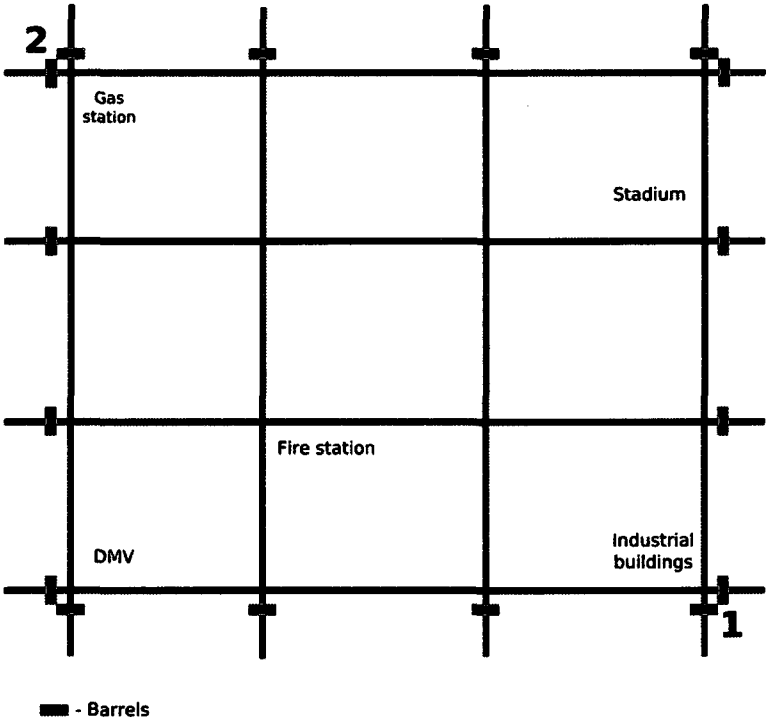


Figure 3.14 Map given to the dispatcher during the training.

3.10 Subjects

The recruitment was performed using flyers and e-mails on university mailing lists. The fliers were posted on bulletin boards at the Durham campus of the University of New Hampshire. The electronic version of the flyer was sent out to the student mailing list of the Electrical and Computer Engineering Department and to the Graduate School of the University of New Hampshire.

The experiment was completed by ten participants (five pairs) between 20 and 43 years of age. The average age of the participants was about 30 years and 30% were female. Subjects received compensation in the form of $10 gift cards.

3.11 Corpus and tools

We recorded the speech of all participants, as well as the car position. Vehicle data were collected at 10 Hz, resulting in about 90,000 vehicle data points for 2.5 hours of driving. We also recorded the time the visual stimuli appeared and synchronized these times with the audio recording of the participants. The five pairs of participants were presented with a total of 286 visual stimuli. Speech Viewer from CSLU toolkit 2.0 was used for audio data annotation. Speech recordings were transcribed by hand. Every interruption had an assigned code for the timing of visual stimulus presentation and the timing of interruption initiation by the driver. SPSS Statistics 17.0 (now called PASW Statistics) was used to perform statistical analysis of the data. We used ANOVA repeated measures to compare measures related to the same subject, such as response time for different urgency levels.

3.12 Results and discussion

We analyzed three aspects of the data. First we looked at the average response time of the driver to urgent and non-urgent visual stimuli. This was a test for hypothesis 5, which stated that urgent interruptions result in a faster response. Figure 3.15 shows the average response times for all subject pairs. We found no significant difference in the average response time depending on the urgency of the interruption ($F(1,4)=0.01, p=0.937$), possibly because participants did not realize that some interruptions were more urgent than others.

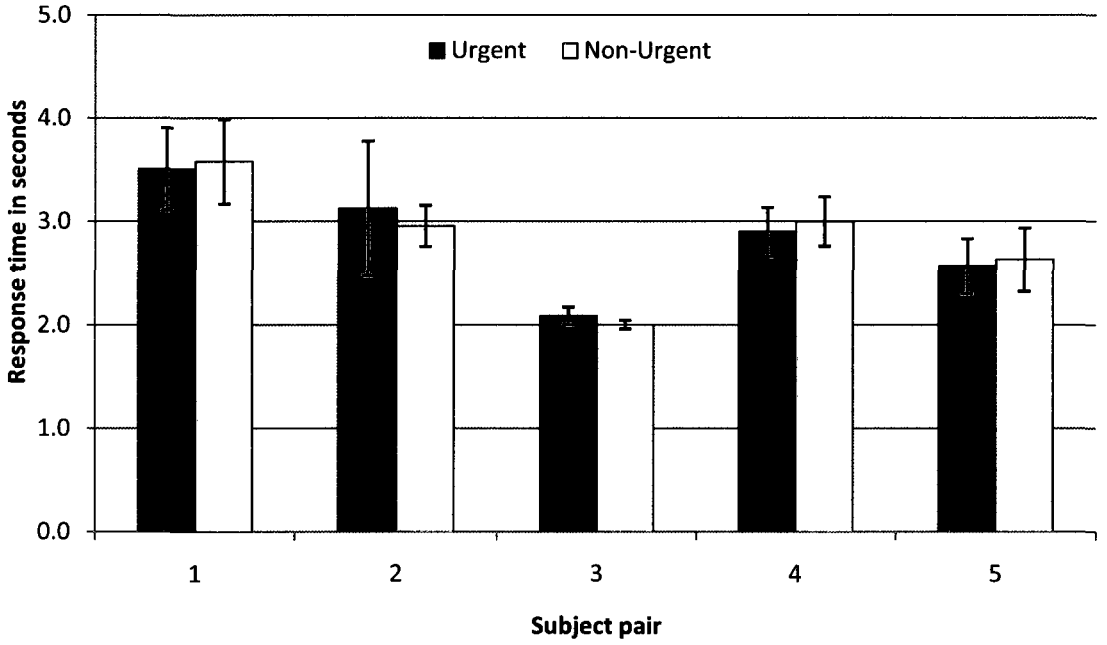


Figure 3.15 Average response times of the drivers.

The response times are slower (average around 2.8 seconds for all cases) than reported by Tsimhoni et al. [62] (average 1.3 seconds), who investigated reading messages on a heads-up display while driving. A reasonable explanation for this is that in our experiment the driver was engaged in verbal communication with the dispatcher and did not pay as close attention to the messages as the participants in the study of Tsimhoni et al. Even more likely, the drivers were complying with established conventions in human-human dialog, and so waited for a suitable point in the interaction. This waiting for an opportunity to speak slowed down their response.

We next analyzed what dialog states allow people to initiate a dialog thread switch (hypothesis 4 – switching behavior). Note that the driver could have ignored the visual stimulus, but this happened only 5 out of 286 times, hence we did not further consider these cases. This left us with 7 x 7 = 49 possible types of interruption (7 parts of

adjacency pairs for visual stimuli presentation and interruptions presentation, section 3.7, pg. 37). We decided to focus on interruptions in which the stimulus occurred during the first part of an adjacency pair ("a" or "e") as this is the point in the local discourse structure that has the longest duration.

When a stimulus is presented during the drivers' first part ("e") 11% of the time the driver interrupts his own first part ("ee") (see Figure 3.16). In 27% of the cases he/she completes the first part and then introduces the interruption ("ef"). In about 2% of the cases the driver introduces the interruption during the dispatcher's second part ("eg"). Most often, in 47% of the cases, the driver waits until after the adjacency pair is over ("ed"). In about 10% of the cases the driver introduces the interruption during the first part of the next adjacency pair when the dispatcher is speaking ("ea"). Finally, in 3% of the cases he/she interrupts after the dispatcher's first part in the next adjacency pair ("eb").

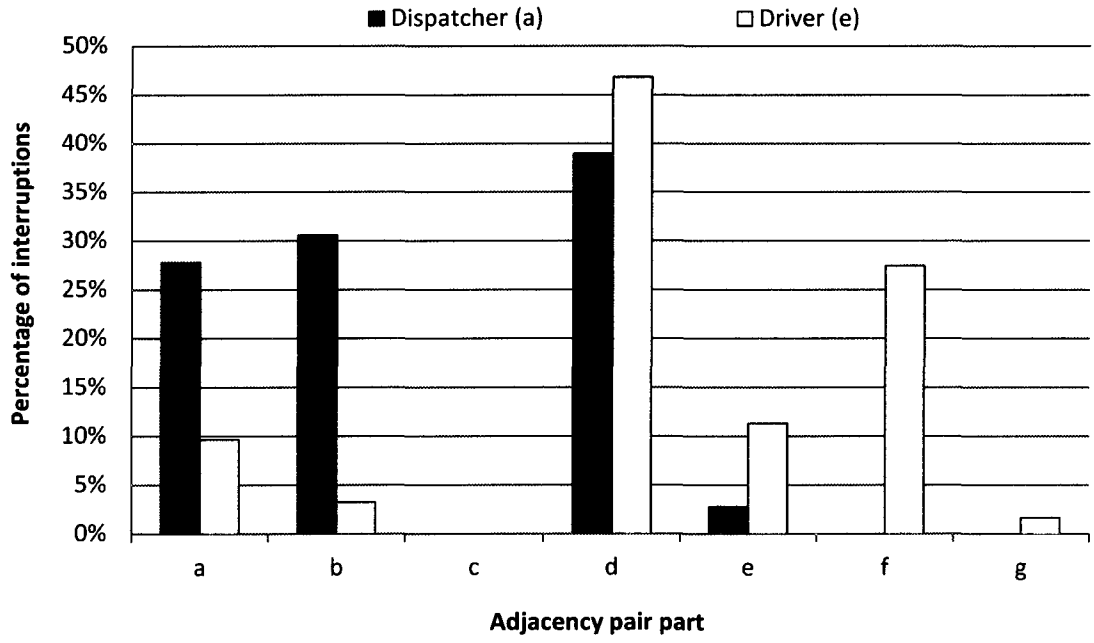


Figure 3.16 Interruption initiation timings.

When the stimulus is presented while the dispatcher is speaking the first part ("a"), the driver interrupts immediately in about 28% of the cases ("aa") and after the first part in about 30% of the cases ("ab") (see Figure 3.16). Again, most often, 39% of the time, the interruption came after the adjacency pair was over ("ad"). In about 3% of the cases each, the interruption came in the next adjacency pair during the driver's first part ("ae").

The above data show that the driver often waited to initiate the interrupting task until after the adjacency pair was done. This might account for the difference between the average response times in this study and the one reported by Tsimhoni et al. [62]. We also looked at the average response time of drivers during difficult and easy driving conditions. We defined difficult driving as driving within a radius of 10 meters of the center of an intersection. The drivers spent only about 8% of their time driving through the intersections and thus, on average this resulted in only 5 visual stimuli out of 57 being presented in difficult driving conditions. Therefore, we were not able to compare performance measures for difficult and easy driving conditions.

3.13 Conclusion

In this experiment, we tried to determine some of the conventions that humans follow in initiating a switch to a new dialog thread. We found that when the stimulus to signal the interruption was in the first part of an adjacency pair, participants either immediately interrupted the first part, or waited until the conclusion of the adjacency pair. This might indicate that participants were trying to avoid having the first part of an

adjacency pair pending during a thread switch, so that there is a simpler discourse context to resume.

The lack of the context build-up in the ongoing task did not allow us to investigate how subjects recover from the interruptions. This happened because the verbal component of the navigation task could be treated as a series of separate steps which do not depend on each other. On the other hand, the interrupting task was very simple and did not allow us to control the difficulty of the interrupting task. Therefore, we decided to modify both the ongoing and interrupting tasks. We wanted to create tasks that are more structured (have better defined adjacency pairs) and allow for a better control over the difficulty of the tasks. During the navigation experiments subjects exhibited a range of behaviors, for example, some subject pairs had a driver that took the initiative and was talking most of the time, while other pairs had a dispatcher that was asking a lot of yes/no questions. Such situations created imbalance in the amount of time the drivers and the dispatchers were talking during the experiments. We intended for the new tasks to be designed in a way that would not allow such a situation to happen.

We also needed to balance the easy and difficult driving segments in order to better understand the impact of driving difficulty on the spoken tasks. Using a city scenario with the traffic and pedestrians created a large variation in the driving data due to the stop signs, traffic, and pedestrians. All of these factors confounded our ability to compare effects of the driving difficulty on the spoken tasks. This meant that the city scenario had to be simplified and transition between the road difficulties had to be clearly marked. In the next chapter we describe the new spoken and driving tasks.

CHAPTER 4

TWENTY QUESTIONS EXPERIMENT

The navigation experiment design suffered from a number of flaws. For instance, the subjects did not build up discourse context as they performed the ongoing task. At the same time, the interrupting task did not allow us to control the difficulty of the task. In addition, the previous experiment was not designed to investigate all of our hypotheses. Our new experiment design aimed to correct the flaws and test our other hypotheses. Namely, the new tasks allow us to test how spoken tasks performance is affected by driving (hypothesis 1) and how driving is affected by the spoken tasks (hypothesis 2). We also designed tasks that allow us to test how timing of a switch between the tasks affects spoken tasks (hypothesis 3). Finally, the new tasks offered a different way to look at the switching behavior of the subjects (hypothesis 4).

In the new experiment one participant was driving a simulated vehicle while conversing with another person situated in a different room. Speech was the only modality of communication available for the participants. This experiment setup is inspired by a real life example: a police officer on patrol. Officers must communicate with a dispatch center using radio, which is a speech-only (unimodal) communication channel. Officers also perform a manual-visual task – driving a vehicle. Dispatchers, on

the other hand, are not driving a vehicle, even though they are using a computer. We selected the spoken tasks based on the constraints that we will describe below.

4.1 Constraints

The following paragraphs describe constraints we worked with when creating the experiment design. These constraints were suggested by the research done elsewhere (described in Chapter 2) and our previous experiences [11,70]. The purpose of these constraints was to be able to select tasks that could address our hypotheses.

To compare performance measures on spoken tasks for both participants, the spoken tasks must require both partners to speak equally. Hence, we avoided tasks which could be accomplished with one of the participants speaking little or not at all. In going through training and then completing the verbal tasks during experiments, participants could easily spend 60 minutes on these tasks. Thus, the tasks had to be complicated enough for the subjects not to run out of things to say, and they had to be engaging enough for participants to be willing to keep talking. In other words, the tasks have to be realistic, because in our previous research we found that tasks that are not realistic lead to poor participant buy-in [70].

Spoken tasks must be designed to have little interference with driving. The 4-dimensional multiple resource model described by Wickens [1] gives guidelines for the design of the tasks done in parallel. The four dimensions of the model are: sensory modalities, codes, channels of visual information, and stages. Figure 4.1 shows three dimensions of the model. The fourth dimension is nested only in visual resources and is not shown to simplify the figure. We decided to separate the manual-visual task from the

multi-threaded dialog as much as possible along these dimensions, in order to remove possible interference between the driving and spoken tasks. It is known that the multiple resource model cannot explain all of the interferences between the tasks [1,52], but using this model as a guideline allows us to better understand the relationship between the manual-visual task (driving) and multi-threaded dialogs (spoken tasks).

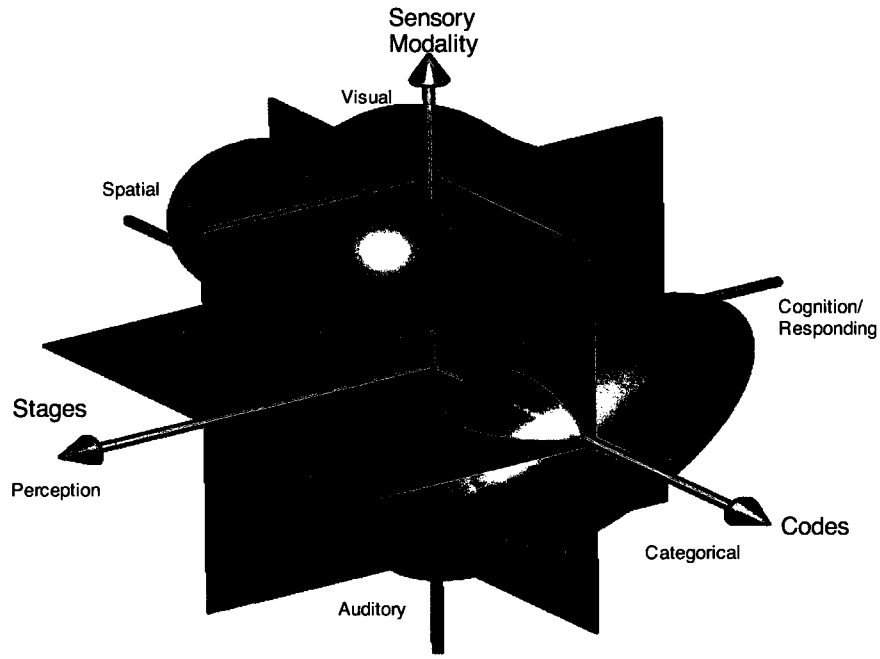


Figure 4.1 Multiple resource model representation (top object represents driving task; the other object represents spoken tasks).

Sensory modalities are divided into visual and auditory modalities (smell, tactile, and temperature modalities [34] are not discussed in this dissertation). Driving is an activity that utilizes visual attention, while the spoken interaction utilizes auditory modality. Given that we focus on command and control type of spoken interaction, there is a need to provide some input to initiate the spoken dialog. In previous experiments with multi-threaded dialogs this input was provided visually [30], or using multiple

modalities [29]. This was possible because participants were not involved in a manual-visual task. Completely removing visual information from the tasks limits the types of possible tasks and makes most of the tasks very challenging for the subjects. For example, most people can play chess while having the board with the pieces in front of them, but it is almost impossible for most people to do the same if they cannot see the chess board with pieces. We experimented with different task combinations in our previous work, and we found that often times the tasks are too easy or too difficult as shown in Figure 4.2. Figure 4.2 shows how difficulty of the tasks changes as a function of some task parameter. From our experience it seems that the general form of the function is exponential. This means that it is hard to choose the proper task difficulty. For instance, rotating a sequence of three letters was easy, but doing the same operation with four letters was much harder. The restrictions on sensory modalities decreased the number of possible tasks that can be used during the experiments to test our hypotheses. We were limited to the tasks that have very low demand for visual resources.

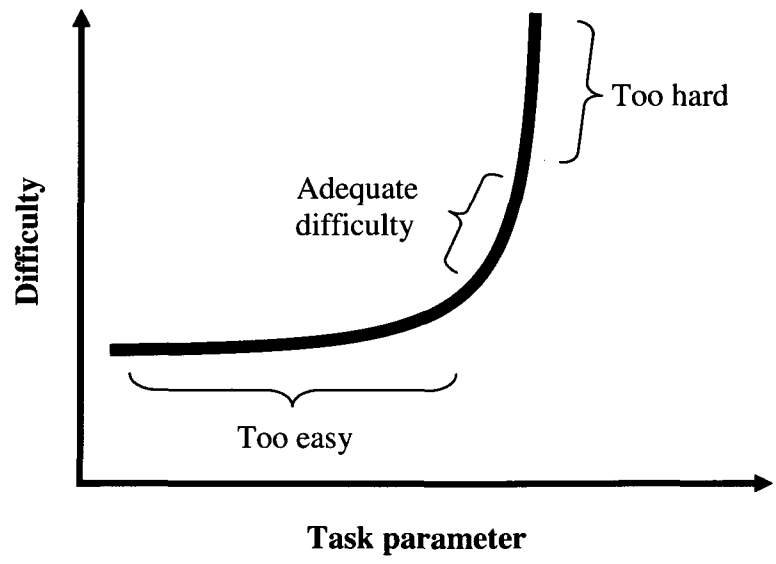


Figure 4.2 Task difficulty vs variation in task parameter.

The code dimension of the multiple resource model differentiates between spatial and categorical (usually linguistic or verbal) processes. Tracking and steering are spatial tasks, while speaking is a categorical task. Navigation can be accomplished using spoken directions, but it might utilize spatial resources. We did not account for such a possibility in our previous experiment setup (Chapter 3). We also decided not to use tasks which would require hand movements. This allowed the driver to keep his hands on the steering wheel at all times.

Visual modality of processing is subdivided into focal and peripheral vision. There is evidence that some driving tasks utilize different types of vision [72]. For example, lane keeping and speed control might utilize ambient vision, but focal vision is utilized for detection and identification of road hazards. This introduces another restriction on the tasks used in the experiments and we should not assume that tasks that use peripheral vision do not influence driving performance.

The stage dimension is divided into a perceptual, cognitive, and response stages. For example, tasks that require perception should interfere less with tasks that require a response, as opposed to tasks that require cognitive effort. Both driving and spoken tasks will require perception, cognition, and response. It is important to notice that perception for visual and audio channels are different. The cognition stage contains different resources for spatial and categorical (verbal) tasks, and driving utilizes manual response resources, while spoken tasks use speech response resources [1].

Figure 4.1 shows a grey object (top object) representing driving and a yellow spheroid (the other object) that represents spoken tasks [66]. The location of the objects serves to illustrate what resources are required for the tasks. It is less informative on how

much of these resources are required. Table 4.1 shows the dependence of different tasks on a given resource, as described above. We assume that a value of 0 indicates that the task does not involve a particular resource. Greater values indicate greater involvement of a resource in the task. For example, the task of keeping a vehicle in its respective lane might involve resource at the perceptual (localizing the lane markers), cognitive (determining the relative position of the vehicle within the lane), and response (turning the steering wheel) levels. Hence, the demand vector across these dimensions is [1,1,1]. Driving at night on the same road might yield in a demand vector [2,1,1], meaning that it is harder to drive at night than during the day. Similarly to Figure 4.1 these numbers only serve to illustrate a relation between different tasks. The demand scalar is an additive combination of the demand vector. The demand scalar illustrates the overall demand of the task.

Task	**Demand vector**								**Demand scalar**
	Perception				**Cognition**		**Response**		
	Vf	Va	As	Av	Cs	Cv	Rs	Rv	
Easy driving	1	1	0	0	1	0	1	0	4
Difficult driving	1	2	0	0	1	0	2	0	6
Spoken task 1	0	0	0	1	0	1	0	1	3
Spoken task 2	0	0	0	1	0	2	0	1	4

Table 4.1 Demand vectors for the driving and spoken tasks (V = Visual, A = Auditory, C = Cognitive, R = Response, f = Focal, a = Ambient, s = Spatial, v = Verbal [1]).

The driving task also had constraints associated with it. For instance, the task of going from point A to point B along a predefined path might require use of a navigation device, which has its own implications [6]. For example, we would have to present the information from the navigation device to the driver during the experiment, which would create an interruption by itself. We decided to avoid driving tasks that would require additional devices.

4.2 Hardware setup

This section describes hardware used in the experiment, such as driving simulator, eye-tracker, audio, and video equipment.

4.2.1 Driving simulator

The experiment involved driving a high fidelity DriveSafety DS-600c simulator described in detail in section 3.2.1 (pg. 27).

4.2.2 Eye tracker

We used the SeeingMachines faceLab 4.6 eye-tracker system, which was installed in the simulator to track the gaze direction of the driver (Figure 4.3). The eye-tracker cameras were positioned on the dashboard above the steering wheel. The eye-tracker provided data at 60 Hz. We collected multiple data channels from the eye tracker (gaze direction, head direction, blinking information, intersection of the gaze with the screen). These data channels are available for future investigation, because we used a limited subset of the data in this research.

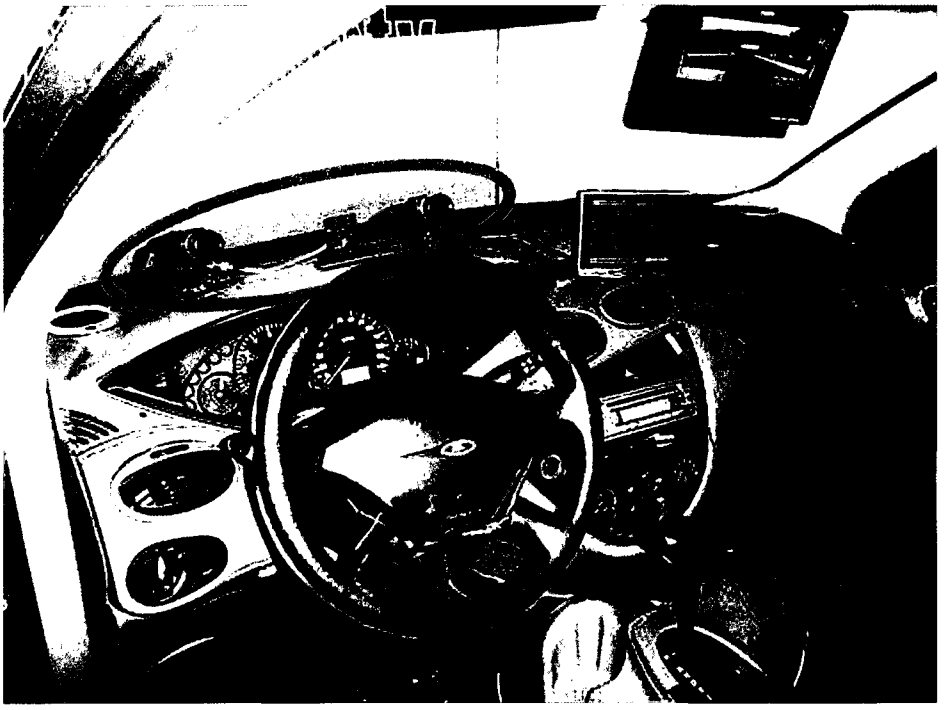


Figure 4.3 Eye-tracker cameras installed inside of the simulator cab.

4.2.3 Audio communication and recording

Two people participated in each experiment. They communicated using headphones and microphones. Their communication was supervised and recorded. Figure 4.4 shows a driver in the driving simulator with headphones and microphone used to communicate with a dispatcher. Figure 4.5 shows the dispatcher wearing headphones. The driver and dispatcher were located in separate rooms and could only communicate using headphones and microphones. All communication was recorded synchronously at 44100 Hz as mono signals in two separate channels (one channel for the dispatcher and another channel for the driver).

Figure 4.4 Driver in the simulator room.

Figure 4.5 Dispatcher in the dispatcher room.

4.2.4 Video recording

The experiment was recorded for presentation and data verification purposes with four video cameras:

- Sony HDR-HC3 HDV 1080i for the eye tracker video;
- Panasonic PV-GS65 for the over-shoulder video;
- Sony DCR-HC28 for the head and hands video;
- Sony DCR-HC52 for the dispatcher video.

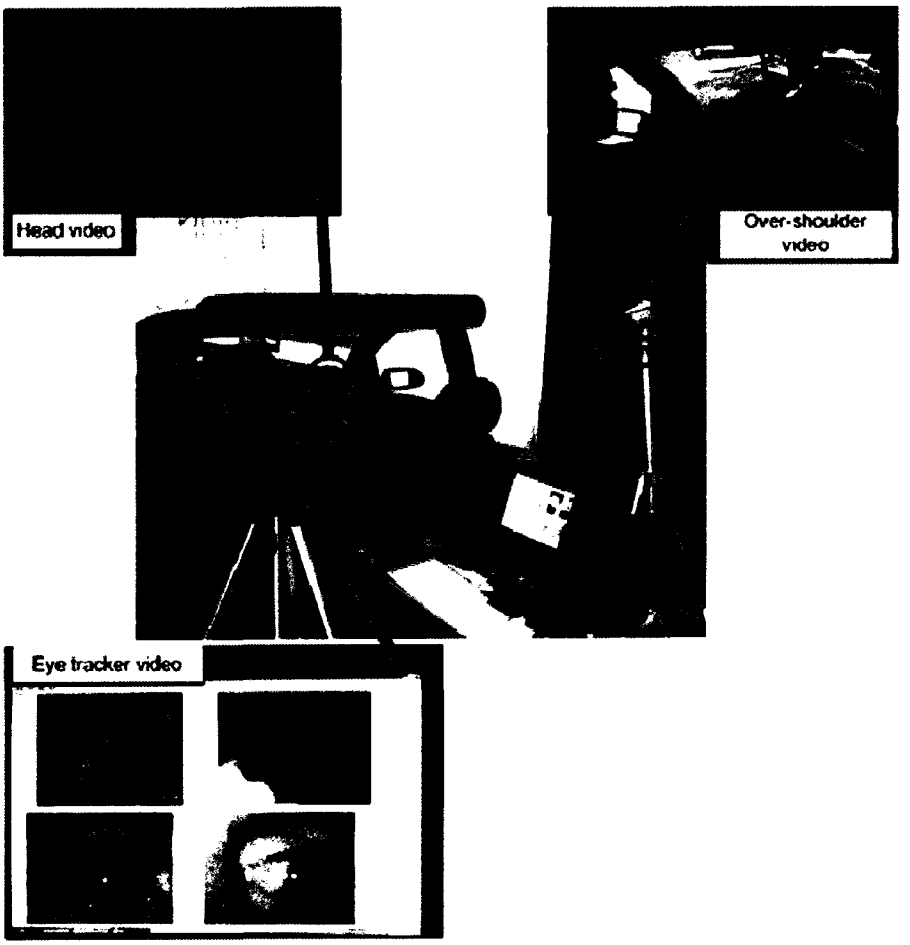


Figure 4.6 Camera setup for drivers [6].

Figure 4.6 shows the positioning of the video cameras and view from these cameras. In situations when the eye-tracker did not a record participant's gazes, e.g. if participant's hand was covering the IR pod, the video recordings could be used to estimate gaze information by visual inspection of the subject's eyes.

We also recorded head video of the dispatcher as shown in Figure 4.7. This recording could be used to confirm the dispatcher's actions in case audio recording fails by listening to the video recording.

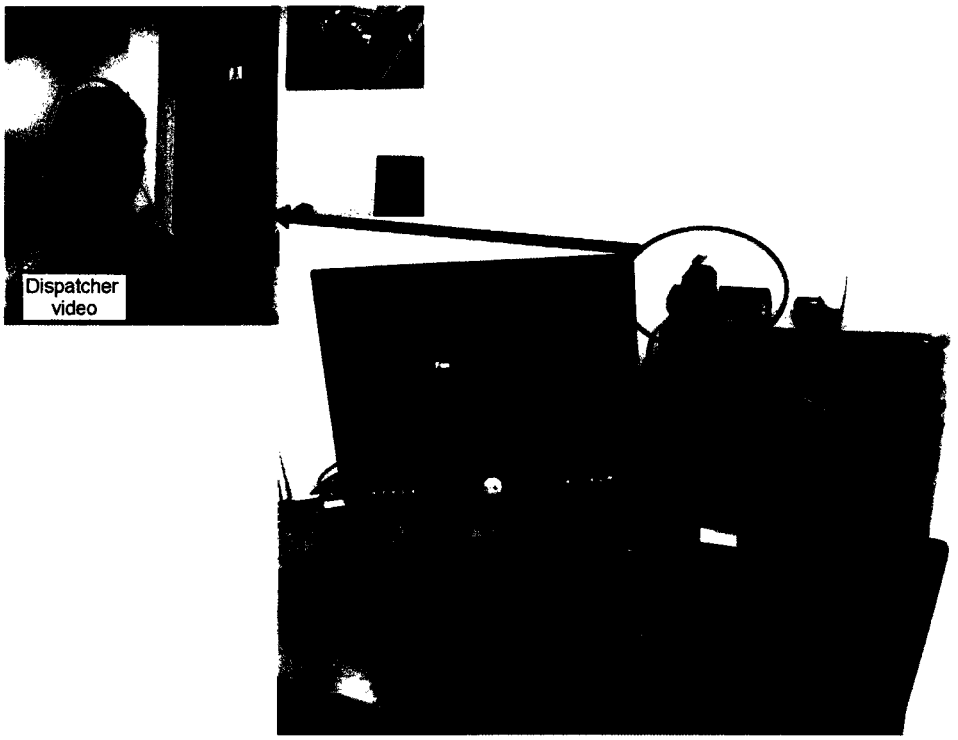


Figure 4.7 Camera setup for dispatchers.

4.3 Ongoing task

The ongoing speech task was based on a game called Twenty Questions. The goal of the game was to discover an object by asking no more than twenty questions. The game is based on the fact that the information (as measured by Shannon's entropy

statistic) required for identification of an arbitrary object is about 20 bits. If each question is structured to remove half of the objects, 20 questions will allow one to differentiate between 1,048,576 objects (2^{20}). Therefore, the most efficient strategy for the twenty questions game is to ask questions that will split the field of remaining possibilities in half. This process is analogous to a binary search algorithm in computer science, which involves creating a tree structure and then traversing this structure until a solution is found [73].

The game allows the players to build a context which must be restored during resumptions. This means that at the time of the resumption the participants already exchanged some information and they need to make sure that both of them remember what that information is after the interruption is over. The solution space of the task can be limited by restricting the number of objects allowed in the game. Hence, participants have a finite number of objects to memorize, which allows us to control the training time for the experiments. Changing the number of objects in the solution space also allows us to control the difficulty of the task. We chose to have 18 objects, as explained below. In addition, the game has clearly differentiated adjacency.

We defined a list of 18 objects that could be described as electric appliances for home use: microwave, stove top, blender, mixer, refrigerator, can opener, TV, radio, fan, heater, vacuum cleaner, main, light, electric shaver, powered toothbrush, hair dryer, washing machine, dryer, and hair trimmer (a fewer than 20 questions is required to complete our variation of the game, but for simplicity we still refer to the game by its original name: twenty questions game). We split all the objects as belonging to three different rooms (6 objects in every room): living room, kitchen, and bathroom. Figure 4.8

shows an example of objects used in the game from the bathroom (see Appendix C for other images). These are common objects, which should be familiar to the subjects. These objects were presented in their common settings, which should ease the memorization process. For example, a toothbrush was in the bathroom, and a TV was in the living room. Subjects were instructed that only the described objects were allowed in the game. This was done to make it clear what to expect during the game. We presented all the items involved in the game in pictures such as Figure 4.8 to create a visual connection between words and real objects. Paivio [33] found that it is easier for people to memorize and retrieve words associated with concrete nouns, especially when they have pictorial representations. Hence, we used concrete nouns with a pictorial representation to ease the memorization process.

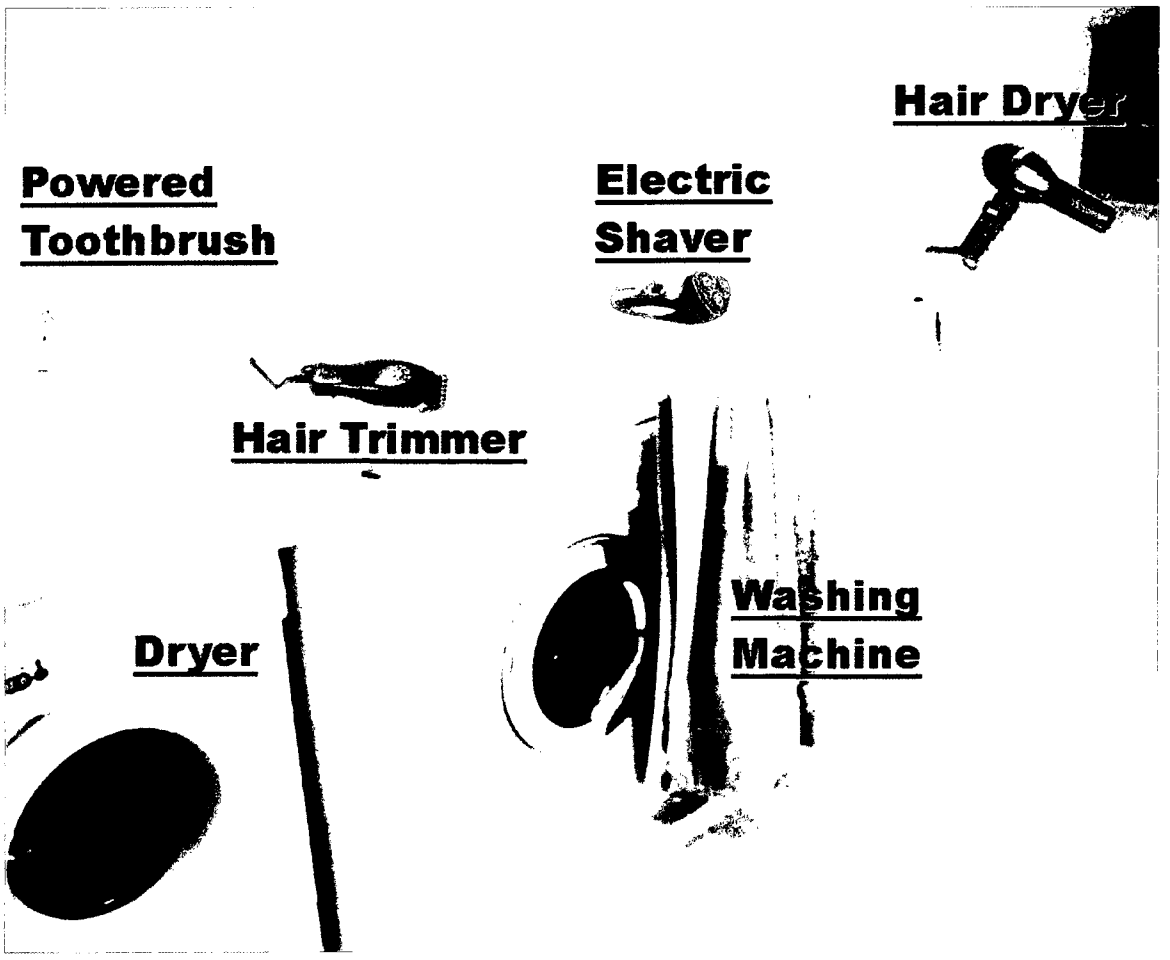


Figure 4.8 Bathroom objects available for the game.

The subjects were given a training tree that they might want to use, which shows all available objects (Figure 4.9). During our pilot studies we found that it is difficult for people to come up with their own trees quickly. By providing an example of a possible way to split objects, we made it easier for people to understand how to play the game. Games were very quick (less than 30 seconds) when people could see this tree in front of them, but during the experiment they had to use their memory, which slowed down the speed with which subjects asked their questions and on average stretched the games to 1 minute and 30 seconds. Allowing drivers to look at the training tree during the experiment would also distract them from the driving task. At the same time we wanted

to compare how driving interfere with this task, which can be done by comparing how the drivers and the dispatchers perform. Therefore, we needed to make sure that the task of driving was the only factor that changed between the drivers and the dispatchers. Hence, both subjects were not allowed to look at the training tree during the experiment.

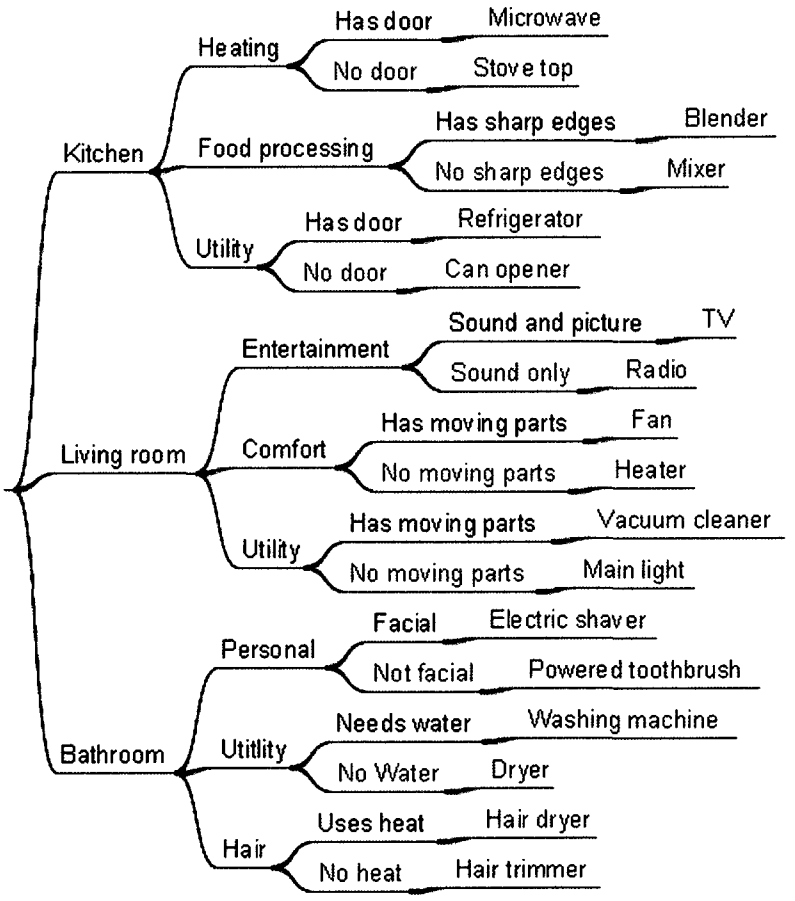


Figure 4.9 Training tree for classification of the appliances.

A single twenty questions game forces one person to ask questions, while the other person only says "yes" or "no". This creates an imbalance in the amount of time the participants are involved in the conversation. In order to resolve this we asked the drivers and the dispatchers to play twenty questions games in parallel by alternating their questions. The driver and the dispatcher were given the words for the other person to discover when the game starts. For the driver, the word was present on the screen below

the horizon level, but above the dashboard. The word location allowed a quick data access, while minimizing interference with the driving and not occluding the leading vehicle (based on the research done by Tsimhoni [62]). Figure 4.10 shows word Microwave that is presented to the driver and should be discovered by the dispatcher (the text was shown in red, white outlines are used to make the word visible in grayscale). Figure 4.11 shows word TV that is presented to the dispatcher and should be discovered by the driver.

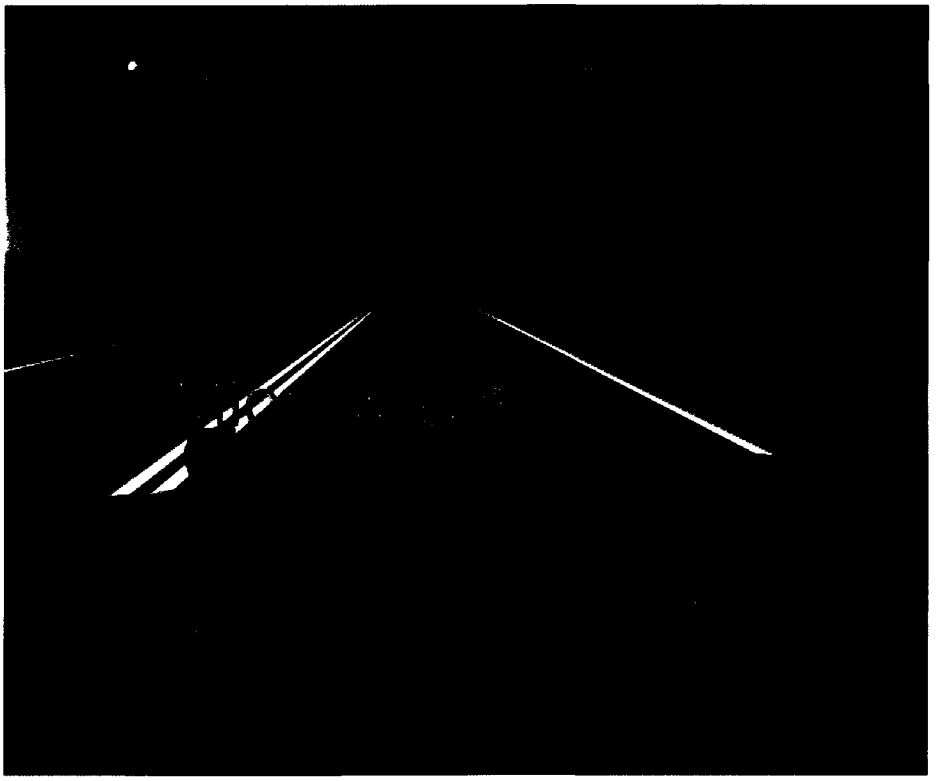


Figure 4.10 Twenty questions game information shown to the driver.

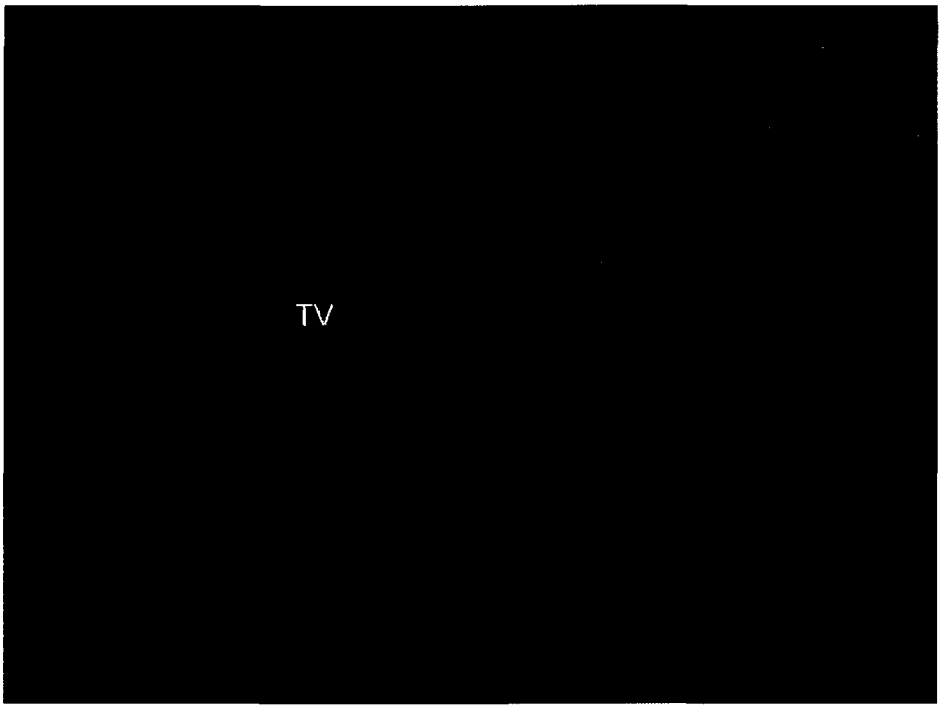


Figure 4.11 Twenty questions game information shown to the dispatcher.

4.3.1 Ongoing task structure

For the ongoing task we call a single adjacency pair a game turn. In this context the term *game* is related to the ongoing task and can be replaced with the phrase *twenty questions game*. There should be no confusion with conversational games which are tied into the discourse structure of a dialog and are used in analysis of task oriented dialogs [74,75]. The term *turn* is defined in relation to the games, as opposed to a speaker. For example, Duncan [76] studied how people signal to each other whose turn it is to speak. In our context, one turn is a question by one person and an answer by the other person. Figure 4.12 shows how questions were alternated within a game. In this sense, the subjects are taking turns when playing two twenty questions games in parallel. We identify whose turn it is by the person who is asking a question. When a driver asks a question it is the driver's turn. When a dispatcher asks a question it is the dispatcher's

turn. This is further illustrated in Table 4.2 that shows an example of playing two twenty question games in parallel. Dispatchers were instructed always to start asking questions first when a new game was started in order to make sure that participants do not spend their time negotiating who should start first.

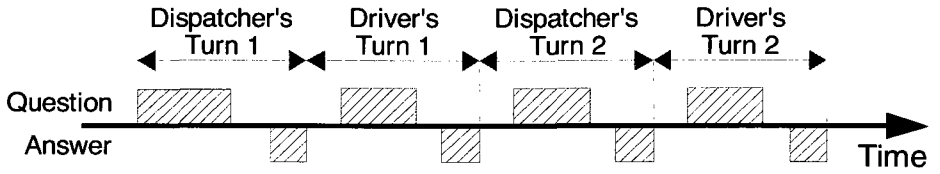


Figure 4.12 Order of turns in twenty questions game.

Code	Speaker	Utterance	Details
U1	Dispatcher	Is it in the kitchen?	Dispatcher's turn 1
U2	Driver	Yes.	
U3	Driver	Is it in the bathroom?	Driver's turn 1
U4	Dispatcher	No.	
U5	Dispatcher	Is it used for heating?	Dispatcher's turn 2
U6	Driver	No.	
U7	Driver	Is it in the living room?	Driver's turn 2
U8	Dispatcher	Yes.	
U9	Dispatcher	Is it used for food processing?	Dispatcher's turn 3
U10	Driver	Yes.	
U11	Driver	Is it a utility item?	Driver's turn 3
U12	Dispatcher	Yes.	
U13	Dispatcher	Does it have a door?	Dispatcher's turn 4
U14	Driver	Yes.	
U15	Driver	Does it have moving parts?	Driver's turn 4
U16	Dispatcher	Yes.	
U17	Dispatcher	Is it a refrigerator?	Dispatcher's turn 5
U18	Driver	Yes	
U19	Driver	Is it a vacuum cleaner?	Driver's turn 5
U20	Dispatcher	Yes	

Table 4.2 Example of parallel twenty questions games.

The subjects were asked to start playing twenty question games as soon as the words appear on the screen. When the words were removed from the screen the subjects were instructed to stop speaking with each other. If the subjects finished the ongoing task,

but words were still on the screen, then they had a choice of chatting with each other until the words disappear.

There were twelve parallel twenty questions games during each experiment for the reasons described in the following sections.

4.4 Interrupting task

For an interrupting task (to simulate a multi-threaded dialog) we use a variation of a last letter word game (a similar task was used as an interruption in a dual task condition in the research by Strayer and Johnston [52]). A person names a word that starts with the last consonant or vowel of the word named by the other person. For example, Table 4.3 shows an interrupting task dialog when a driver sees an interruption and asks the dispatcher to name a word starting with the letter A.

Code	Speaker	Utterance
U1	Driver	Name a word starting with A.
U2	Dispatcher	Apple
U3	Driver	Exit
U4	Dispatcher	Tomb
U5	Driver	Beak
U6	Dispatcher	Kite
U7	Driver	Enter

Table 4.3 Example of an interrupting task.

The time duration of this task can be controlled by increasing the number of words to be named or/and by limiting what type of words can be used. During our preliminary studies we found that naming three 4 or 5 letter words provided us with 10 to 20 seconds of game duration. Words with less than 4 or more than 5 letters resulted in longer time spent on the game. No limitation of the word length often resulted in a very short completion time (less than 10 seconds). We also instructed subjects not to use the

words that were already used. This ensured that the subjects try to come up with the new words instead of reusing the same words. We assumed that the chosen game duration was long enough to create interference with the ongoing task to simulate a multi-threaded dialog.

We instructed subjects to attempt to finish last letter word games in 30 seconds. A progress bar showing how much time is left to play the game was shown on the screen to the person who starts the last letter word game. This was done to motivate subjects to switch to the interrupting task before the ongoing task is complete. At the same time, subjects did not have to interrupt immediately, which allowed them to pick the timing of the interruption presentation. Figure 4.13 shows the letter "A" with a progress bar presented to the driver for the last letter game, while Figure 4.14 shows the letter "B" with a progress bar presented to the dispatcher.

Figure 4.13 Interrupting task shown to the driver.

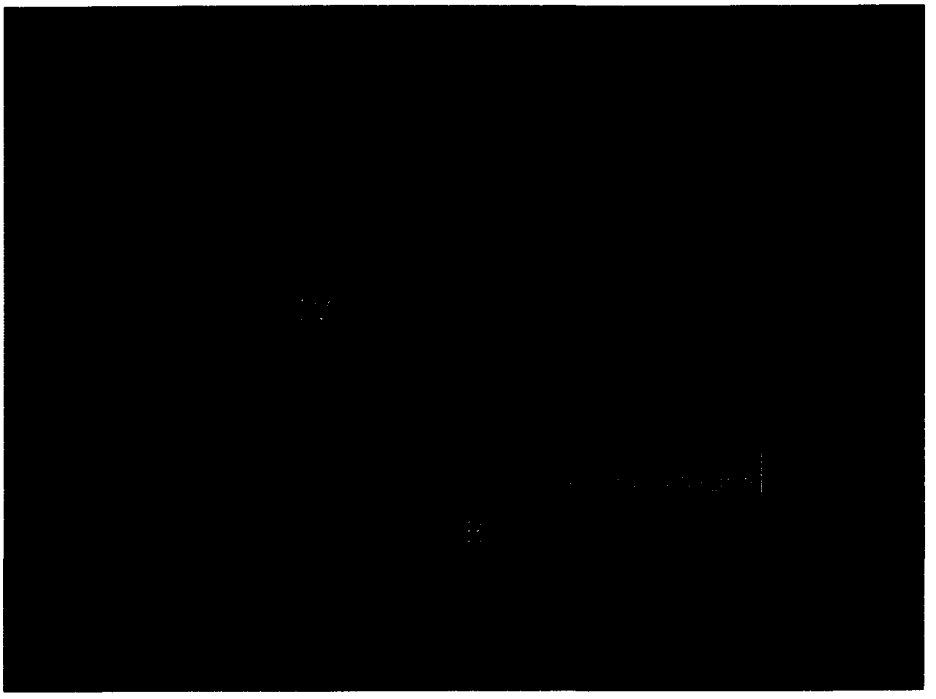


Figure 4.14 Interrupting shown to the dispatcher.

When subjects saw an interruption they had to prompt the partner to name a word that starts with the given letter. This way there was no cognitive load on the subject who received the interruption to come up with the word before the introduction of the interruption. This ensures that any pause between the presentation of the interruption to the subject and the subject mentioning it is not affected by the difficulty of the interrupting task itself. In other words, repeating a prompt does not require as much time as thinking of a word and then saying it [24].

Subjects needed at least four questions to complete a twenty questions game (as described in the section 4.3, pg. 58). We presented an interruption after the first, second, or third questions (different interruption timings). We also present an interruption to the driver or to the dispatcher. Each of the twelve twenty questions games was interrupted. One half of the twelve interruptions were presented to the driver and the other half to the dispatcher. Therefore, the driver was presented with six interruptions, and the dispatcher was presented with six interruptions. We decided to have two occurrences of each interruption timing for each subject. This gave us four interruptions (two for the driver, and two for the dispatcher) that were initiated after the first pair of questions; four interruptions that were initiated after the second pair of questions; and four interruptions that were initiated after the third pair of questions. This added up to 12 interruptions per experiment.

Each interruption was presented after a certain number of turns as explained above. The experimenter kept track of the number of turns in every twenty questions game. Once the required number of turns in a twenty question game was done by the driver the experimenter pressed a button and an interruption was shown after a delay

randomly chosen from 0 to 10 seconds. This ensured that the experimenter did not introduce a bias into the procedure. From our pilot studies we found that it takes about 10 seconds to complete a game turn. Thus, the random delay introduces the interruption during the next turn of the twenty questions game, which is what we would like to happen.

For the interrupting task we considered naming a single word to be a game turn. Similar to the definitions in section 4.3.1, pg. 64, the term *game* is related to the interrupting task and the term *turn* is defined in relation to the last letter word games (similar to explanations in section 4.3.1, pg. 64). A turn starts when the other person requests to name a word or when the other person names a word. The turn ends when the person finishes saying a word. When the driver must name a word it is the driver's turn, and when the dispatcher must name a word it is the dispatcher's turn. Given the rules of the game each subject must take three turns before an interrupting task is complete.

4.5 Multi-threaded dialog

Figure 4.15 shows an ongoing task interrupted by an interrupting task . Once the interrupting task is complete subjects resume the ongoing task. Completion of the ongoing task finishes the game. The first part of the twenty questions game is called *before interruption*, and the second part of the twenty questions game is called *after interruption*. Notice that it is possible for the subjects to run out of time and the ongoing task will not be resumed. In this case there is no resumption activity present for such a game. We minimized such situations by providing enough time for participants to

complete both tasks. We found how much time should be enough based on the data from our pilot studies.

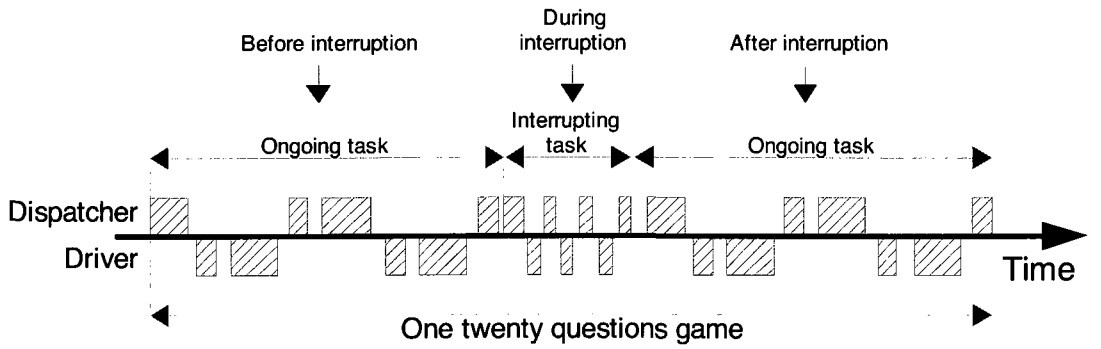


Figure 4.15 Ongoing and interrupting tasks.

We limit the time a person drives during training to 10-15 minutes and during the experiment to 30-40 minutes. We concluded that this duration is satisfactory for our experiments based on the previous research done in our laboratory [11,77,10]. This allows for proper training and does not fatigue drivers to the extent that the fatigue starts affecting the results of the experiment. Using data from pilot experiments we calculated that two minutes is enough time for participants to complete parallel twenty questions games. With a short break between the games (30 seconds) and added time for the interrupting task (30 seconds), the participants played 12 parallel twenty questions games during a 30 to 40 minute long experiment. This number of the twenty questions games matches the number of interruptions that we decided to have during experiments.

4.6 Driving

All drivers were instructed to follow a lead vehicle, which traveled at 89km/h (55mph). The task of following a vehicle forced the drivers to maintain the speed required for the experiment. The leading vehicle was positioned 20 meters ahead of the

subject's vehicle at the beginning of the experiment. The drivers were instructed not to lose sight of the leading vehicle, but there were no instructions as to what distance must be maintained from the leading vehicle. There was another vehicle positioned 20 meters behind the subject's car at the beginning of the experiment. The rear vehicle encouraged the drivers to check the rear and side view mirrors as drivers would in real life driving. The rear vehicle also traveled with the same speed as the leading vehicle, but it slowed down to keep a safe distance from the subject's car when necessary. No other traffic was present on the road to avoid additional variability in driving difficulty.

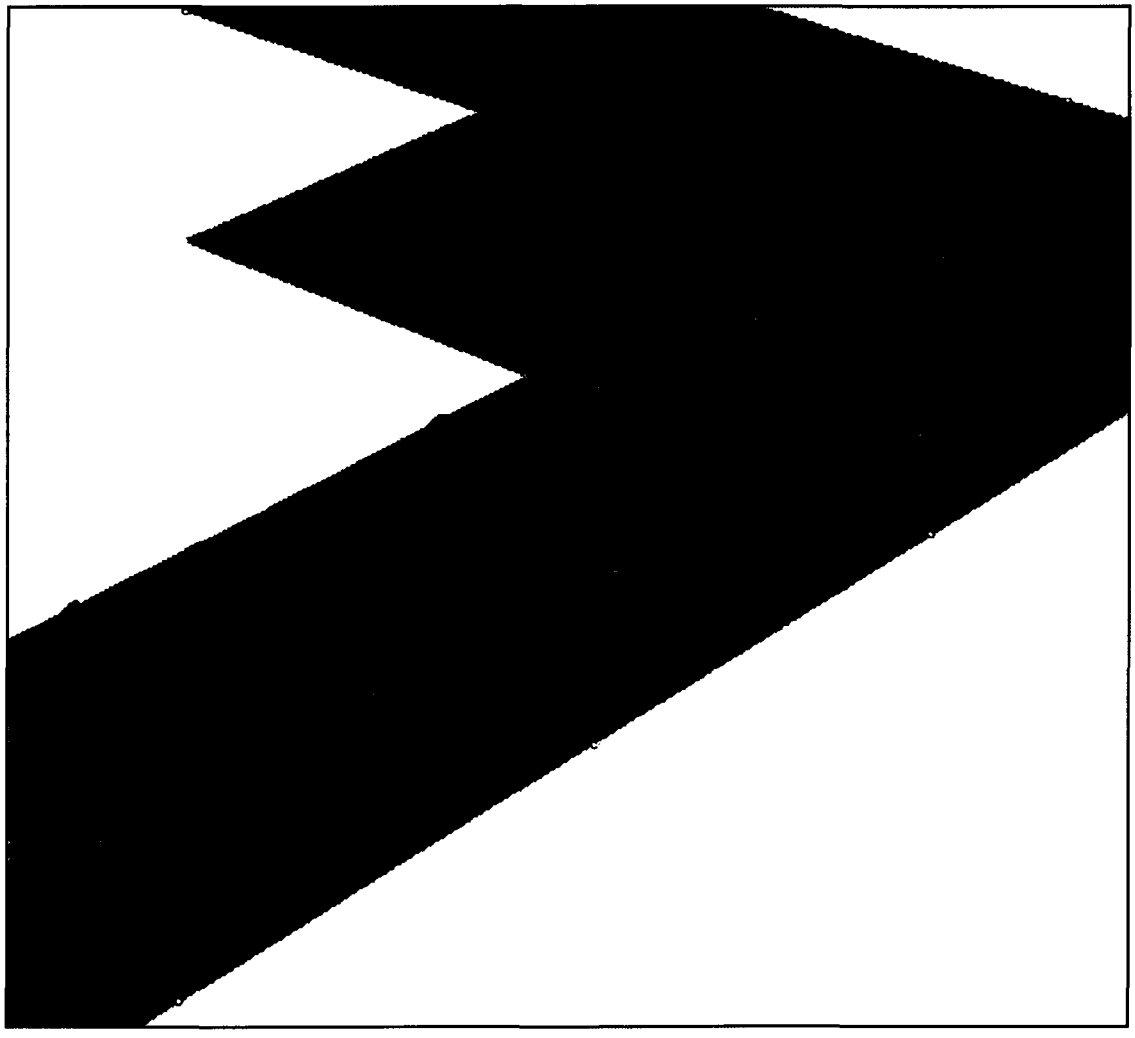


Figure 4.16 Road with trees and houses along it.

The drivers drove on a two-lane road (one lane 3.6m wide in each direction) representing a rural highway in daylight, as shown in Figure 4.10. The separating road marker line between the lanes was full during all times. There were buildings and trees along the road as shown in Figure 4.16.

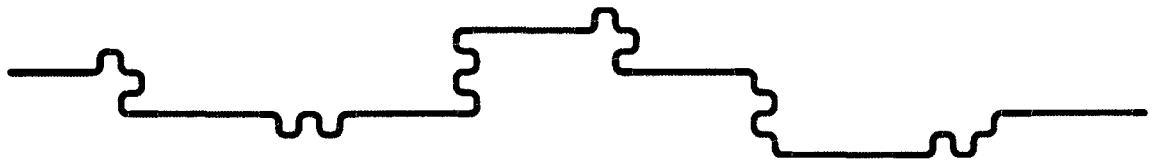


Figure 4.17 Overview of the road.

Each driver traveled along the road that had six straight and six curvy road segments. Figure 4.17 shows a sequence of alternating straight and curvy road segments traversed by a driver in an experiment. Straight segments were 3.4km long and curvy segments were 3.75km long. The difference in distance was due to constrains of the software for the road design. At the beginning and the end of the road we introduced two short regions during which the subjects did not communicate with each other, in order to allow the drivers to transition from one road difficulty to another. We also allowed the drivers to drive for 1.5km when the simulation started to make sure that the drivers adjust their speed to the speed of the leading vehicle. Overall, the road was 47km long.

Each curvy road segment had an equal number of left and right turns. Each turn introduced a 90 degree change in heading over 320 meters of travel (radius of 230 meters). After the change of the direction was complete there were 160 meters of straight road before the start of the next turn. The straight segment before the next turn made sure that two consequent right turns are not different from a left turn followed by a right turn. The previous experiments [10,77] showed that this road geometry at 89km/h does not

cause motion sickness for the majority of the subjects. Tsimhoni and Green [57] found that the driving difficulty increases with the road curvature. According to their model visual demand for curvy roads with the radius of 230m should be 30% larger than for the straight roads. We assumed that this difference in visual demands should provide us with an increased driving difficulty for curvy road segments as compared to the driving on straight road segments.

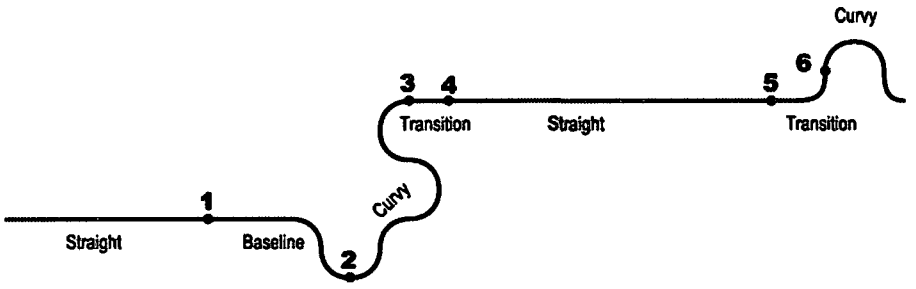


Figure 4.18 Sample of road segments.

Figure 4.18 shows the sequence of a few road segments. Before the point 1 the driver communicates with the dispatcher while driving on a straight road segment. From point 1 to point 2 we have 1km of the baseline section, which included straight and curvy regions. To point 1 and from 4 to 5 there are straight road segments. From point 2 to point 3 there are curvy road segments. 3 to 4 and 5 to 6 are transitional segments, during which subjects were not supposed to talk. From point 6 on there is a curvy road segment. The participants are presented with the twenty questions game words when the driver passes points 2, 4, and 6. The words are hidden when the driver reaches the points 1, 3, and 5. Interruptions are presented somewhere before 1, in between points 2 and 3, 4 and 5, and after point 6. Subjects were instructed to play the twenty questions games only when they saw words on the screen and they had to stop talking when the words disappeared from the screen. This means that subjects could play the twenty questions

game only during 3km length inside of each segment (shown in red in Figure 4.17), and the subjects were requested to be silent during transitions from one segment to another.

It is important to notice that the interrupting task had an explicit time limit with a progress bar shown to the subjects (section 4.8.2, pg. 81). The ongoing task had a "distance" limit, meaning that the participants played twenty questions games only while the drivers drove inside of a 3km range within each road segment (as explained above). Given that the drivers on average had to maintain a constant speed (set by the leading vehicle), the "distance" limit was mostly constant in time (about two minutes). This limit for the twenty questions game was not visually presented to the subjects. The participants were not explicitly informed about this "distance" limit, but they knew from training that they have to stop playing twenty questions games when the words disappear from their screens.

4.7 Independent variables

We focused on three independent variables in this study: subject role, road type, and timing of interruptions. We had five factors for the ongoing and interrupting tasks that could have introduced ordering artifacts: timing of interruptions, twenty questions game words, interruption letter, subject for interruption presentation, and starting road segment. It would take too many experiments to counterbalance all of these factors. Hence, we chose to counterbalance the two factors we assumed could have the most confounding effect on the experiments. The first factor is the type of the starting road segment during which the driver is engaged in the ongoing task for the first time. The second factor is the twenty questions game words. The other factors such as the order

of the interruption timing, interruption letter, and subject for the interruption presentation were coupled with the twenty questions game words as described below. Every ongoing task had the objects to be discovered by the subjects (one for the driver, and one for the dispatcher), an interruption timing (after which turn the interruption was presented), an interruption letter (what letter should be used to start the interrupting task), and the subject role for the interruption presentation (who sees the interruption letter: driver or dispatcher). For example, during game 1 the driver must discover Fan while dispatcher is discovering Can opener; the interruption is presented after the third turn of the game; the interruption has letter B and is presented to the driver. The next game has different words, different interruption timing, letter, and who is presented with the interruption. We created two sequences of these combinations, which are shown in the Table 4.4 and Table 4.5. Both sequences of word pairs for twenty questions games utilized all possible objects. Each sequence for interruptions was designed using the three rules described below.

	Sequence 1				
	Ongoing tasks		**Interrupting task**		
#	**Driver**	**Dispatcher**	**Timing**	**Letter**	**Person**
1	Can opener	Fan	3	B	Driver
2	Stove	Powered toothbrush	2	A	Dispatcher
3	TV	Refrigerator	2	C	Driver
4	Dryer	Radio	1	D	Driver
5	Heater	Washing machine	1	B	Dispatcher
6	Blender	Main light	3	D	Dispatcher
7	Hair trimmer	Mixer	1	C	Driver
8	Microwave	Electric shaver	3	A	Driver
9	Fan	Hair dryer	2	C	Dispatcher
10	Vacuum cleaner	Refrigerator	2	B	Driver
11	Dryer	Radio	3	A	Dispatcher
12	TV	Can opener	1	D	Dispatcher

Table 4.4 Combination of game parameters for the experiment sequence 1.

	Sequence 2				
	Ongoing tasks		Interrupting task		
#	Driver	Dispatcher	Timing	Letter	Person
1	Radio	Hair dryer	1	D	Dispatcher
2	Powered toothbrush	Microwave	3	A	Dispatcher
3	Fan	Washing machine	2	B	Driver
4	Hair trimmer	Mixer	2	C	Dispatcher
5	Heater	Blender	3	A	Driver
6	Vacuum cleaner	Hair dryer	1	C	Driver
7	Main light	Electric shaver	3	D	Dispatcher
8	Dryer	Stove	1	B	Dispatcher
9	TV	Can opener	1	D	Driver
10	Microwave	Washing machine	2	C	Driver
11	Refrigerator	Powered toothbrush	2	A	Dispatcher
12	Radio	Hair dryer	3	B	Driver

Table 4.5 Combination of game parameters for the experiment sequence 2.

Rule 1 stated that the change of the person to whom the interruption is present must not happen more than three times in a row. Otherwise subjects might anticipate the next interruption. For example, if the interruption would be presented to a different participant every single time, the subjects could learn it and, as a result, anticipate who will be interrupted next.

Rule 2 stated that all interruption timings must be presented before they can be repeated, to make sure that most of the interruption timings are separated from each other as much as possible. For instance, there are four interruptions that happen after the third turn of the twenty questions game, and we wanted to make sure that all of these interruptions do not happen at the very beginning or the end of the experiment.

Rule 3 stated that the interruptions after the second and third turns must be as far away (time wise) from each other as possible. This allows us to capture how subjects react to the different interruption timings at the beginning and at the end of the experiment. We expected that more game turns provide more context and consequently more interesting behavior for resumptions and interruptions. Thus, we made the

interruptions after the second and third turn to be far away from each other in time. This should account for possible learning, and/or fatigue effects.

During the experiment each interruption requested to name a word starting with one of the letters: A, B, C, and D. Each letter was used by three interruptions presented to the each subject during the experiment. All of the letters were used before they could be repeated. This ensured that we can see learning effects if any, because the same letters were used at the beginning, middle and the end of the experiment. We used the reverse order of the sequence to counterbalance for the ordering effect and satisfy the rules described above at the same time (as shown in Table 4.4 and Table 4.5).

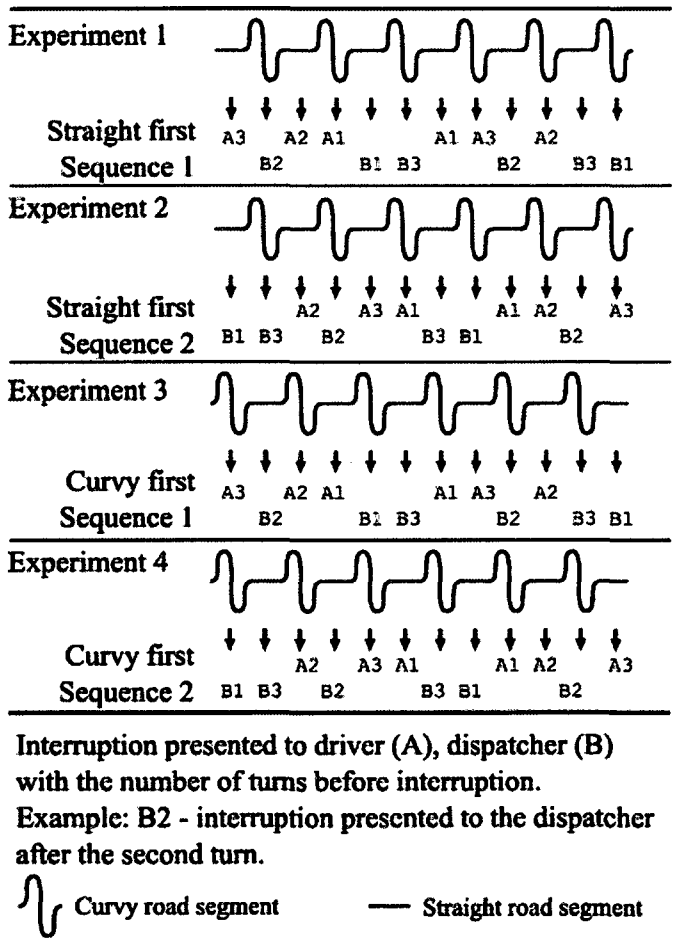


Figure 4.19 Four different experiment sequences (each was done by four subject pairs).

Two types of the starting road segments (curvy and straight) with two different sequences for spoken tasks gave us four different experiment setups that are shown in Figure 4.19. In the experiments 1 and 2 drivers started with driving on a straight road segment, and in the experiments 3 and 4 drivers started with driving on a curvy road segment. Interruption timing for the experiment 1 is the same as for the experiment 3, and interruption timings for the experiment 2 is the same as for the experiment 4. Notice that the order of interruption timings for sequence 2 is the reverse of sequence 1, as explained before. Experiment 1 and 3 used one sequence and Experiment 2 and 4 used the other sequence of words. This means that all pairs of words were tested against different road conditions. For example, twenty questions games with Can opener and Fan was played while driver drove on a curvy road in one experiment and while driver drove on a straight road during another experiment. Each subject pair was assigned a single experiment sequence, so that each of these four experiment sequences were done by four different subject pairs.

4.8 Dependent variables for the spoken tasks

The following sections describe dependent variables for the spoken tasks. The dependent variables for the ongoing and the interrupting task allow us to test hypotheses 1 and 3, which focus on, respectively, how the spoken task performance changes while driving, and how timing of a switch influences the spoken tasks. Modeling switching between the tasks allows us to test hypothesis 4, which focuses on switching behaviors.

4.8.1 Ongoing task

A twenty questions game (ongoing task) can have one of three outcomes: correct object is named (win), incorrect object is named (fail), and the subject runs out of time (timeout). When the word is properly guessed we consider the game to be successfully completed. The ongoing task had the following dependent variables: game outcome, number of turns in a game, pause length before asking a question, length of the utterance containing a question, pause length before providing an answer, length of the utterance containing an answer, and speaking rate for the question and the answer.

Figure 4.20 shows measurements for every turn of the ongoing task. Speaking rate was calculated as number of syllables per second for every word in an utterance and then it was averaged to get a single value for the complete utterance for the question and answer in the turn. Measurements for every variable for every turn in a game were averaged to obtain a single variable value for the game. For example, question pause measurements were averaged over every turn in a twenty questions game to obtain the question pause measurement for this game. Game outcome, number of turns in a game, and the averaged turn variables were averaged to obtain a single measurement for the subject. For example, number of turns in a game was averaged over the twelve twenty questions games to obtain a single measurement for the subject.

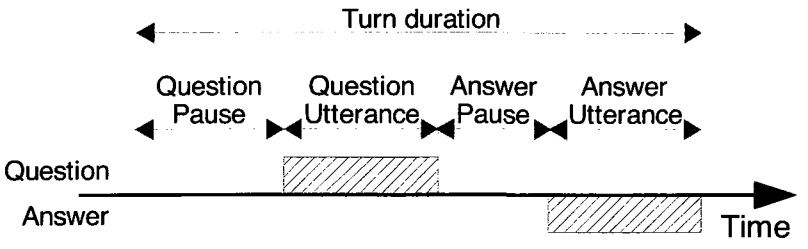


Figure 4.20 Twenty questions game turn related dependent variables.

We considered a turn everything from the end of the previous turn or beginning of the first utterance for the very first turn of the game, until the end of the answer for this turn or beginning of interruption if the turn was interrupted. We consider the last complete sentence that formed a question as a question utterance, and the last complete sentence that formed an answer as an answer utterance. Time from the beginning of the turn until the beginning of the question is considered the question pause. Time from the end of the question utterance to the beginning of the answer utterance is considered the answer pause. Figure 4.21 shows how we defined the turn measurements in a speech sequence.

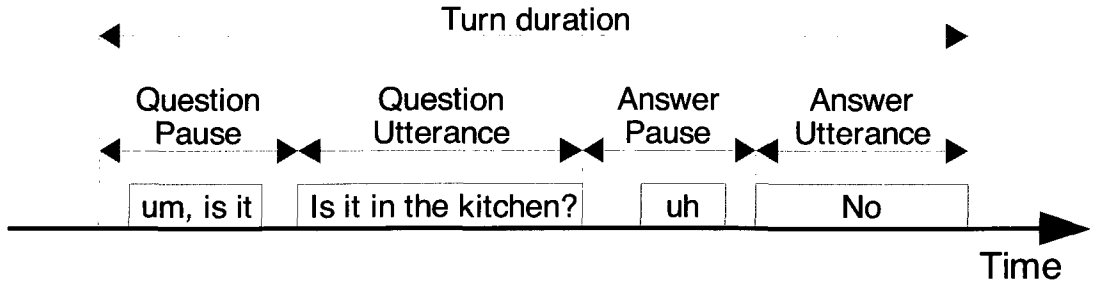


Figure 4.21 Example of twenty questions game turn measurement assignment.

4.8.2 Interrupting task

The interrupting task (last letter word game) had the following dependent variables: pause to provide a word, length of the utterance containing a word, number of turns (words named), and speaking rate. We consider the last word named during the current turn as the utterance. Speaking rate was calculated as number of syllables per second for every utterance. Time from the beginning of the turn to the beginning of the utterance is considered a pause.

Figure 4.22 shows variables for the interrupting task for every turn, and Figure 4.23 shows how we defined these measurements in a speech sequence. These measurements along with the speaking rate were averaged among the turns of a single game to obtain a single measurement for a particular game. For example, pauses for all turns of an interrupting task were averaged to obtain a single measurement for this game. The number of turns in a game and the averaged turn measurements were averaged to obtain a single measurement for a subject. For example, number of turns in a game was averaged over the twelve interruptions to obtain a single measurement.

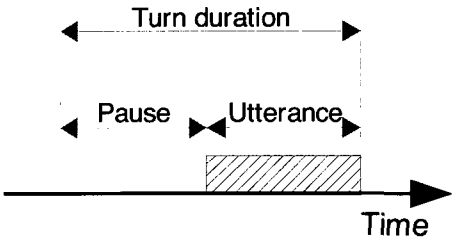


Figure 4.22 Last letter game turn related dependent variables.

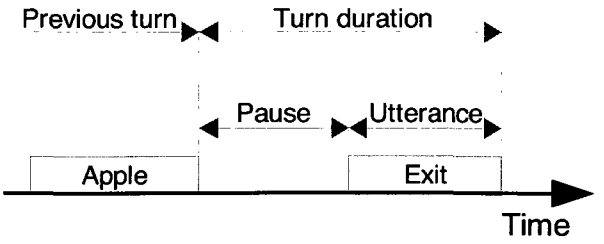


Figure 4.23 Example of last letter word game turn measurement assignment.

4.8.3 Switching between the tasks

Based on our pilot studies we modeled switching between two spoken tasks using the following scheme. First the ongoing task that is the twenty questions game (TQG) is interrupted by initiating a switch to the last letter word game (LLG). Once both

parties agree that the LLG is complete the switch to TQG is performed and TQG is continued. This model is shown in Figure 4.24.

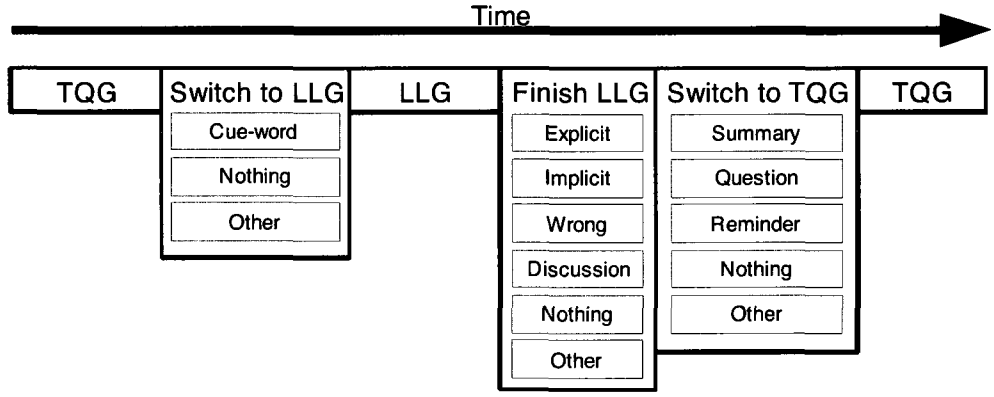


Figure 4.24 Interruption/resumption of a twenty questions game.

As shown in Figure 4.24, when TQG is interrupted to switch to LLG, the interrupting person can take one of the following actions: use a cue-word to indicate the interruption (Okay, Wait, Sorry, etc.) or start the interruption without a cue-word (Nothing). Which cue word is used characterizes a switch from the ongoing to the interrupting task. This parameter is associated with the person who is initiating the interrupting task.

Once the interrupting task is completed, both participants must agree that it is indeed complete. This can be done by a combination of the following: explicitly acknowledging the end of the interrupting task, for instance "We are done" or "That's my three"; implicitly acknowledging the end of the interrupting task, for instance "Okay"; wrongly acknowledging the end of the interrupting task, for instance "We are done, oh, I have another word"; discussing if the interrupting task is complete, by posing a question, for example "Are we done?"; or no acknowledgment that the interrupting task is done by simply resuming the ongoing task. These parameters are associated with both

participants. Each participant could choose how to signal the completion of the interrupting task, for example, the driver might say "We are done" (explicit confirmation) and the dispatcher might say "Okay" (implicit confirmation).

When the interrupting task is complete the context of the ongoing task could be restored. This can be done by: providing a summary of one's own state, for instance "I was in the living room"; asking a question, for example "Was I in the living room?"; reminding what the state of the other participant was, for instance "Yours have a door"; or no context restoration. These parameters are associated with both participants. Each participant could choose how to restore the context, for example, the driver might say nothing (no context restoration) and the dispatcher might say "I am in the living room, you are in the kitchen" (summary and reminder).

4.8.4 Interruption initiation

Following our prior work [11] described in Chapter 3, the ongoing task is modeled as a sequence of adjacency pairs [23]. Section 3.7 (pg. 37) has detailed explanation of our modeling for adjacency pairs. Figure 4.25 shows the summary of the model.

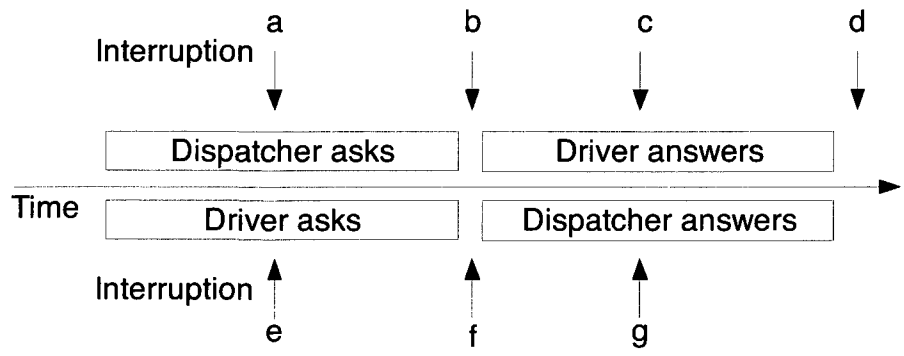


Figure 4.25 Interruption timing.

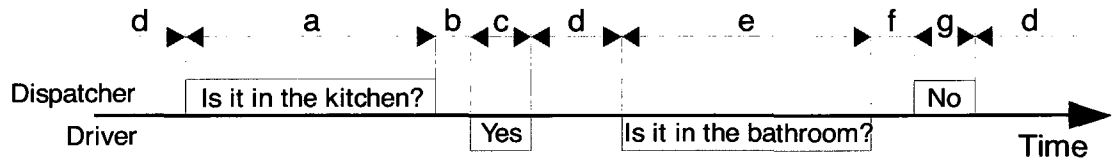


Figure 4.26 Example of codes assigned to adjacency pairs.

Figure 4.26 shows an example of how timings are assigned to a segment of speech. Before the dispatcher asks a question, there is no communication and it is "d" part of the adjacency pair. When the dispatcher asks a question "Is it in the kitchen?" it is "a" part of the adjacency pair. Pause before the driver provides response is marked as "b", and the driver's response itself is "c". Now the first adjacency pair is done and in between the adjacency pairs we have pause "d". When the driver asks "Is it in the bathroom?" it is "e" part. This part is followed by the pause "f" before the dispatcher provides the answer "No", which is "g" part. This is the end of the second adjacency pair.

4.9 Dependent variables for driving

The DriveSafety DS-600c driving simulator allows us to record standard driving measures, such as lane position, vehicle velocity, steering wheel angle, and distance to the leading vehicle at 60 Hz. We calculated variances for each measure. The detailed description of lane position, vehicle velocity, and steering wheel angle variables is given in section 3.8 (pg. 39).

Distance to the leading vehicle is the distance between the center of the leading vehicle and the center of the simulated vehicle and is measured in meters. Higher variance characterizes poor driving performance, since it indicates that the participant did not keep a constant distance from the leading vehicle.

All variables were assigned to corresponding road segments and tasks that were performed during these segments. After that the average was found for these variables. For example, all curvy and straight roads have their averaged values, which allow us to compare driving performance on curvy and straight roads. At the same time, as shown in Figure 4.15 (pg. 71), every curvy and straight segment contained a duration of time when the subjects played the twenty questions game before an interruption, when the subjects played the last letter word game, and when the subjects played the twenty questions game after an interruption. Variables were also averaged for these three distinct regions for every road segment to obtain averages for before, during, and after interruption task segments.

4.10 Experiment procedure

The Experiment Wizard application [78] was used to set up and run the experiment. The following steps were taken during the experiment:

1. Subject preparation: consent forms, questionnaires, and introductions;
2. Training for the twenty questions game (not parallel games): 4 games each;
3. Training for the last letter game: 4 games;
4. Training for playing the twenty questions games in parallel interrupted by last letter word game: 2 games, 4 interruptions;
5. Training for driving and playing the games: 3 games, 3 interruptions;
6. Experiment: 12 games, 12 interruptions;

7. Subject release: questionnaires, debriefing, and reward.

Subjects were presented with computerized questionnaires using the LimeSurvey software [79] before and after the experiment. The text of the questionnaires can be found in Appendix B. The text of the game instructions as given to the participants can be found in Appendix C.

Training included nine twenty questions games, which ensured that subjects played using all the allowed objects. This was done to help the subjects learn the objects. During training the first four twenty questions games were done sequentially, meaning that only one person would ask questions and the other would only answer. After a game was done the roles were reversed. The last five training games were done in parallel as they would be done during the experiment.

Each experiment lasted about 1.5 hours, including paper work, subject training, data collection, and debriefing. Data were recorded on average for about 35 minutes, during which the driver traveled for about 47km.

4.11 Subjects

The recruitment was performed using flyers and e-mails on university mailing lists. The fliers were handed out in personal contacts and posted on bulletin boards at the Durham campus of the University of New Hampshire. The electronic version of the flyer was sent out to the student mailing list of the Electrical and Computer Engineering Department and to the Graduate School of the University of New Hampshire.

The experiment was completed by 32 participants (16 pairs) between 18 and 38 years of age. Each pair was formed by two people who have never met each other

before. The average age of the participants was 24 years and 28% were female. Subjects were promised a $15 compensation for participating in the experiment. They were also told that if they perform well (attempt to finish all the games and interrupting tasks according to the rules) they would be given a bonus of $5. By providing a monetary incentive we tried to motivate subjects to perform well during the experiment. All subjects were given the bonus regardless of their performance. The reward was given as gift card certificates.

CHAPTER 5

RESULTS AND DISCUSSION FOR THE TWENTY QUESTIONS EXPERIMENTS

This chapter describes the data, data analysis methods, and results, as well as the discussion of the results obtained during the twenty questions experiments described in the previous chapter. This experiment was designed to answer the following questions (hypotheses described in section 1.3, pg. 7): Does driving influence performance of the spoken tasks? Does timing of switching between the spoken tasks affect the spoken tasks? Do the spoken tasks affect driving performance? What switching behaviors are exhibited by the drivers and the dispatchers? How do subjects resume the interrupted ongoing task? The following sections show the data we used and the methods we employed to answer these questions.

5.1 Corpus and tools

The experiment was completed by 32 participants (16 pairs) between 18 and 38 years of age. Each pair was formed by two people who have never met each other before. The average age of the participants was 24 years and 28% were female. During

the experiments we collected 9.3 hours of speech interactions with synchronized simulator and eye tracker data. The driving and eye-tracker data were collected over 800km traveled.

We choose to use 16 subject pairs, because we had four different experiment setups (section 4.7, pg. 75) and we decided that each experiment setup had to be done by multiple subject pairs. In general, a sample size of less than 16 experiments was commonly used in previous research involving driving simulators [10,48,62].

We collected data from 384 games (12 games for 32 subjects) for the ongoing task. Half of these games (192) were played by the drivers and the other half by the dispatchers. The same statistic applies to the interrupting task with 384 games. During the experiments 25% of the time the subjects were saying something to each other. The audio files were annotated in order to extract the values for dependent variables (section 4.8, pg. 79). Data annotation was done by the author. In addition, two undergraduate students participated in the annotation of the switching behavior. The disagreements in the transcription of the switching behavior were resolved by consensus. The corpus contains 5752 utterances (about 360 utterances per experiment and 180 utterances per subject).

Speech Viewer from CSLU toolkit 2.0 was used for audio data annotation. Speech recordings were transcribed by hand. Every utterance in the ongoing task was assigned a game number (1 to 12) and a turn number (1 to 10, as explained in sections 4.3.1 and 4.4. Every game was marked with the outcome (win, timeout, fail). Every turn was marked as: being normal (question/answer pair), or containing a switching activity, such as resumption, reminder, etc. (as explained in section 4.8.3, pg. 82), or interrupted (an interrupting task was initiated during this turn). Unless the turn was interrupted, it had

four parts as shown in Figure 4.20: pause before the question, question utterance, pause before an answer, and answer utterance. In addition, speaking rate was calculated for the question and answer utterances. Section 4.8.1 (pg. 80) explains how we define these measures. Every question in the ongoing task was assigned a level one to four based on the explanations in section 5.1.1 (pg. 94).

Every interruption game was classified with the number of the last complete turn before the interruption, and the level of the question in the last complete turn before the interruption. In addition, every interruption had two codes attached to it: when the interruption was visually presented (shown to a subject), and when the interruption was initiated (the subject initiated the interruption). These codes indicated when the interruption occurred in relation to the closest adjacency pair. Section 4.8.3 (pg. 82) provides more explanations of these codes along with examples.

For every switch from the ongoing task to the interrupting task we marked the switch as containing or not containing a cue word (no other methods of switching were observed). For every switch from the interrupting task to the ongoing task we marked the switch as containing summaries, reminders, questions, no activity, or something different from all the previous activities.

Speaking rate was calculated with help of Tcl scripts provided by Peter Heeman. These scripts used CSLU toolkit to find the syllables and their durations in the annotated data. The scripts were used previously by Yang et al. [18]. Driving performance measures were extracted using SEAT application developed by Oskar Palinko for internal use in Project54.

SPSS Statistics 17.0 (now called PASW Statistics) was used to perform statistical analysis of the data. The drivers and the dispatchers worked together during the experiments, and, consequently, their performance measures cannot be considered independent. Because measures for the drivers and the dispatchers depend on each other, we obtained dependent samples, therefore, we decided to conduct a paired (dependent) t-test for comparing measures for the drivers and the dispatchers [80-82] (also see section 5.8, pg. 149). We also used ANOVA repeated measures to compare measures related to the same subjects, for example, when comparing driver's performance on curvy and straight roads. The post hoc analysis was adjusted for multiple comparisons using Fisher's protected LSD test.

Code	Speaker	Utterance	Details	Task
U1	Dispatcher	Is it in the kitchen?	Dispatcher's turn 1	TQG
U2	Driver	No.		TQG
U3	Driver	Does it have sharp edges?	Driver's turn 1	TQG
U4	Dispatcher	No.		TQG
U5	Dispatcher	Is it in the bathroom?	Dispatcher's turn 2	TQG
U6	Driver	No.		TQG
U7	Driver	Does it produce heat?	Driver's turn 2	TQG
U8	Dispatcher	No.		TQG
U9	Dispatcher	Is it on the ceiling?	Dispatcher's turn 3	TQG
U10	Driver	No.		TQG
U11	Dispatcher	Letter, word beginning with B	Interrupting task	LLG
U12	Driver	Ball.	Driver's turn 1	LLG
U13	Dispatcher	Like.	Dispatcher's turn 1	LLG
U14	Driver	Kite.	Driver's turn 2	LLG
U15	Dispatcher	Time.	Dispatcher's turn 2	LLG
U16	Driver	Move.	Driver's turn 3	LLG
U17	Dispatcher	Voice.	Dispatcher's turn 3	LLG
U18	Driver	Okay.	Implicit signal	Switch
U19	Dispatcher	Your turn to ask.	Reminder	Switch
U20	Driver	Does it have a door?	Driver's turn 3	TQG
U21	Dispatcher	Yes.		TQG
U22	Dispatcher	Does it produce sound?	Dispatcher's turn 4	TQG
U23	Driver	Yes.		TQG
U24	Driver	Does it preserve food?	Driver's turn 4	TQG
U25	Dispatcher	Yes.		TQG
U26	Dispatcher	Does it produce picture?	Dispatcher's turn 5	TQG
U27	Driver	Yes		TQG
U28	Driver	Is it the refrigerator?	Driver's turn 5	TQG
U29	Dispatcher	Yes		TQG
U30	Dispatcher	Is it the TV?	Dispatcher's turn 6	TQG
U31	Driver	Yes.		TQG

Table 5.1 The ongoing task with the interrupting task for game 3, subject pair 11.

Table 5.1 shows an example of one game (game 3, subject pair 11). The interruption is presented to the dispatcher. This example was chosen to illustrate that sometimes subjects negotiated (3 out of 16 subject pairs) that the dispatcher will always ask the first question about the room where his own object is. This way the driver did not have to ask a question about a room. The negotiation happened during the training period.

5.1.1 Assigning interruption levels

The design of the twenty questions game is such, that not all game questions progress a subject through the game equally. For example, it is possible to find out what room an object is after the first question or after the third question. This means that amount of information that must be retained during the interrupting task about twenty questions game could be the same if the person is interrupted after the first question or after the third question. We assume that the amount of information that must be retained increases the cognitive load, which in turn, might affect the performance measures for the spoken tasks or driving. Thus, we decided to keep track of where in the game a person is using levels assigned to every question as described below. We structured the twenty questions game so that the subjects had to discover the room with the object first (we call this level 1 question), then the general function of the object (we call this level 2 question), then the particular feature of an object (we call this level 3 question), and the final question is to guess the object (we call this level 4 question). Four questions is the minimum number of questions required to discover an object if the twenty questions game is played by our rules. Levels must not be skipped and therefore all four levels should be represented with at least a single question. For example, if a "microwave" is the object to discover, then the shortest set of questions/answers could be (following the training tree in Figure 4.9, pg. 62) such as shown in Table 5.2.

Code	Speaker	Utterance	Details
U1	Person A	Is it in the kitchen?	Level 1
U2	Person B	Yes.	
U3	Person A	Is it used for heating?	Level 2
U4	Person B	Yes.	
U5	Person A	Does it have a door?	Level 3
U6	Person B	Yes.	
U7	Person A	Is it a microwave?	Level 4
U8	Person B	Yes.	

Table 5.2 The shortest set of question/answers in a twenty questions game.

Within each level there can be three or two possible questions (as given by the training tree in Figure 4.9). The participant must guess what question to ask first for every level. Thus, the longest set of questions without repeated questions would be nine questions. For example, if the object is a "hair trimmer" and the participant follows the training tree from top to bottom, then the sequence of questions/answers shown in Table 5.3 would occur.

Code	Speaker	Utterance	Details
U1	Person A	Is it in the kitchen?	Level 1
U2	Person B	No.	
U3	Person A	Is it in the living room?	Level 1
U4	Person B	No.	
U5	Person A	Is it in the bathroom?	Level 1
U6	Person B	Yes.	
U7	Person A	Is it for personal use?	Level 2
U8	Person B	No.	
U9	Person A	Is it a utility?	Level 2
U10	Person B	No.	
U11	Person A	Is it used on hair?	Level 2
U12	Person B	Yes.	
U13	Person A	Does it use heat?	Level 3
U14	Person B	No.	
U15	Person A	The object does not use heat?	Level 3
U16	Person B	Yes.	
U17	Person A	Is it a hair trimmer?	Level 4
U18	Person B	Yes.	

Table 5.3 The longest set of questions in a twenty questions game.

Participants can deduce that if they asked questions about two out of the three rooms and they received "No" as answers, then the third room is the only choice and there is no need to explicitly ask if that is the room. Such an approach would reduce the longest sequence of questions from nine to six.

In general, we used the following rules to determine a level of the question:

1) Level 1 questions are related to rooms. For example, "Is it in the kitchen?"

2) Level 2 questions differentiate between two groups of objects. For instance, "Does it have a door?" There is a group of objects that has a door and another group that does not;

3) Level 3 questions differentiate between two objects. For example, "Does it use sound and picture?" This question differentiates between TV and Radio;

4) Level 4 questions are about a particular object. For instance, "Is it a mixer?"

We used the level of the question from the last complete turn to assign the level to an interruption. For example, if the last complete turn had question "Does it have a door?", then the interruption was assigned as happening at level 2.

5.2 Design verification

During the data processing we first set out to confirm that the ongoing and interrupting tasks were performed by the participants as we intended them to be performed. Specifically, we wanted to confirm that the number of turns in the ongoing task was around six according to the game design (section 4.3, pg. 58). Figure 5.1 shows the distribution of the number of turns in the ongoing task. This plot shows that out of

384 games only 2.6% (10) of the games had less than four turns and only 4.4% (17) of the games had more than nine turns. This is consistent with the twenty questions game design as explained in section 4.3 (pg. 58).

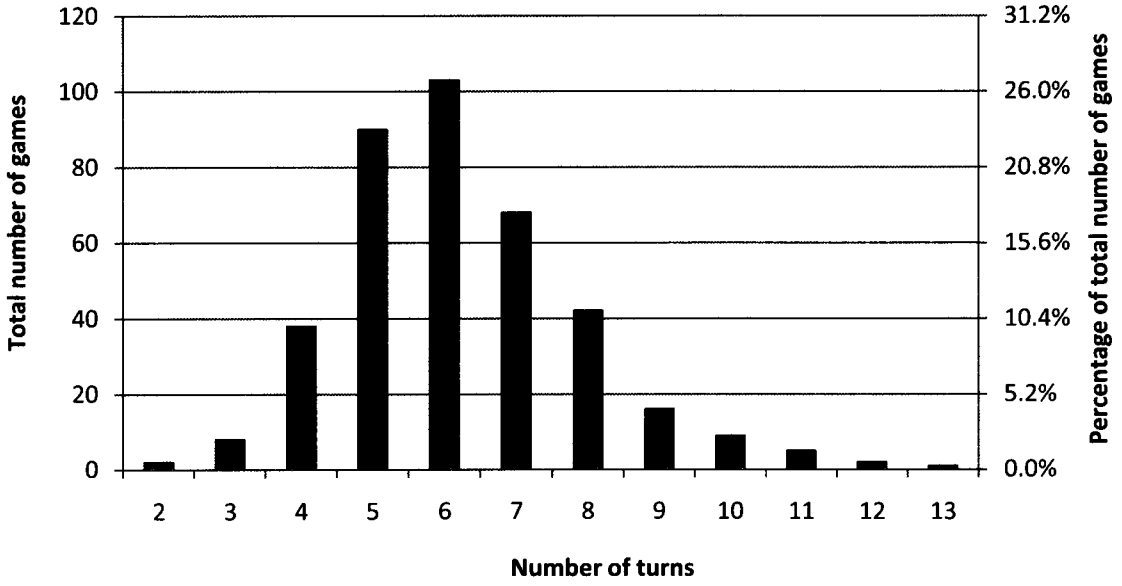


Figure 5.1 Distribution of number of turns in a twenty questions game.

Similarly, we wanted to confirm if the interrupting task was played according to the rules of the last letter word game. The interrupting task required participants to have three turns each. Figure 5.2 shows the number of turns in the interrupting task. We can see that the majority (87%) of the games were done according to the rules (section 4.4, pg. 66).

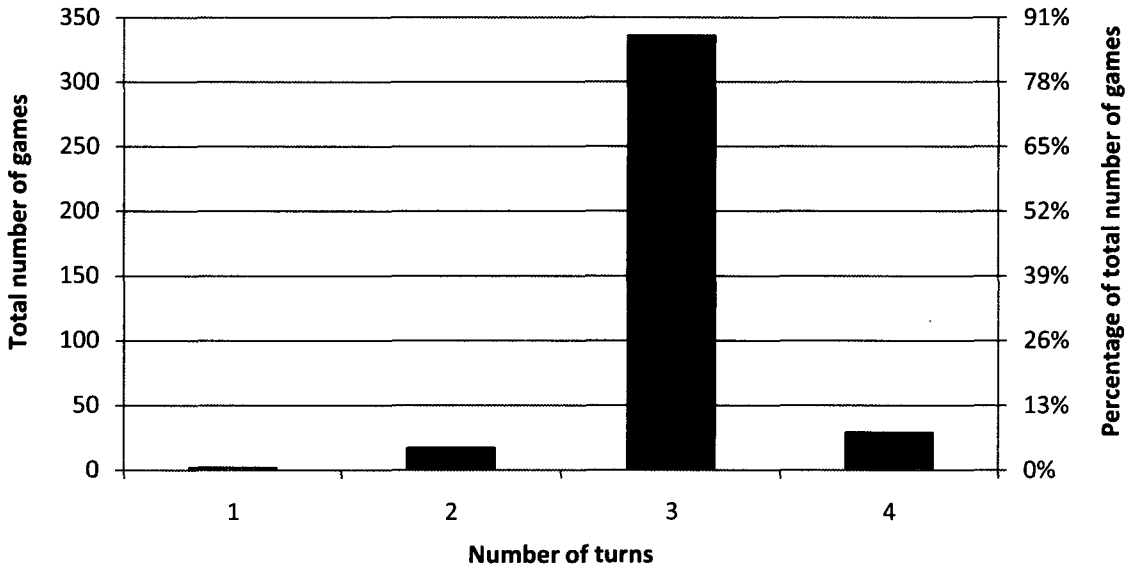


Figure 5.2 Number of turns in a last letter word game.

On average the drivers and the dispatchers finished playing their TQG in 62 seconds and LLG in 28 seconds. These values indicate that two minute time allocated for the games was sufficient for most of the subjects. This is consistent with the experiment design as described in section 4.5 (pg. 70). Table 5.4 lists mean values with their standard deviations for some dependent variables.

Variable Name (unit)	**Drivers**		**Dispatchers**	
	Mean	STD	Mean	STD
TQG pause before asking a question (s)	1.87	±0.88	1.47	±0.88
TQG question utterance duration (s)	1.53	±0.32	1.45	±0.36
TQG pause before answering a question (s)	0.74	±0.28	0.78	±0.18
TQG answer utterance duration (s)	0.55	±0.13	0.58	±0.14
LLG pause before naming a word (s)	5.49	±1.78	5.23	±1.51
LLG utterance duration (s)	0.68	±0.23	0.71	±0.32
TQG number of turns	6.07	±0.95	6.36	±0.82
LLG number of turns	3.02	±0.13	3.03	±0.09
TQG question speaking rate (syllables/s)	8.07	±1.31	8.45	±1.30
TQG answer speaking rate (syllables/s)	2.60	±1.00	2.79	±0.93
LLG speaking rate (syllables/s)	2.80	±0.70	2.87	±0.61
Delay from interruption presentation to interruption initiation (s)	2.59	±0.14	2.35	±0.16

Table 5.4 Average values and standard deviations for some dependent variables.

We did not have precise control over the timing of the interruptions with respect to the progress of TQGs, because different subjects progressed through the ongoing task with different speeds (see section 4.4, pg. 66 for detailed explanation). Figure 5.3 shows how interruption timings were distributed for the dispatchers and the drivers. The differences in the distributions are due to the fact that the dispatcher always started the game first (all dispatchers were instructed to do so). Hence, it was very unlikely for them to be interrupted right after the first turn. Overall, the distribution does cover the points of interest for us, which are interruptions after turns two, three, and four as explained below.

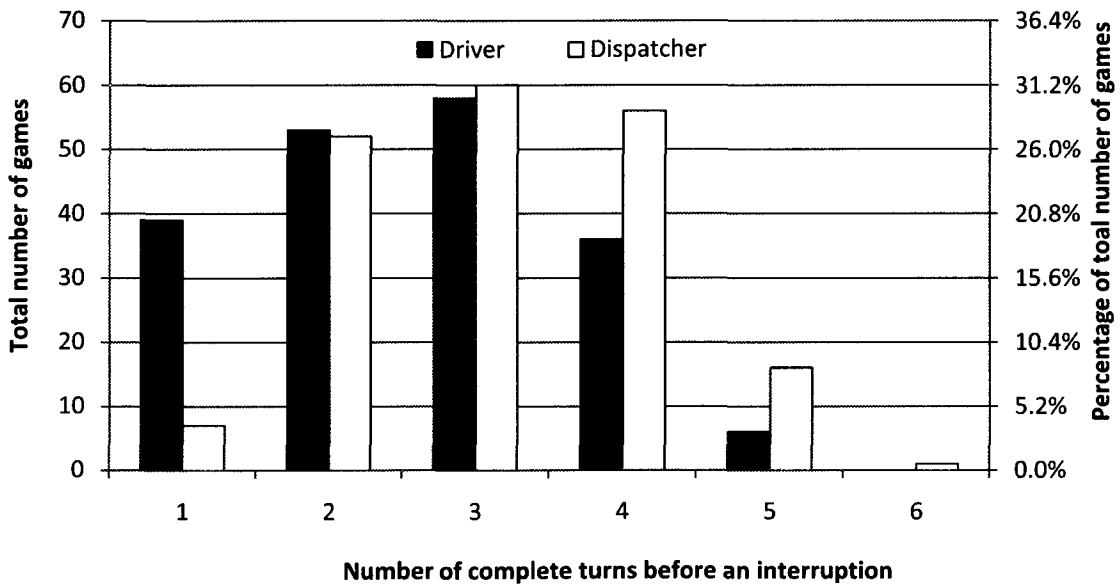


Figure 5.3 Distributions of number of turns before an interruption.

We hypothesized that the subjects build up the context with the progression of the ongoing task. As a result, the interruptions of the ongoing task with different amount of context might be treated by the subjects differently. We labeled interruptions that happen between turns two and three as *early*, interruptions that happen between turns

three and four as *middle*, and interruption that happen between turn four and five as *late*. To clarify, the same interruption may be marked as middle for the dispatcher and early for the driver, depending on when it happened during the twenty questions game. For example, if both the dispatcher and the driver completed their second turn and an interruption happened, then both of the participants have a game with the early interruption. On the other hand, if the dispatcher completed the third turn, but the driver did not, then the interruption is marked as middle for the dispatcher and as early for the driver.

Games with the interruptions before turn two (3.6% of the data) or after turn five (8.8% of the data) were discarded during the analysis that involved timing of interruptions. Removal of these interruptions eliminates possible bias. For example, the drivers had more interruptions right after the first turn than the dispatchers did. As a result, uneven number of data points does not allow us to balance effects of subject variability in the data. At the same time, this leaves 87% (336) of the games for comparison. Figure 5.4 shows the distribution of the timing of interruptions for the drivers and the dispatchers (subset of data from Figure 5.3).

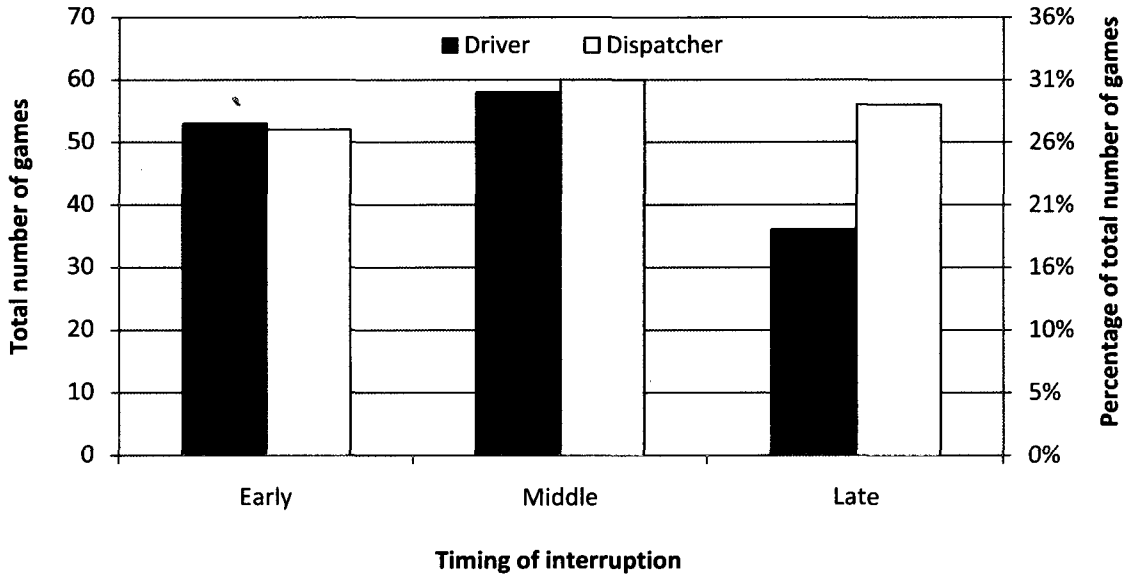


Figure 5.4 Number of games for different timing of interruption.

Figure 5.5 shows the ongoing task outcomes for all 384 games. A total of 296 games (77%) resulted in a successful completion. This shows that the difficulty of the ongoing task was selected in a way that did not cause the subjects to be frustrated about their performance, but at the same time the subjects knew that it was possible to lose games.

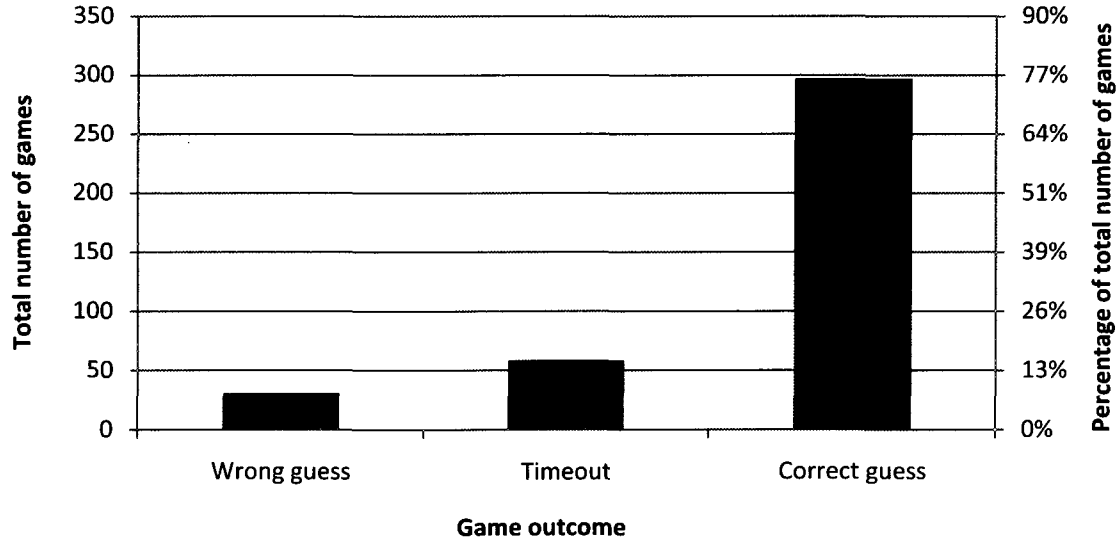


Figure 5.5 Outcomes of the ongoing tasks.

Figure 5.6 shows the average duration of a pause before a question over the game duration (averaged over 384 games). Error bars in this figure and others show standard error unless otherwise noted. We could expect to see the subjects slow down with time if the subjects became tired. Instead we observe that both the drivers and the dispatchers provided responses faster with time, as demonstrated by the slope of the fitted line (driver: R^2=0.19, 11 d.f., p=0.158; dispatcher: R^2=0.66, 11 d.f., p=0.001), which may be due to learning effects.

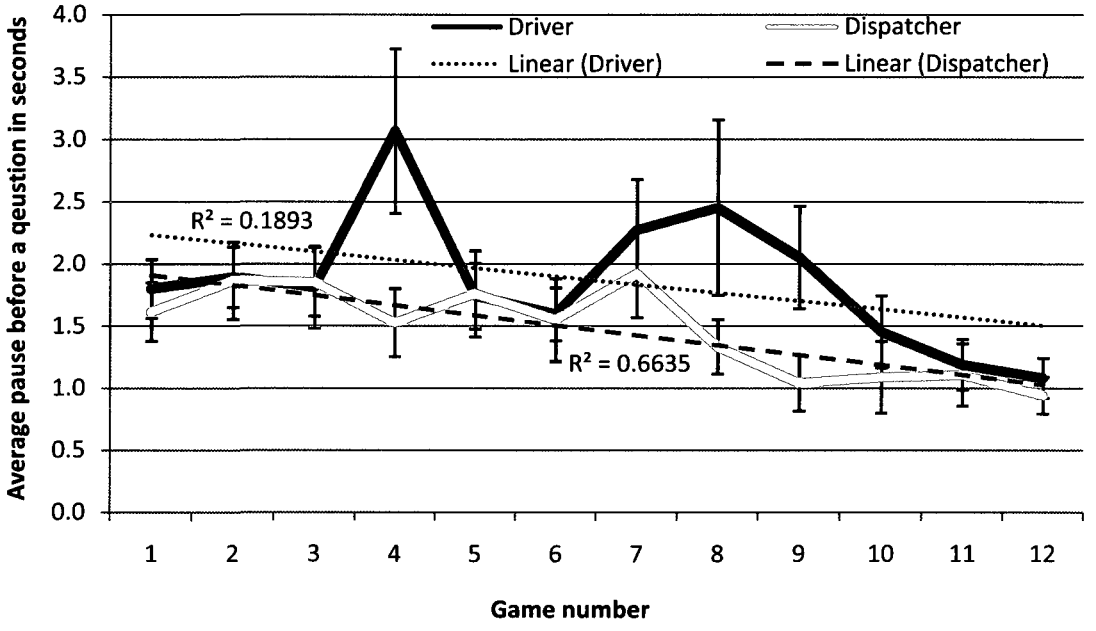


Figure 5.6 Average pause duration before a question over the duration of the experiment with a linear fit.

Figure 5.7 shows the average pause before an answer (driver: R^2=0.53, 11 d.f., p=0.007; dispatcher: R^2=0.37, 11 d.f., p=0.036), which also demonstrates the learning trend. We do not have an explanation for the spikes in the average pause before asking a question, as shown in Figure 5.6. For instance, game four, on average, has the pause duration before asking a question that is significantly different between the drivers and

dispatchers (t(15)=2.6,p=0.02), while we failed to observe any difference between characteristics of game four and other games. Or using a reverse argument, it is not clear why some games have the same pause duration before asking a question for both the drivers and the dispatchers. For instance, game five, on average, have virtually the same pause duration before asking a question (t(15)=0.05,p=0.96).

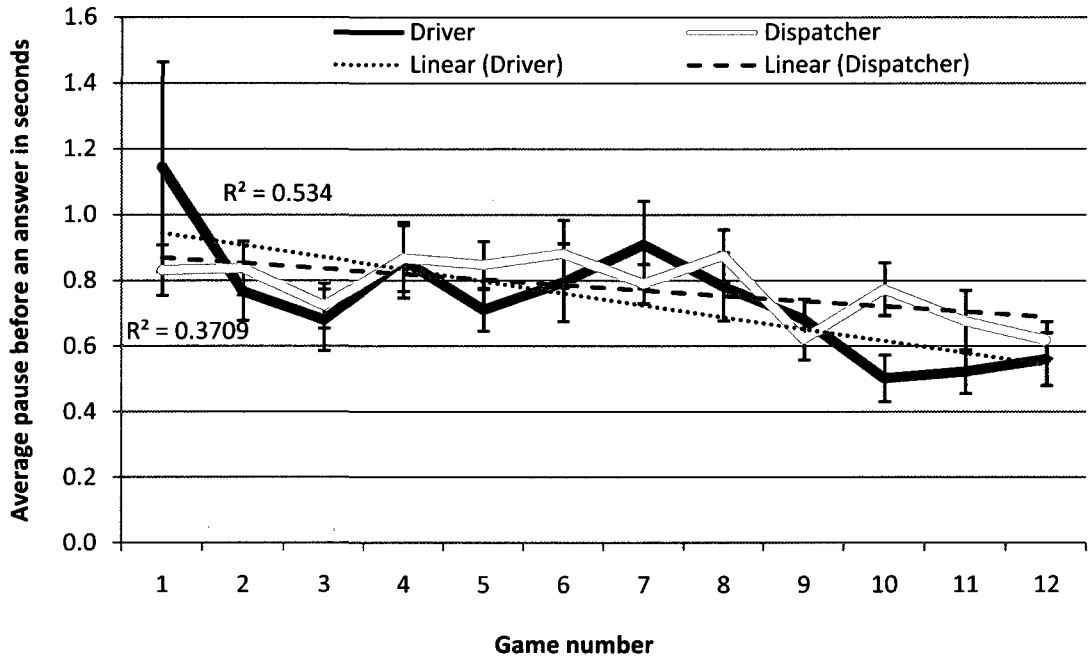


Figure 5.7 Average pause duration before an answer over the duration of the experiment with a linear fit.

In contrast to the learning effects for the ongoing task, Figure 5.8 shows that the averaged pause before naming a word during an interruption (LLG) becomes longer over the duration of the experiment (driver: R^2=0.63, 11 d.f., p=0.002; dispatcher: R^2=0.54, 11 d.f., p=0.007). This can be explained by the fact that the participants had to come up with the words that they did not use before, and, therefore, had to think more. This is consistent with the experiment design.

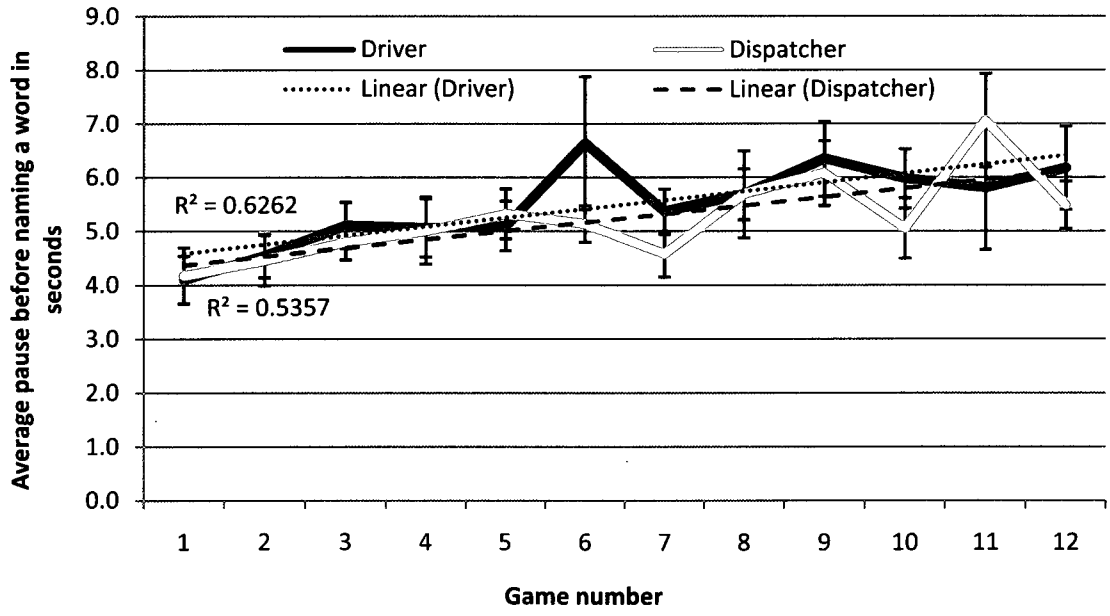


Figure 5.8 Average pause before naming a word (during the interrupting task) over the duration of the experiment with a linear fit.

We also looked at the percent dwell time [6,83] at the road ahead for the drivers using the eye tracker data. We found that 96% of the time the drivers look at the road ahead of them. The other 4% included times when the eye tracker did not track the data, as well as glances at the rear view mirrors and speedometer. There were no additional traffic on the road or other distracting events along the road, and that is why we expected the drivers to look at the road ahead of them most of the time. The eye tracker data confirmed our expectations.

5.3 Performance on the ongoing spoken task

We compared performances of the drivers to the performances of the dispatchers on the ongoing spoken task. This test is driven by hypothesis 1, which focuses on the interaction between the spoken tasks and driving. We hypothesized that there would be differences in the performances due to the fact that the drivers are

engaged in the manual-visual task. The first measure we looked at was the number of successfully completed games for the drivers and the dispatchers. There are three possible outcomes for a twenty questions game: correct guess, wrong guess, or timeout. Figure 5.9 shows the game outcomes for the drivers and the dispatchers. Statistical analysis showed that the differences between the drivers and the dispatchers are not significant ($t(15)<1.373, p>0.19$).

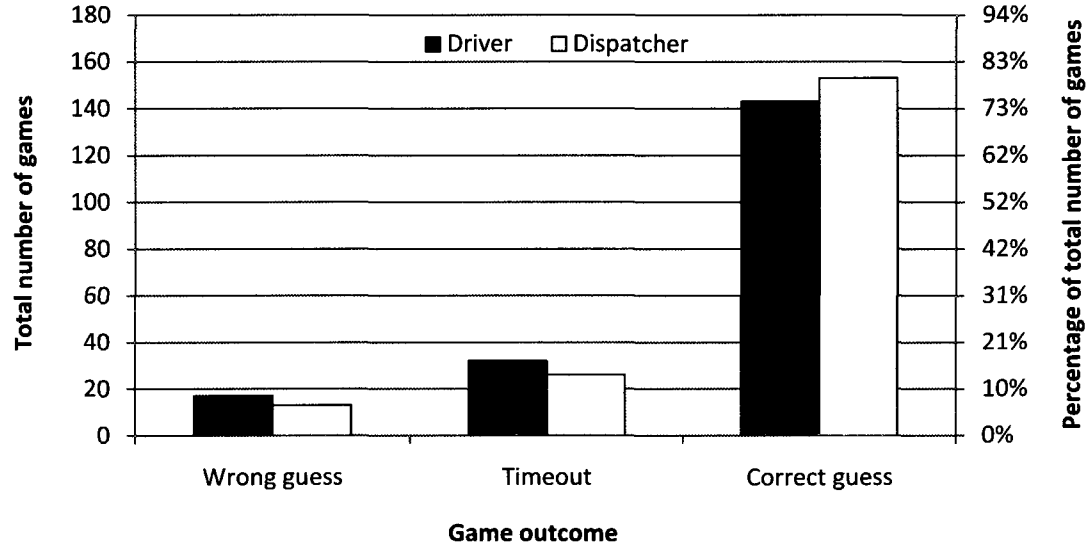


Figure 5.9 Game outcomes for driver and dispatcher.

Figure 5.10 shows how games with wrong guesses were distributed over the 16 subject pairs. It is interesting to notice that 13 out of 16 drivers had at least one game that ended in a wrong guess, while only 7 out of 16 dispatchers had at least one game that ended in a wrong guess. However, statistical analysis did not show that the drivers and the dispatchers have a significant difference in the number of games that ended with a wrong guess. The number of games that end with wrong final guesses is very small (8% or 30 games), and, thus, we focused on games with timeouts and correct guess only (354 games).

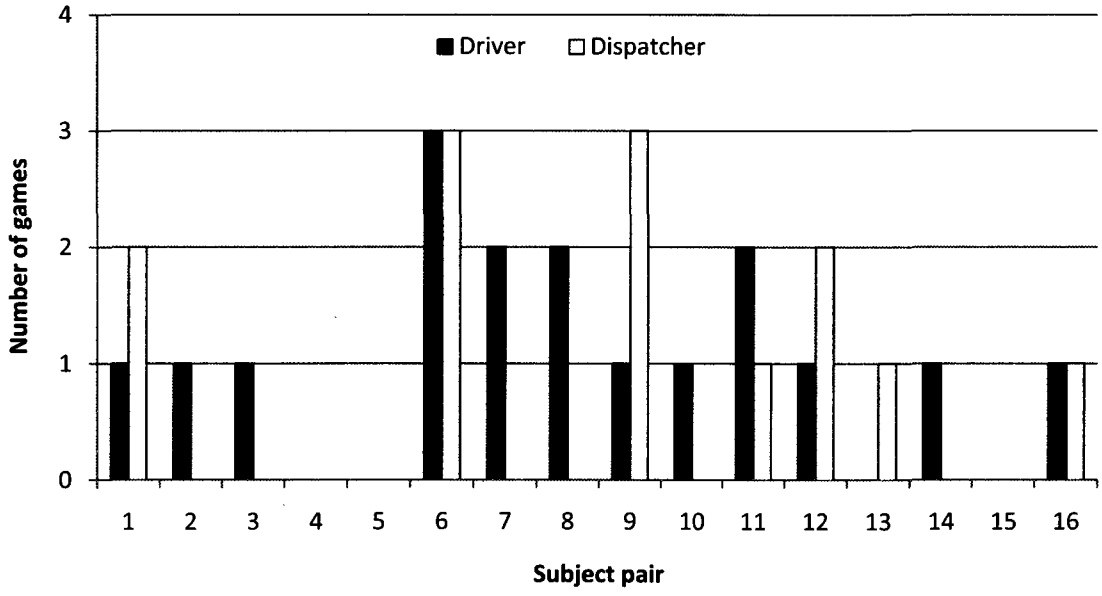


Figure 5.10 Wrong guesses over the experiments for driver and dispatcher.

We were expecting the dispatchers to perform better than the drivers, because we hypothesized that the additional task of driving should not allow the driver to perform the ongoing task as well as the dispatcher could. Figure 5.10 shows that, overall, the drivers won less of their games than the dispatchers did. The trend toward this conclusion is visible in the data, but it is not significant. One possible explanation is that the ongoing task was easy enough for the drivers to perform while driving at the given level of difficulty. Increasing the difficulty of the ongoing or the driving task could emphasize the observed trend. On the other hand, Tsimhoni et al. [9] also found that the driving workload did not influence the spoken task performance. In their experiments, the subjects were listening to the different types of messages (news, email) while driving a simulated vehicle on roads with two difficulty levels (straight segments and constant radius curve segments). After listening to a message the comprehension of the message was assessed by asking subjects a series of questions. The time to answer a question was used as one of the performance measures. The authors did not specify the radius of the

curves they used in their experiments to control the driving difficulty. The spoken tasks in our experiment are different from those used by Tsimhoni et al., but it could be that we are finding similar results.

Similarly, we found that there is no significant difference for the pause duration before asking a question between the drivers and the dispatchers ($t(15)=1.83, p=0.87$). The duration before answering a question was also not significantly different between the drivers and the dispatchers ($t(15)=-0.4, p=0.63$). The interrupting task measures did not show significant differences either, for example, pause before naming a word did not have significant differences for the drivers and the dispatchers ($t(15)=-1.5, p=0.3$). Given the lack of differences between performances on the spoken tasks for the drivers and the dispatchers when all 384 games were treated equally, we decided to see how the timing of interruptions affects the performance measures.

5.3.1 Timing of interruptions by turn number

We decided to split the twenty question games according to the interruption timing to test the hypothesis 3, which states that the timing of interruptions affects spoken tasks. Figure 5.11 shows the percentage of games won for different interruption timings (number of games for different interruption timings is shown in Figure 5.4). The statistical analysis showed that the dispatchers won more of their games when an interruption happens *early* as compared to the games with early interruptions that the drivers won ($t(15)=2.13, p=0.049$). But there is no significant difference for the *middle* and *late* interruptions for the dispatchers and the drivers ($t(15)<1.985, p>0.069$). It is important to notice that the p values for these observations are very close to 0.05,

meaning that it is possible to have false positive for the games with *early* interruptions and false negative for the games with *middle* and *late* interruptions. The next step was to understand why the drivers lose more of their games than the dispatchers when the interruption happened early. This analysis should reveal if the observed difference is indeed present and is not false positive.

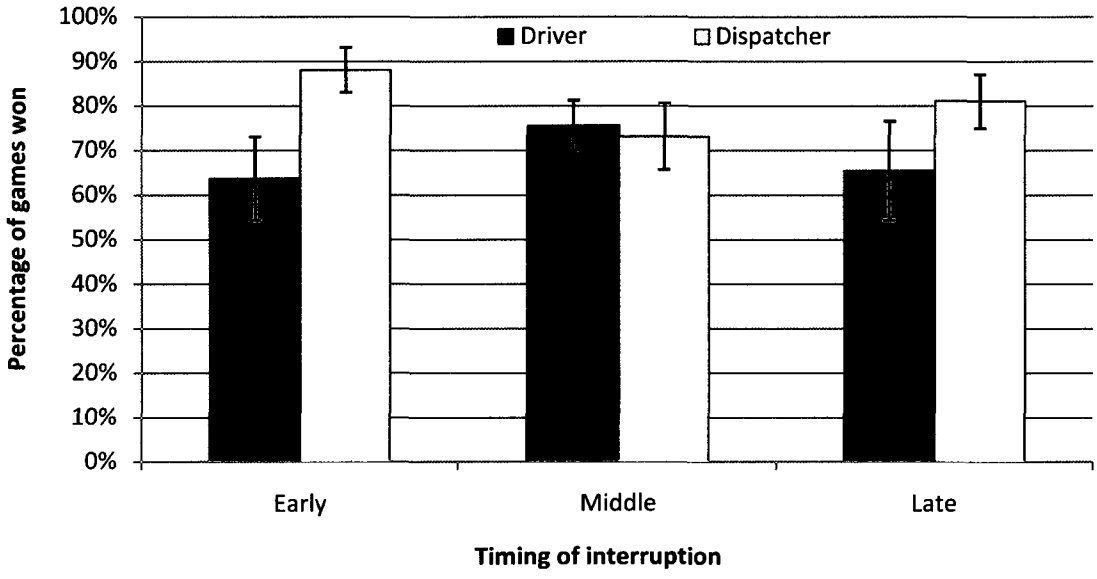


Figure 5.11 Percentage of wins by timing of interruption.

Figure 5.12 shows the average duration of a pause before a question for the drivers and the dispatchers for games when interruptions happened at different times. The difference between the drivers and the dispatchers is significant for games with *early* interruptions ($t(15)=3.1, p=0.007$) and is not significant for games with *middle* ($t(15)=0.5, p=0.637$) and *late* ($t(13)=1.3, p=0.215$) interruptions. The high significance level of the comparison for the games with *early* interruptions indicate that there is indeed a difference between the drivers and the dispatchers and it is not likely to be a false positive. It is interesting to notice that statistical analysis shows that the drivers have different pauses before asking a question ($F(2,13)=4.86, p=0.027$) when the pauses are

compared between different interruption timings (early vs middle p=0.006, early vs late p=0.071, middle vs late p=0.439). In contrast, the dispatchers have the same duration of the pause for all interruption timings (F(2,13)=2.33,p=0.137). This indicates that the timing of the interruption had a larger impact on the drivers than on the dispatchers.

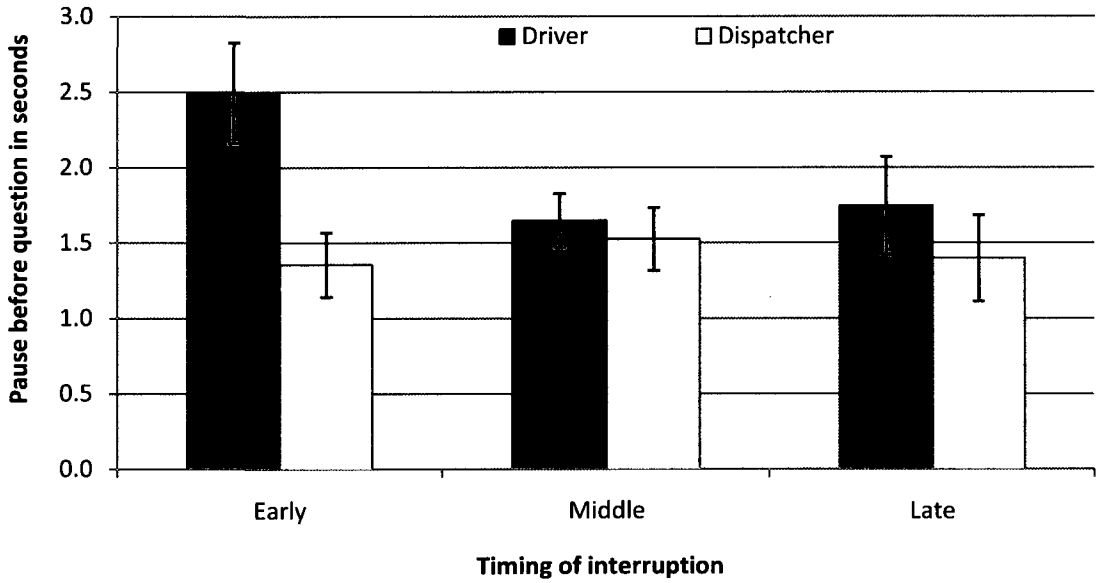


Figure 5.12 Pause before question by timing of interruption.

The number of turns for the ongoing task (t(15)<1.1,p>0.289) and the interrupting task (t(15)<1.7,p>0.108) are not significantly different for the drivers and the dispatchers. Hence, the drivers lose because it takes them longer to ask a question and the drivers run out of time before they can finish the TQG. To test this conclusion we compared the average pause before asking a question between the games that were lost by timeouts and the games that were successful.

Figure 5.13 shows the average pause before asking a question and the average pause before answering a question for the drivers for *early* games only. Statistical analysis showed that there is a significant difference (F(2,29)=20.49,p<0.001) in the

pause before asking a question during games that end with a timeout and games that end with a correct guess. The difference in the pause before answering a question for these games is also significant (F(2,29)=4.74,p=0.017). It is important to notice that for the drivers, as Figure 5.9 shows, there were more games that ended with correct guesses (75% or 143 games) than games ended with timeouts (17% or 32 games). For *early* interruptions only, there are 36 (68% of 53) games that end with a correct guess and 16 (30% of 53) games that end with a timeout. The fact that there are two times as many games with the correct guesses than with the timeouts might bias the results, because the smaller data set may not capture the possible range of individual variations between the subjects. Nevertheless, the trend is clearly visible.

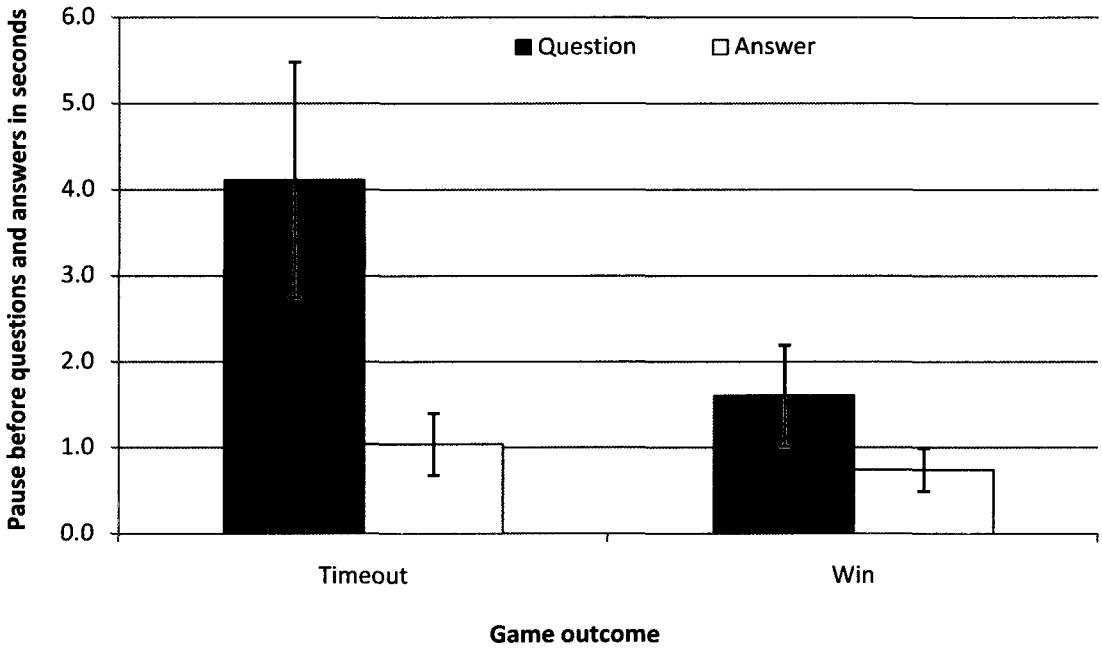


Figure 5.13 Pause before driver's questionsand answers for games interrupted early.

Similar analysis was performed for the dispatchers. Figure 5.14 shows the average pause before asking a question and the average pause before answering a question for the dispatchers during the games with *early* interruptions only. Statistical

analysis showed that there is a significant difference (F(2,22)=5.37,p=0.009) in the pause before asking a question during games that ended with timeouts and games that ended with correct guesses. The difference in the pause before answering a question for these games is not significant (F(2,22)=2.6,p=0.095). Again, it is important to notice that for the dispatchers, as Figure 5.9 shows, there are more games that ended with a correct guess (80% or 153 games) than games ended by a timeout (14% or 26 games). For *early*, interruptions there were 45 (87% of 52) games that ended with correct guesses and only two (4% of 52) games that ended with timeouts. The small number of games that end with a timeout does not capture the range of individual variations between the subjects, and, and for this reason cannot be used to draw a definite conclusion.

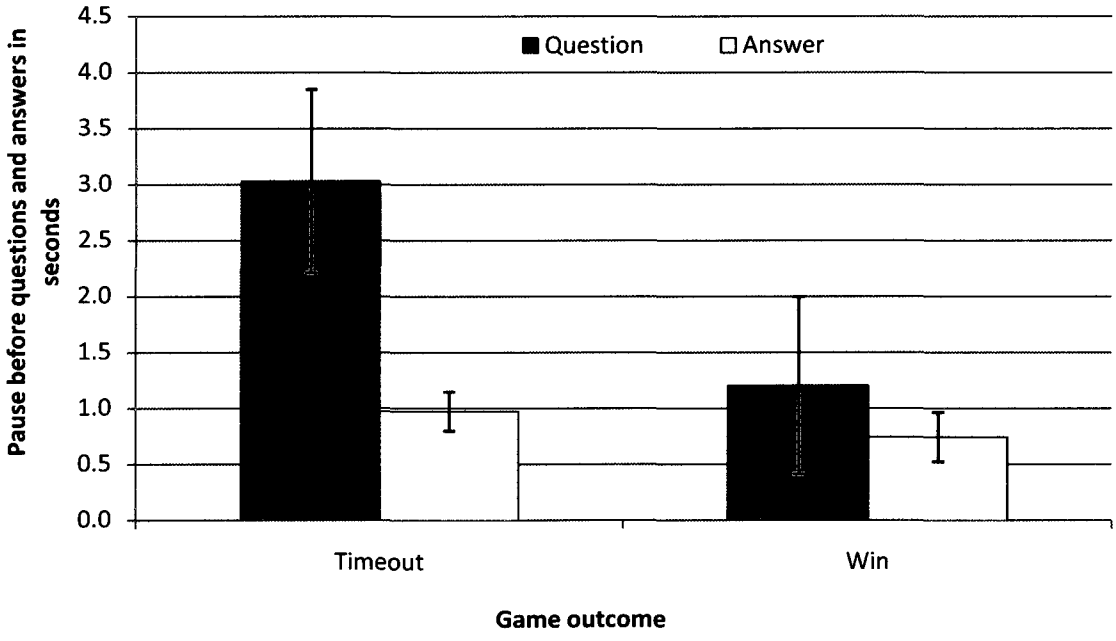


Figure 5.14 Pause before dispatcher's questions and answers for games that were interrupted early.

Figure 5.15 shows the pause before naming a word in the interrupting game depending on the timing of the interruption. The data suggests that for the *early*

interruptions it could take longer for the drivers to name a word for the interrupting task, but this difference is not significant ($t(14)<1.1, p>0.286$).

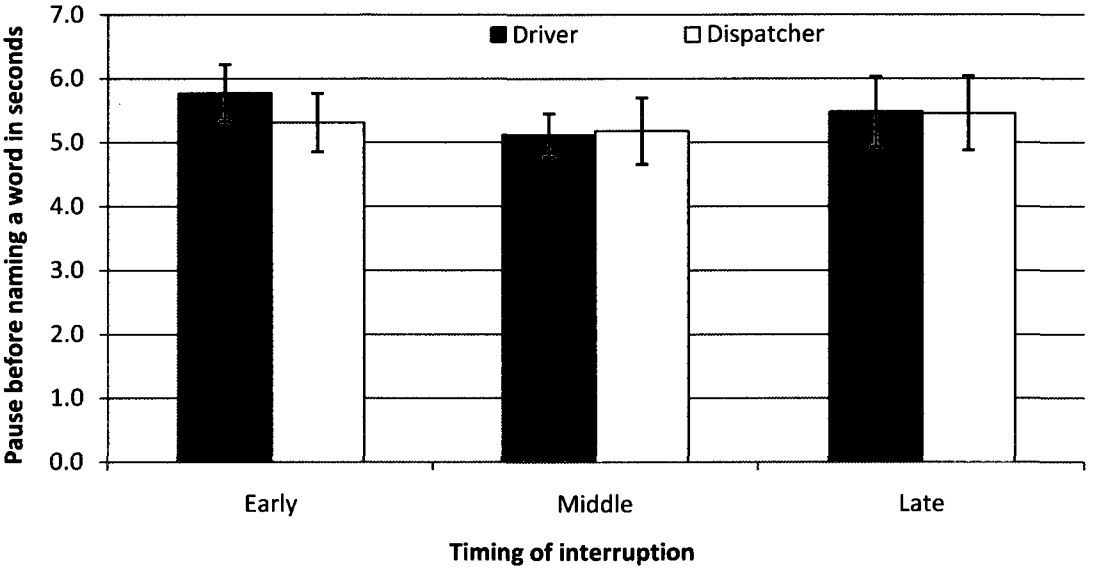


Figure 5.15 Effect of interrupting timing on the interrupting task.

We expected the interruption timing to affect both tasks. However, the data shows that the interruption timing affects the ongoing task, but not the interrupting task. This can be due to the differences in the tasks, or due to the priorities that participants assign to the tasks. The interrupting task had an urgency associated with it, because it had to be done in a limited amount of time. It is also interesting to notice that only *early* interruptions had an effect on the ongoing task. The reason for this could be that *early* interruptions did not create as much time pressure as the *middle* and *late* interruptions. We also confirmed that the duration of questions or speaking rate during question was the same for all conditions. Therefore, the pause before asking a question was the reason why the drivers lost more games during *early* interruptions. Another observation is that the interruption timing affects the drivers but not the dispatchers, which indicates that the driving might affect the spoken tasks. In order to investigate this issue from a different

angle we proceeded to explore if the interruption timing associated with the question levels would provide us with more insight.

5.3.2 Timing of interruptions by level

As discussed in section 5.1.1 (pg. 94) the design of the twenty questions game is such that not all game turns progress a subject through the game equally. This means that amount of information that must be retained during the interrupting task about the twenty questions game does not directly depend on the turn number. It is possible that the amount of information retained during the interruption might affect the cognitive load of the subjects. Using the levels we can classify interruptions based on when they happen in relation to the progression within the game, as opposed the interruption timing based on turns that is described in the previous section. This is a different way of testing how interruption timing influences the spoken tasks (hypothesis 3).

There can be no interruptions before level 1 and if an interruption happens after level 4 we cannot treat it as an interruption, because the ongoing task is complete. We define interruptions at level 1 as *early*, at level 2 as *middle*, and at level 3 as *late*. Figure 5.16 shows the distribution of the games that have interruptions after different levels of questions. Interruptions after level 4 signify the twenty questions games that were completed before an interruption could happen. There is no significant difference between the distribution for the drivers and the dispatchers ($t(15)<-1.23, p>0.24$).

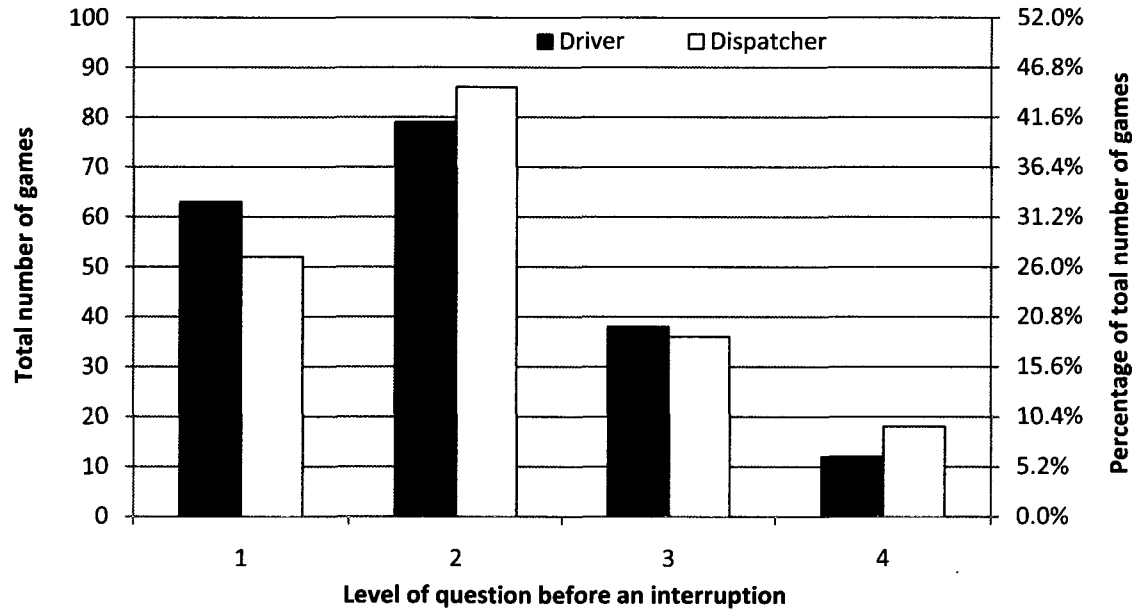


Figure 5.16 Timing of interruption using level of questions.

Statistical analysis showed that the timing of interruptions according to the level does not significantly influence any performance measure of the ongoing task for the drivers and the dispatchers. On the other hand, the timing of interruptions according to the level does influence the last letter word game for the drivers, but not the dispatchers. Figure 5.17 shows the average duration of a pause before naming a word for the drivers and the dispatchers. Statistical analysis showed that the timing of interruptions has a significant effect on the pause duration during the interrupting task for the drivers ($F(1,13)=5.56, p=0.035$). Post hoc comparisons confirmed that the drivers were thinking longer (had longer pauses before naming a word) during the interrupting task if the interruption happened early ($p=0.048$).

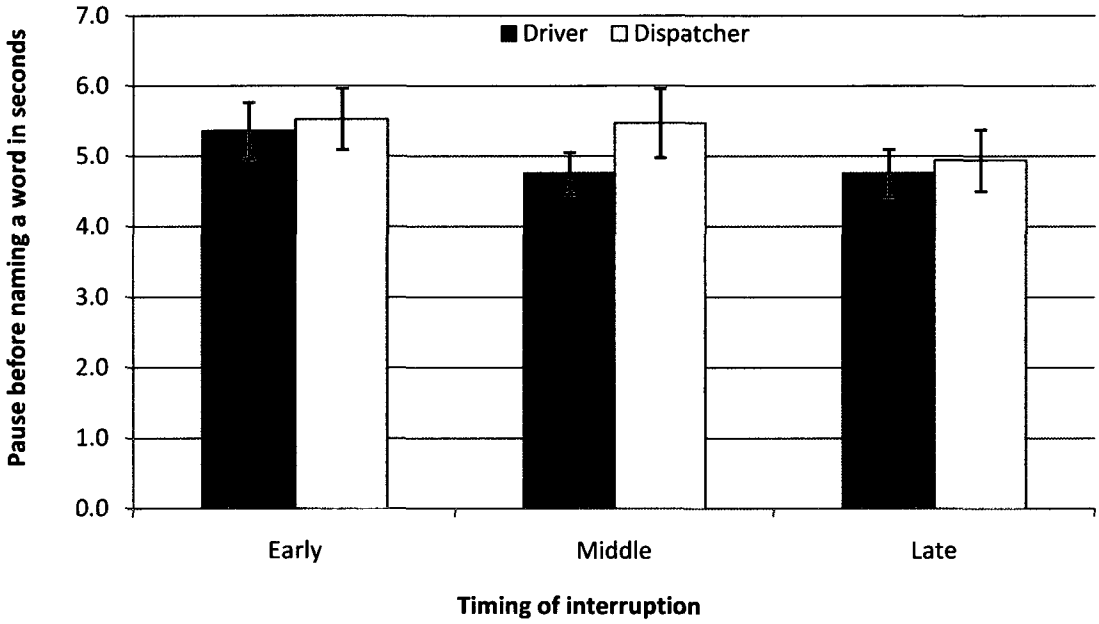


Figure 5.17 Last letter word game pauses (interrupting task).

We assumed that the subjects experience changes in cognitive load as the twenty questions game progress. Given that the driving increases overall cognitive load, we can observe the effects of different interruption timings on the drivers, but not on the dispatchers. On the other hand, a different explanation could be that the drivers knew that the ongoing task just started and there is no need to rush with the interrupting task. Hence, they took the time to think about the interrupting task. In other words, drivers did not experience as much time pressure during *early* interruptions as they did during *middle* and *late* interruptions. If this explanation is correct, then it is not clear why the dispatchers did not exhibit the same behavior. In addition, this trend was not found for the turn based interruption timings for the interrupting task described in the previous section.

Similar to the conclusion in the previous section we see that the drivers are affected by the interruption timing more than the dispatchers. We conclude that both how

long ago a game started and where in the game a subject is could be factors that contribute to the decision of how to perform the spoken tasks. It is not clear to us how these two factors interact with each other. But the data confirm that there is an interaction between the timing of a switch and the spoken tasks performance.

5.4 Driving

In order to test hypothesis 2 (which focuses on how the spoken tasks affect the driving performance), we compare the driver's performance on the ongoing and interrupting tasks. Figure 5.18 shows the lane position variance on different road types during different tasks. Statistical analysis revealed that there is a significant difference in the lane position variance when comparing measurements before, during and after interruptions ($F(2,30)=10.0, p<0.001$) for curvy roads and ($F(2,30)=6.3, p=0.005$) for straight roads.

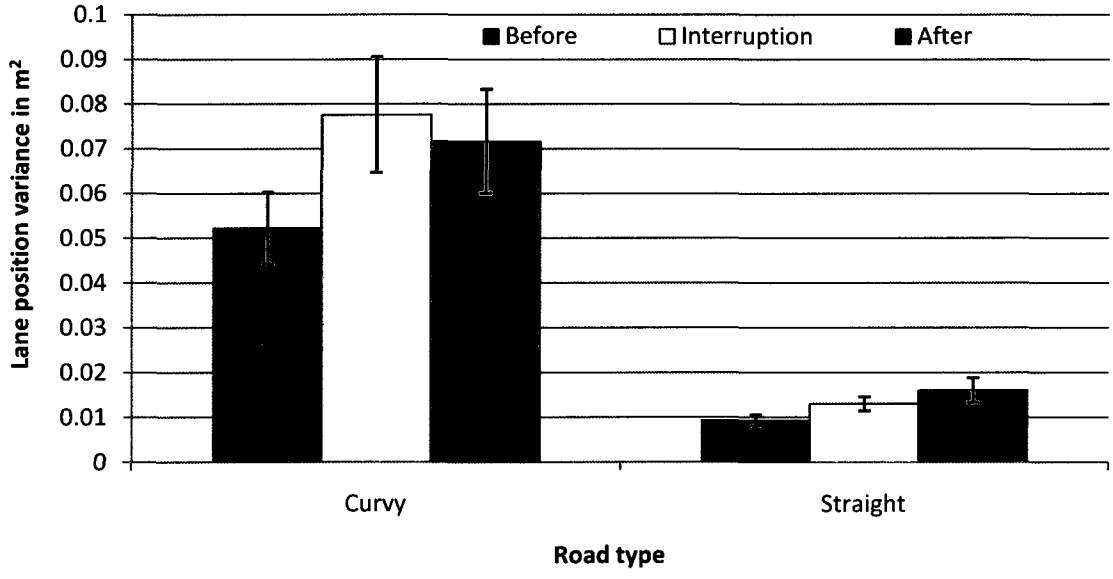


Figure 5.18 Lane position variance on different road types.

Post hoc comparison showed that on curvy roads the lane position variance during the interruption is larger than before interruptions (p=0.002), and the lane position variance is larger before than after interruptions (p=0.007), but the difference between the lane position variance during and after interruptions is not significant (p=0.175). Post hoc comparison showed that on straight roads the lane position variance has significant increase when comparing the lane position variance before and during interruptions (p=0.002), and when comparing before and after interruptions (p=0.005), but the lane position variance during interruptions is not significantly different from the lane position variance after interruptions (p=0.225).

It seems that the lane position variance on curvy and straight roads was affected similarly by the presence of the interruptions (in both cases driving performance decreased during the interruption). We attribute this difference in the lane position variances before and during interruptions to the increased attention demands caused by the interrupting task. The drivers focus on the interrupting task and, consequently, neglect the driving. It is not clear if this affect is associated with a choice, meaning that drivers choose to neglect the driving because the interrupting task is urgent, or the interrupting task is so difficult that the drivers cannot maintain driving performance. We do know that a similar task was used as an interruption in a dual task condition in the research by Strayer and Johnston [52]. The authors showed that indeed this task interfered with a simulated driving task. The current experiment setup does not allow us to make a distinction between driving performance decrements due to the task urgency or the task difficulty, because we do not change how instructions are given to the subjects and we do not change the difficulty of the interrupting task. Changing how we give instructions to

the subjects can change how they perceive the interrupting task. For instance, explicitly telling the drivers that the driving must have the ultimate priority might force the drivers to focus more on the driving and think of the interrupting task as not urgent.

Figure 5.19 shows the velocity variance on curvy and straight roads. Statistical analysis showed that there was no significant difference in the velocity variance on curvy and straight roads ($F(1,15)=0.416, p=0.528$). Only the velocity variance on curvy roads after interruptions is significantly different ($p=0.007$) from the velocity variance before and during interruptions. Figure 5.20 shows the average velocity on curvy and straight roads. Statistical analysis showed that there were no significant differences for the average velocity on different road types and for different tasks ($F(1,15)<1.65, p>0.227$).

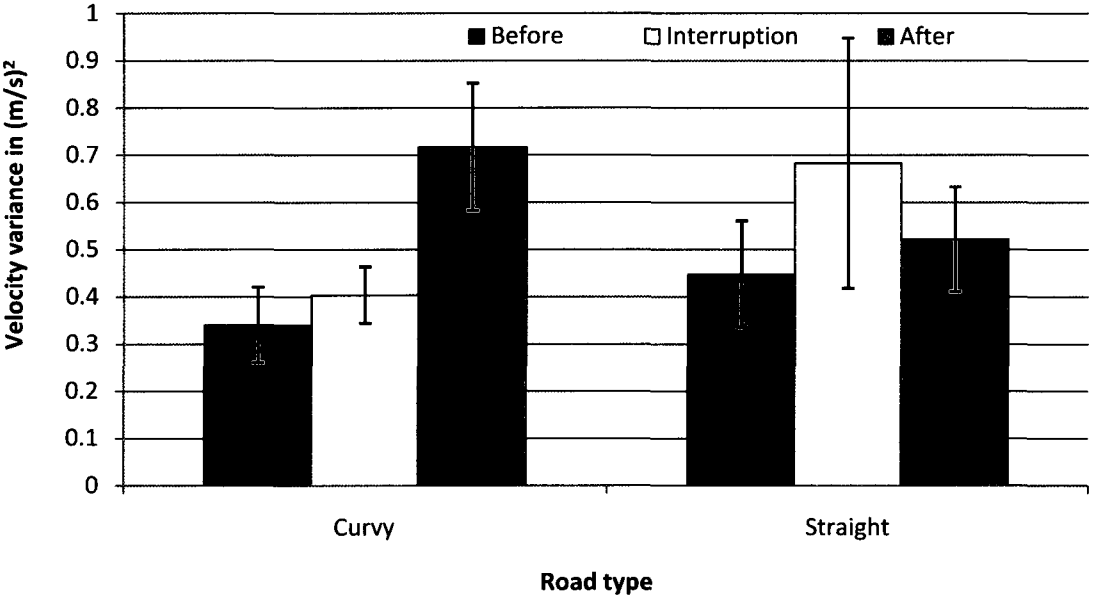


Figure 5.19 Velocity variance on different road types.

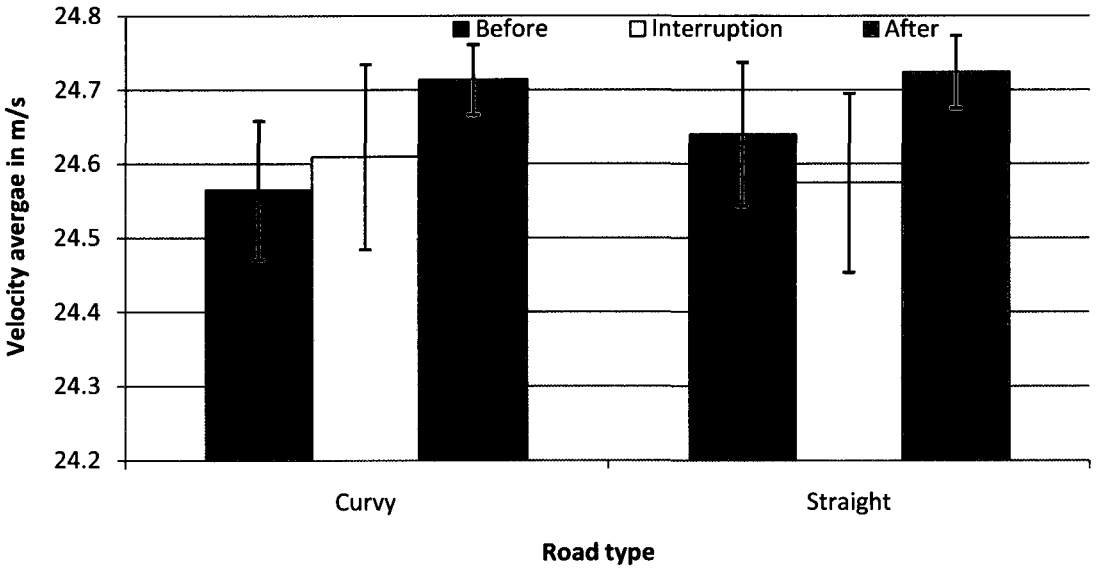


Figure 5.20 Average velocity on different road types.

Vollrath [66] found that the velocity with which subjects drove a vehicle decreased as the complexity of the spoken task increased. Interestingly, Figure 5.20 indicates that on straight roads subjects decrease their average velocity during interruptions as compared to their velocity before interruptions, while such a change did not happen on curvy roads. It could be that the velocity was affected differently by curvy and straight roads. Alternatively, the high data variation is the likely source of the pattern shown on Figure 5.19 and Figure 5.20. The performance measures for the spoken tasks (shown in sections below) do not indicate that curvy roads created a significantly different road difficulty as compared to straight roads which supports the later conclusion. The data also show that the average velocity increased after the interruption for both road types. We suggest that the drivers tried to get closer to the leading vehicle and, therefore, chose to increase their speed. That is also the reason why the velocity variance increased on curvy roads after the interruption. This conclusion is supported by the variance of the distance to the leading vehicle as shown later in this section.

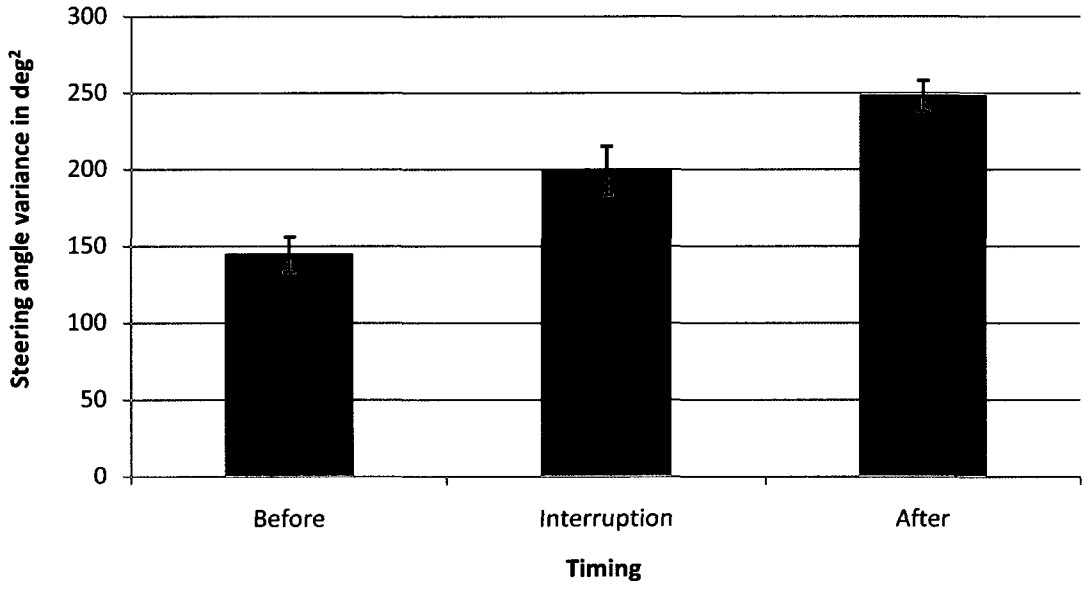


Figure 5.21 Steering variance on curvy roads.

Figure 5.21 shows the steering angle variance on curvy roads for before, during, and after interruptions. Statistical analysis shows that the steering angle variance significantly changes on curvy roads when comparing the steering angle variance before, during, and after interruptions ($F(2,30)=25.0, p<0.001$). Post hoc comparisons revealed that all differences are significant (before vs interruption $p=0.006$; interruption vs after $p=0.004$, before vs after $p<0.001$). It could be that the time when the task is done is a more significant factor than the task itself, i.e. if the interrupting task was present first, it would have the smallest steering variance. This could be caused by the fact that people become more and more tired. On the other hand, the data were extracted from games that happen throughout the experiment from the beginning to the end, which should counterbalance the effects of being tired.

Another possible explanation is that interruptions introduced urgency, because they had to be completed on time. For this reason, the drivers allocated less attention to driving. Once an interruption was over, the participants knew that they could run out of

time to finish the twenty questions game (the perceived urgency by subjects), and that is why the driving performance did not return to the same level as it was before the interruption. This is consistent with our explanation of why the timing of interruptions affected the ongoing task (section 5.3.1, pg. 107). On the other hand, as shown in Figure 5.20 the average velocity on curvy roads was increasing for different tasks in a similar way. Even though the difference in the average velocity before, during, and after interruptions were not significant on curvy roads it is plausible to suggest that a higher average velocity on curvy roads results in a higher steering angle variance. This would mean that the changes in the driving performance are due to the fact that the drivers attempted to catch up with the leading vehicle.

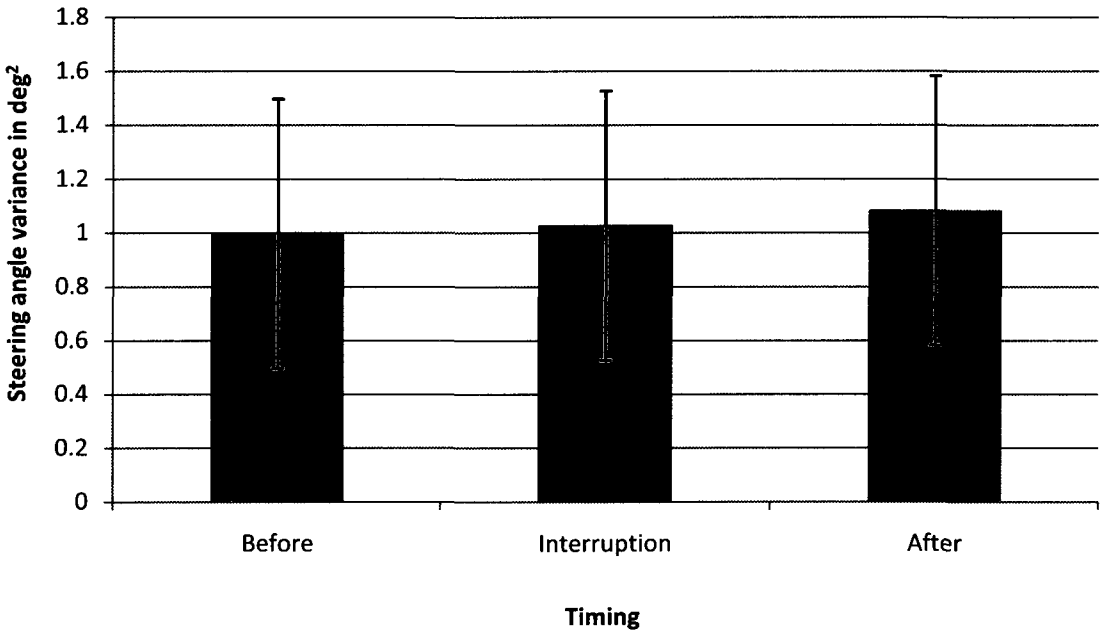


Figure 5.22 Steering angle variance on straight roads.

Figure 5.22 shows the steering angle variance on straight roads for before, during, and after interruptions. The steering angle variance on straight roads exhibit similar trend as on curvy roads (increase from before to during and from during to after

interruptions), but the difference in the steering angle variance before, during and after interruptions is not significant ($F(2,30)=0.14, p=0.870$). The difference in the steering angle variance on curvy and straight roads could be caused by the fact that driving on straight roads is much easier as compared to driving on curvy roads. This is consistent with the previous research by Kun et al. [83].

An argument can be made that the steering angle variance between straight and curvy roads cannot be compared directly due to the presence of turns on curvy roads. Therefore, we filtered the low frequency maneuvers from the steering angle data. We used 0.3Hz to 0.6Hz band to compare the data between curvy and straight roads. Jamson and Merat [55] used similar values to focus on the high frequency variation in the steering angle. Their work was based on the research by McLean and Hoffman [84] who found that normal steering activity to maintain the heading of a vehicle is contained below 0.3Hz. Filtering the signal above 0.6Hz reduces the noise. There is a significant difference ($t(15)>5.449, p<0.001$) between filtered steering angle variance on curvy and straight roads as shown in Figure 5.23. We expected the filtered data for curvy and straight roads to be similar, but because it is not, the argument can be made that filtering values are not chosen properly to remove steering variation due to the turns. It is interesting to notice that the filtered steering angle variance for curvy roads does not exhibit significant change ($F(1,15)=0.1, p=0.923$) when comparing before, during, and after interruptions. This means that the variation observed in Figure 5.21 is due to the low frequency steering control which is used to maintain the vehicle heading [55].

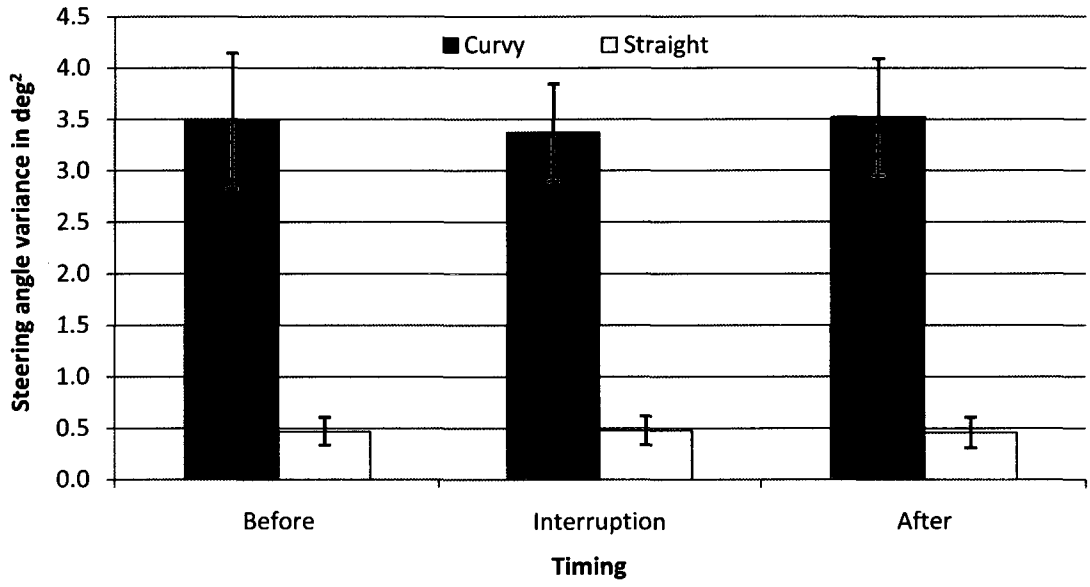


Figure 5.23 Filtered steering variance.

The fact that the steering angle variance significantly changes on curvy roads but not on straight indicates that spoken tasks has greater influence on driving with increased driving difficulty. Hence, the decrements in driving performance due to the interrupting task are more prominent during difficult driving conditions. This is consistent with findings by Strayer and Johnston [52].

Figure 5.24 shows the variance of the distance to the leading vehicle on different road types. The data follow the same pattern as for the velocity variance (Figure 5.19). Similarly, the differences in the distance variance are not significant ($F(1,15)<2.99, p>0.066$). The exhibited trend does show that the distance to the leading vehicle on curvy roads is changing the most after interruptions. The largest variation of the distance to the leading vehicle is during interruptions on straight roads, which implies that on straight roads the drivers allocated the least amount of attention to the driving during interruptions.

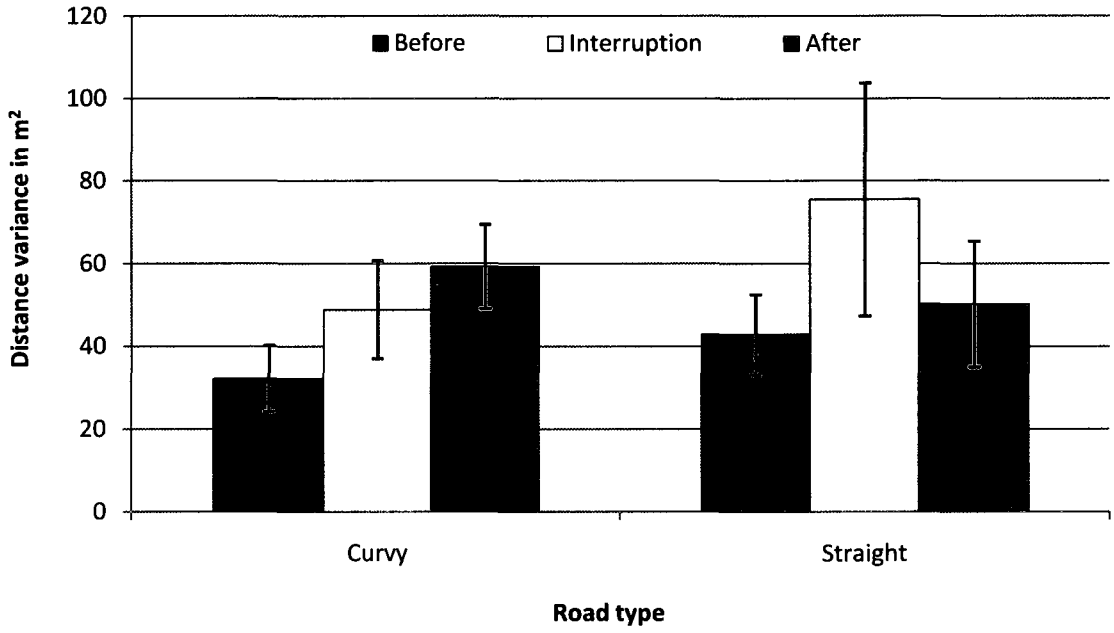


Figure 5.24 Variance of the distance to the leading vehicle on different road types.

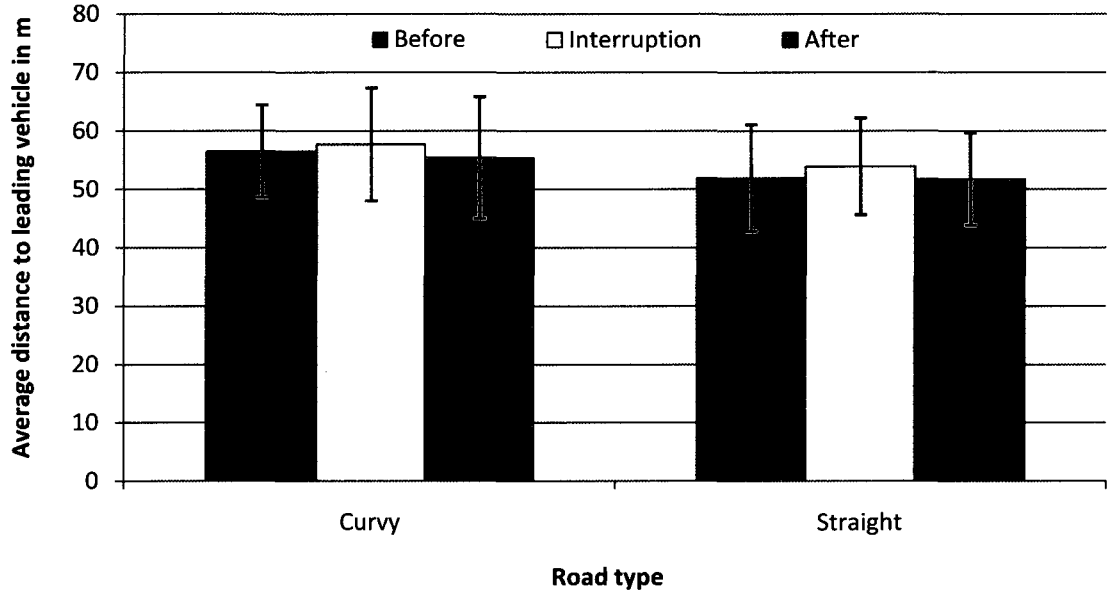


Figure 5.25 Average distance to the leading vehicle on different road types.

Figure 5.25 shows the average distance to the leading vehicle on different road types. Statistical analysis showed that there is no difference in average distance to the leading vehicle for different road ($F(1,15)=1.14, p=0.071$) types or tasks

($F(1,14)=2.19, p=0.238$). This indicates that on average the drivers did maintain the same distance to the leading vehicle during the experiment, but the amount of corrective actions (indicated by the variance) was increasing during the interrupting task.

We also considered comparison of the driving performance between short periods of time. For example, we could compare driving performance when the drivers ask questions with driving performance when the drivers answer questions. Unfortunately, for the driving performance measures that we use in this dissertation (section 4.9, pg. 85) such a comparison yields an ambiguous interpretation in our experiment setup. This is due to the fact that the driving performance measures at any particular short period of time do not necessarily correspond to the actions of a driver during that period of time. For instance, if we observe a change in a driving performance measure when a driver asks a question, there could be multiple contradicting explanations. On one hand, the change could have happened because the driver focuses less on driving and has larger errors. On the other hand, the change could have happened because the driver focuses on driving more and is correcting errors introduced during the previous action, such as answering a question. Given that both interpretations are valid we cannot make the distinction between these two cases. In addition, most of the research done with the similar driving performance measures does not involve averaging over short periods of time [66,55,57,77,83]. Alternatively, there are other driving performance measures, such as a reaction time to a braking leading vehicle, that can be used to avoid this ambiguity, because they require immediate reaction from the driver and, therefore, can be assigned to a particular period of time [85,86]. We did not utilize these performance measures in our experiment setup. Once the experiment setup is modified to

include such performance measures or new methods of processing for the existing measures are available, then it will be possible to compare driving performance between short time periods.

The driving performance measures can also be correlated with cognitive load estimations. For example, the cognitive load estimated using pupillometric measurements [87]could show the interaction between the changes in driving performance and changes in cognitive load.

5.5 Driving difficulty

Hypothesis 1 predicted that more demanding driving conditions should negatively influence the spoken tasks. To study this influence we compared the number of games won by the drivers on curvy roads with the number of games won by the drivers on straight roads. Figure 5.26 shows the outcomes of the games for different road types. Statistical analysis did not show that the road difficulty has a significant effect on the outcomes [Wrong guess ($F(1,15)=1.77, p=0.203$); Timeout ($F(1,15)=0.517, p=0.483$); Correct guess ($F(1,15)=1.31, p=0.723$)].

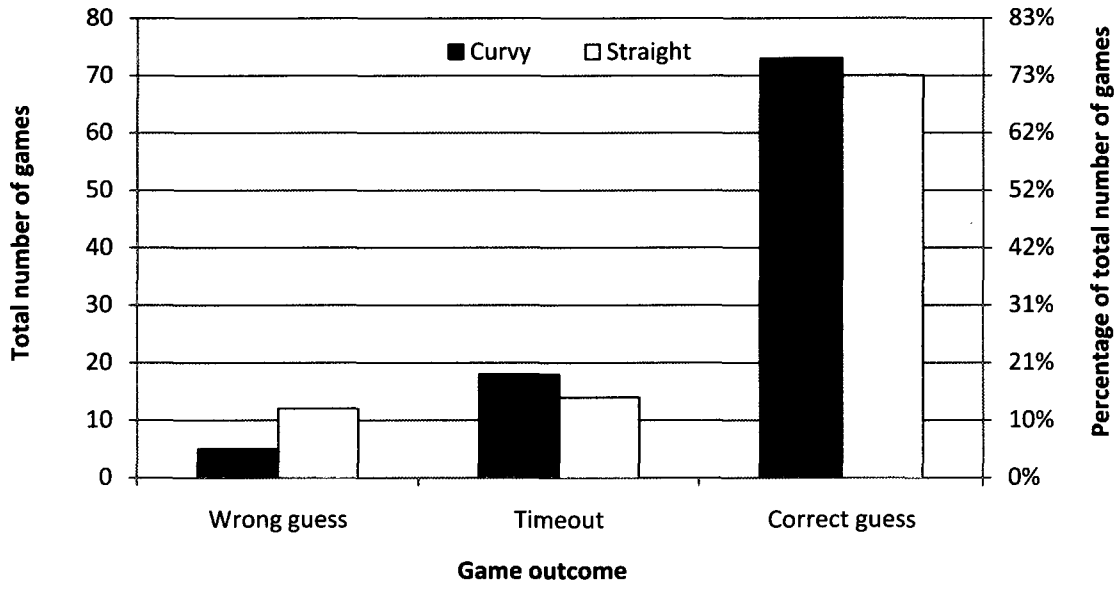


Figure 5.26 Game outcomes for different road types.

Following the same procedure that we used in section 5.3 (pg. 104), we split the games according to the interruption timing. Figure 5.27 shows the percentage of games won for different interruption timings. Statistical analysis did not show a significant difference in the percentage of the games won during different interruption timings according to a turn number for curvy and straight roads ($F(1,15)<0.216, p>0.649$). Figure 5.28 shows percentage of games won for different interruption timings according to turn levels for curvy and straight roads. Statistical analysis did not show significant differences between curvy and straight roads ($F(1,15)<0.4, p>0.535$). The data show that the difficulty of the road did not affect the number of games the drivers win for different interruption timings. It could be that the difference in driving difficulty was not sufficient to create visible changes in the ongoing task performance.

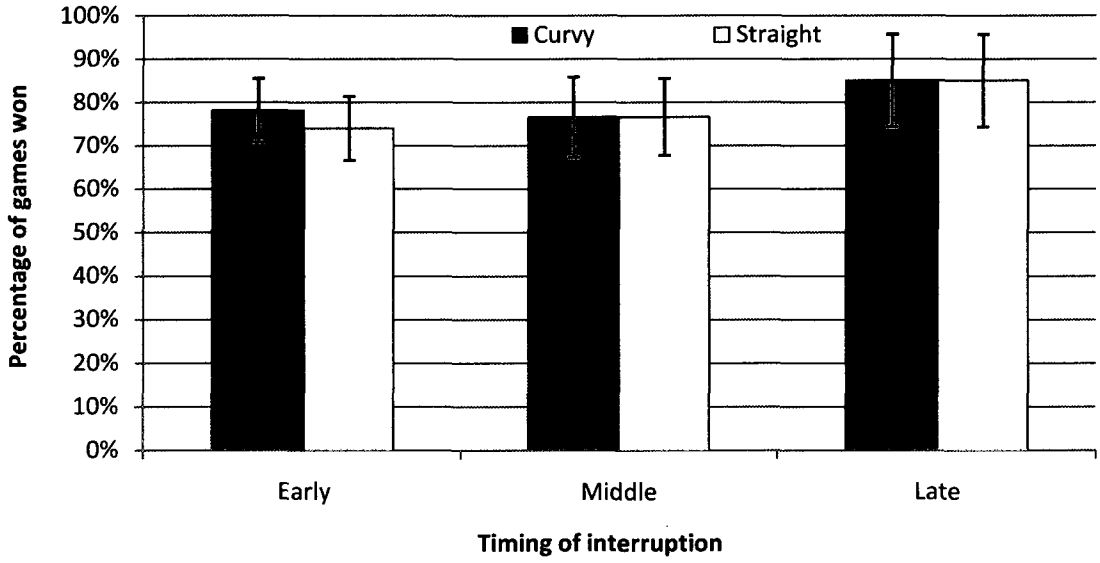


Figure 5.27 Percentage of games won for different interruption timings according to a turn number.

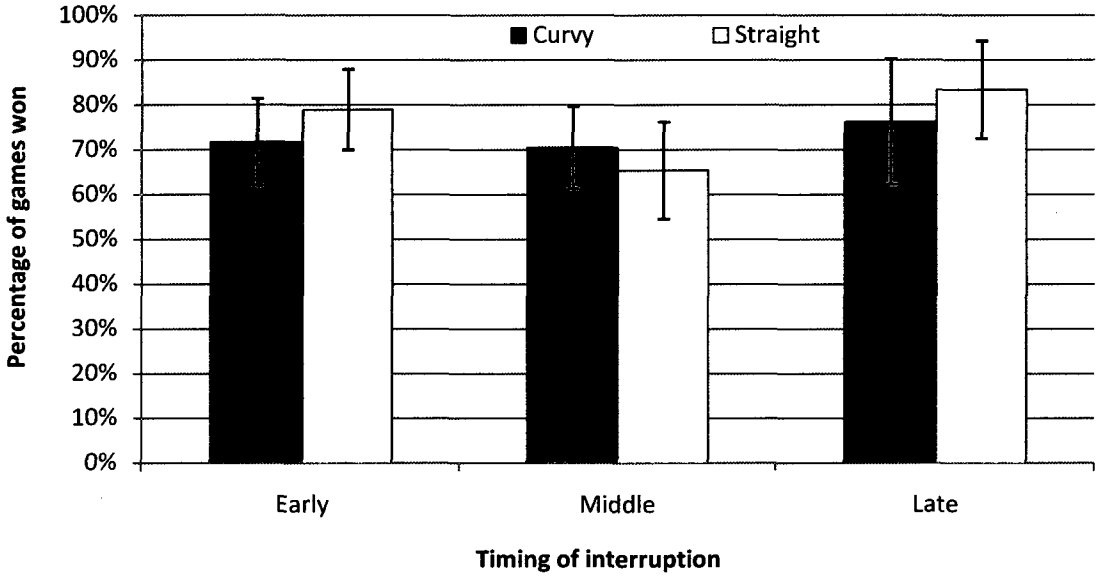


Figure 5.28 Percentage of games won for different interruption timings according to a turn level.

The drivers could have the same number of wins on different road types, but they still could have played slower on curvy roads. To test if that was happening we compared the pauses in the ongoing and the interrupting tasks. Figure 5.29 shows the

average duration of pauses before asking a question during the ongoing task for different road types and different interruption timings according to a turn level. Statistical analysis showed that there is no significant difference between curvy and straight roads for any interruption timing ($F(1,14)<3.6, p>0.080$). Figure 5.29 shows that the pauses are shorter during the *late* interruptions as compared to *early* or *middle* interruptions. ANOVA repeated measures model (with the timing of interruption, the type of road, and the interaction between these two variables) revealed that neither the timing of interruption ($F(2,6)=0.55, p=0.16$), nor the type of road ($F(1,3)=2.15, p=0.239$), nor their interaction ($F(2,6)=0, p=0.99$) has significant effect on the pause before asking a question in the ongoing task. This confirmed that the road difficulty did not influence the ongoing task in our experiment.

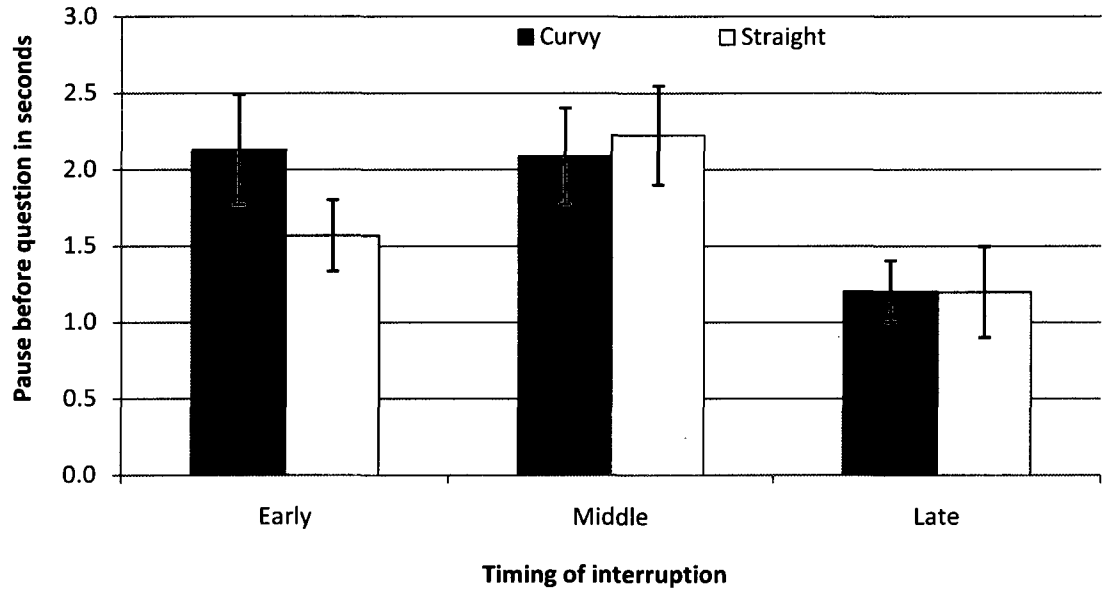


Figure 5.29 Pause before asking a question (ongoing task) for different interruption timings according to a turn level.

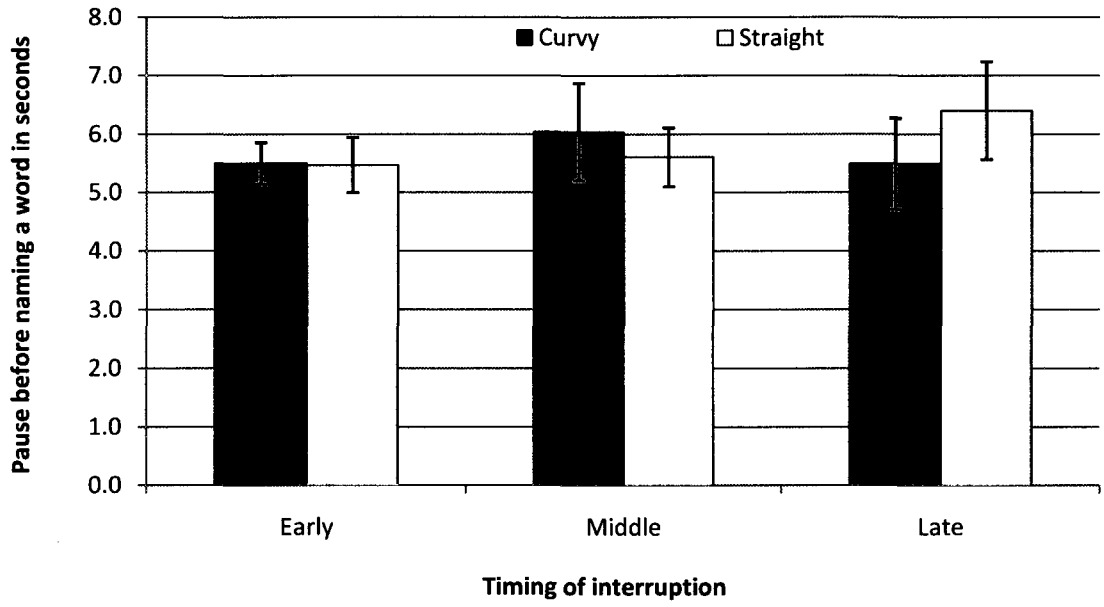


Figure 5.30 Pause before naming a word (interrupting task) for different interruption timings according to a turn level.

Figure 5.30 shows the average duration of pauses before naming a word during the interrupting task for different road types and different interruption timings according to a turn level. Statistical analysis showed that there is no significant difference between curvy and straight road types for any interruption timing ($F(1,6)<2.14, p>0.194$). This, again, confirmed that the driving difficulty did not affect the interrupting task in our experiment.

Similarly, the statistical analysis of the data using the interruption timings according to a turn number did not show any significant effects of the road type on the ongoing and the interrupting tasks. The data presented in this section suggests that driving difficulty did not influence the spoken tasks. On the other hand, it could be that the curvature of curvy roads did not increase the difficulty of the driving as compared to the straight roads to create visible effects. Tsimhoni and Green [57] found that the driving

difficulty increases with the road curvature. We suggest that our assumption about the road difficulty was not correct and, therefore, we do not observe the effects of driving difficulty on the spoken tasks. On the other hand, Strayer and Johnston [52] showed that both the driving difficulty and the spoken task difficulty affect the driving performance. It could be that the spoken task difficulty was not chosen properly to illustrate an interaction between the road difficulty and the spoken tasks.

5.6 Multiple task management

The following sections outline how the interruptions were initiated by the subjects and how the subjects switched between the ongoing task and the interrupting task. Explanations of the models are given in section 3.7 (pg. 37) and 4.8.3 (pg. 82). The purpose of the following analysis is aimed to understand different switching behaviors, which is the focus of hypothesis 4.

5.6.1 Interruption initiation

We coded the interruption initiation based on where it happened with respect to the model in Figure 5.31 (copy of Figure 4.25). There were 93 interruptions presented to the driver (3 out of 96 interruptions were presented after the ongoing task was complete) and 84 interruptions presented to the dispatcher (12 out of 96 interruptions were presented after the ongoing task was complete). For the drivers, there were 45 interruptions presented on curvy roads (3 out of 48 interruptions were presented after the ongoing task was complete) and 48 interruptions presented on straight roads.

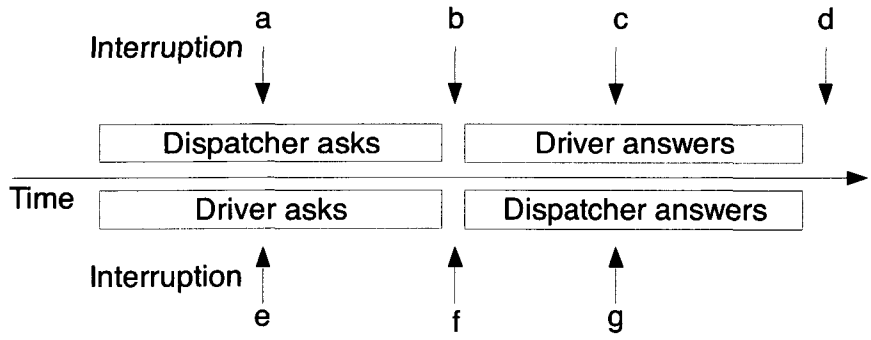


Figure 5.31 Interruption timing.

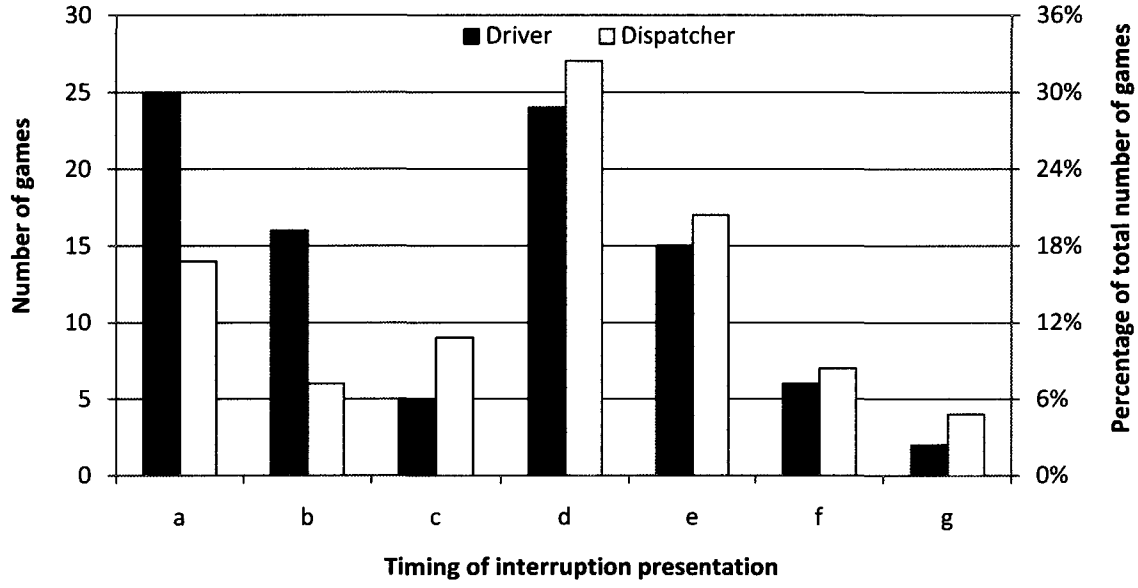


Figure 5.32 Interruption presentation timing.

Figure 5.32 shows when interruptions were presented to the subjects on the screens in relation to the most recent adjacency pair. This figure shows that b, c, f, and g had the smallest number of presentations. This is due to the fact that these are the shortest periods in adjacency pairs. Answers marked as c and g are "yes/no" answers and have very short duration. This distribution is consistent with our previous research [11] (section 3.12, pg. 43) and the task design (section 4.3, pg. 58).

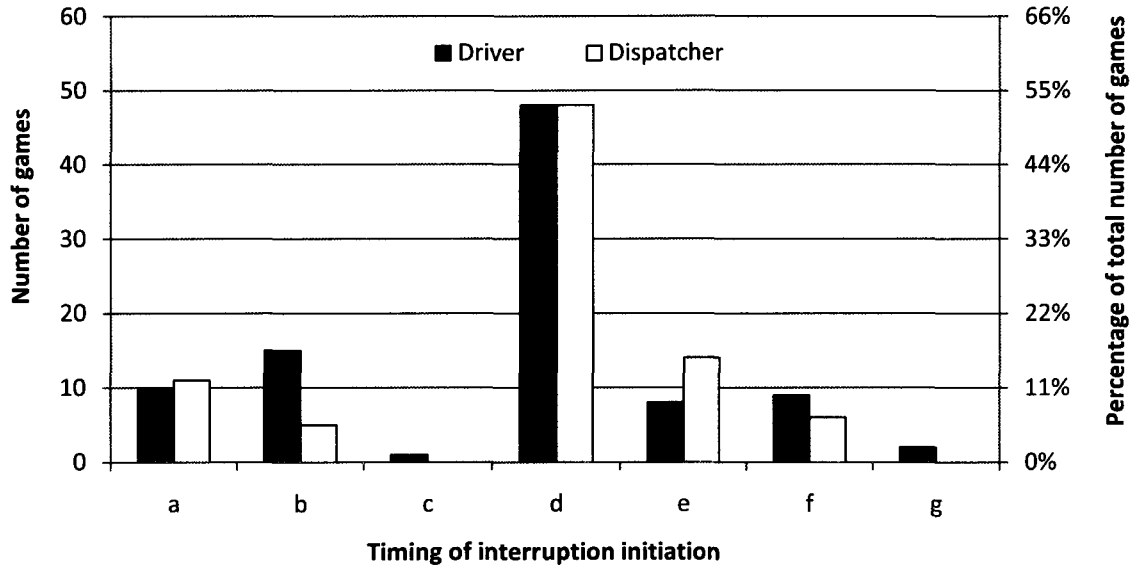


Figure 5.33 Interruption initiation timing

Figure 5.33 shows when interruptions were initiated by the subjects on the screens in relation to the most recent adjacency pair. The plot demonstrates that both the drivers and the dispatchers chose to interrupt when no one was speaking (during the pause between adjacency pairs "d"), which is consistent with our previous research [11] (section 3.12, pg. 43). Statistical analysis showed that the drivers and the dispatchers were equally likely to interrupt each other or themselves (initiate interruptions during parts "a" or "e"). We attribute no differences in the behaviors to the fact that both the drivers and the dispatchers treated the interruption as a priority. For this reason, driving did not change how the drivers introduced interruptions. Given that driving performance decreased during the interrupting task (for example, as shown in Figure 5.18, pg. 116) we can suggest that the drivers behaved as if the driving task did not have a priority (thus the same behavior as dispatchers for the interrupting task). This implies that in order to see how driving affects interruption introduction, the drivers must be instructed to maintain

driving performance as the priority, or the driving difficulty should be harder not to allow the subject to be distracted from the driving task.

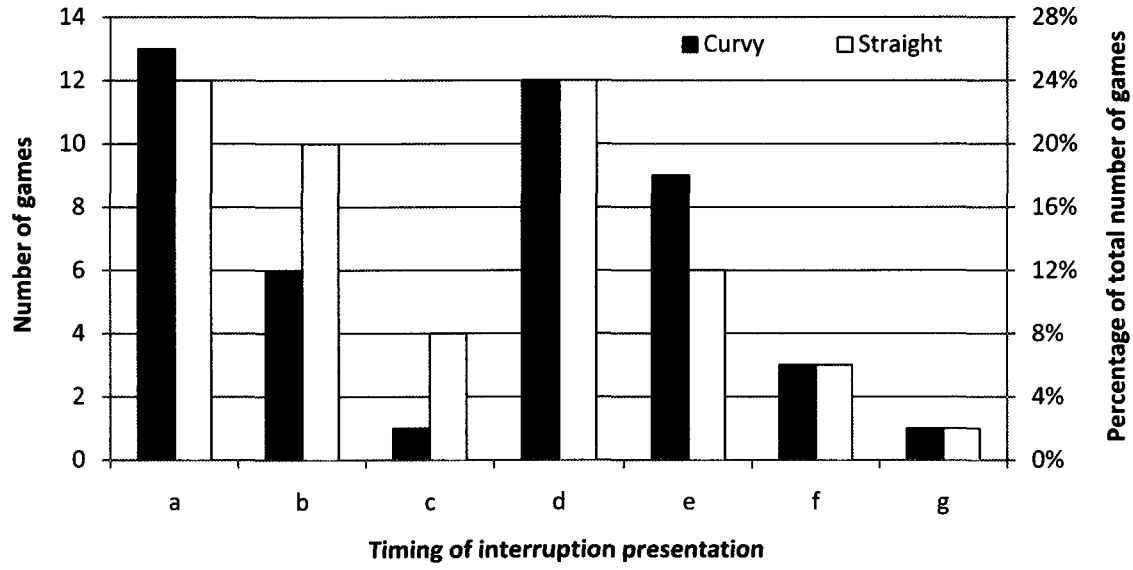


Figure 5.34 Interruption presentation timing on curvy and straight roads.

Figure 5.34 shows how the interruption presentations were distributed for curvy and straight roads and Figure 5.35 shows the distribution of the interruption initiations for curvy and straight roads. These distributions demonstrate that the drivers preferred to wait for the end of an adjacency pair to introduce interruptions on both road types. Statistical analysis did not show any significant effect of road difficulty on the timing of interruption initiation ($F(1,15)<4, p>0.05$). Such results can be interpreted in support of our conclusion that the interruptions had priority over driving for the drivers.

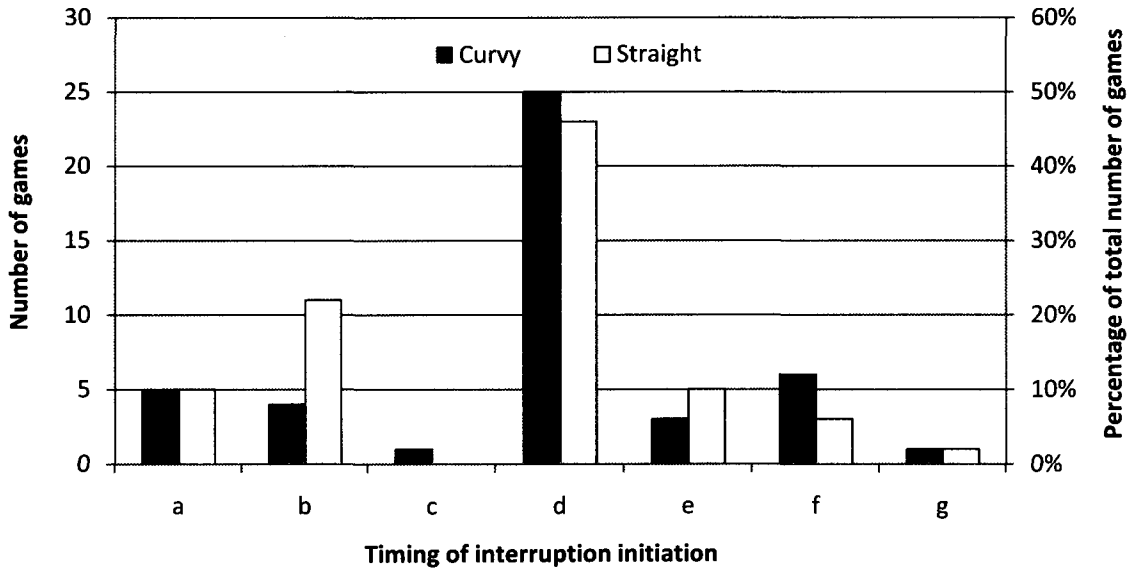


Figure 5.35 Interruption initiation timing on curvy and straight roads.

We also looked at the distribution of the interruption initiations for different interruption timings (early, middle, and late). The sparse number of data points and their uneven distribution among these interruption timings did not allow us to draw a conclusion about how different interruption timings affected the interruption initiations. The reason for that is that in our experiment setup we did not control the distribution of the interruption initiation in relation to the interruption timings.

5.6.2 Task switching

The model of switching between the ongoing and the interrupting tasks is explained in the section 4.8.3 (pg. 82) and is aimed at understanding different switching behaviors, which is the focus of hypothesis 4. Figure 5.36 (copy of Figure 4.24) shows the summary of this model.

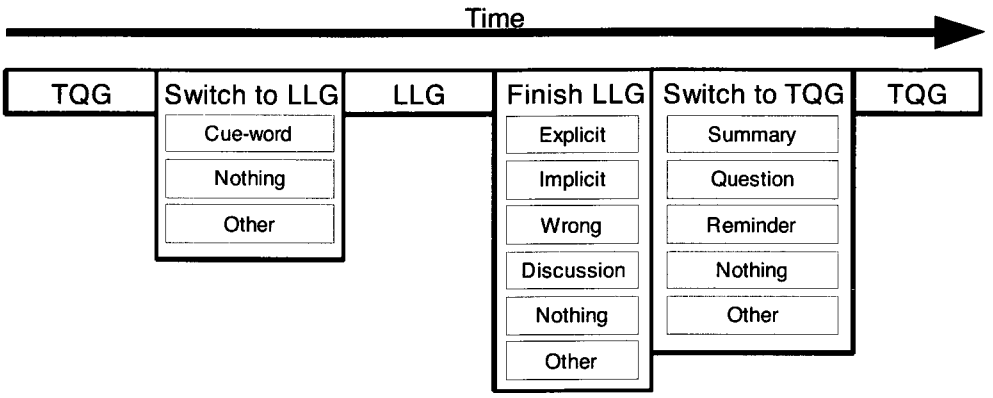


Figure 5.36 Interruption/resumption of a twenty questions game.

We found that subjects used a cue word in only four out of 192 interruptions. This could indicate that the tasks were very different, and, therefore, did not require additional cue words. In addition, there was only one interrupting task, and this might be the reason why subjects did not need to cue each other about the switch. This model ignores the fact that it is possible to have multiple switches between TQG and LLG, for example, when asked a question the person initiates an interruption by requesting to name a word, but then immediately answers the question. These cases were infrequent (3% or 6 interruptions) and were excluded from the analysis.

When the interrupting task was completed each participant took one of the actions shown in Table 5.5 (explained in section 4.8.3, pg. 82). Table 5.6 shows an example of the interrupting task followed by the finish of the interrupting task and the switch to the ongoing task (game 6, subject pair 4). In this example, the dispatcher explicitly signaled the end of the interrupting task, while the driver implicitly confirmed it. This example contains no context restoration activity before the subjects continued the ongoing task.

Action	Example
Explicit	That's my three.
Implicit	Okay.
Wrong	That's my three, oh no, I need one more.
Discussion	Are we done?
Nothing	

Table 5.5 Finishing LLG actions.

Code	Speaker	Utterance	Details	Task
U1	Dispatcher	Begins with D	Interrupting task	LLG
U2	Driver	Dude.	Driver's turn 1	LLG
U3	Dispatcher	Easy.	Dispatcher's turn 1	LLG
U4	Driver	Yarn.	Driver's turn 2	LLG
U5	Dispatcher	Nate.	Dispatcher's turn 2	LLG
U6	Driver	Early.	Driver's turn 3	LLG
U7	Dispatcher	Yell.	Dispatcher's turn 3	LLG
U8	Dispatcher	I think that's three for us.	Explicit signal	Switch
U9	Driver	Yep	Implicit signal	Switch
U10	Dispatcher	Is it in the living room?	Ongoing task	TQG

Table 5.6 Interrupting task for game 6, subject pair 4.

Figure 5.37 shows the average percentage of games for each type of finishing the interrupting task. The statistical analysis showed that all of these actions were employed by the drivers and the dispatchers equally often ($t(15)>1.72, p>0.106$). Nevertheless, the data exhibit a trend that the drivers chose to provide fewer confirmation signals than the dispatchers. This can be explained by the increased workload induced by the driving task. As a result, we suggest that increasing the driving difficulty will create more differences in the switching behavior for the drivers and the dispatchers.

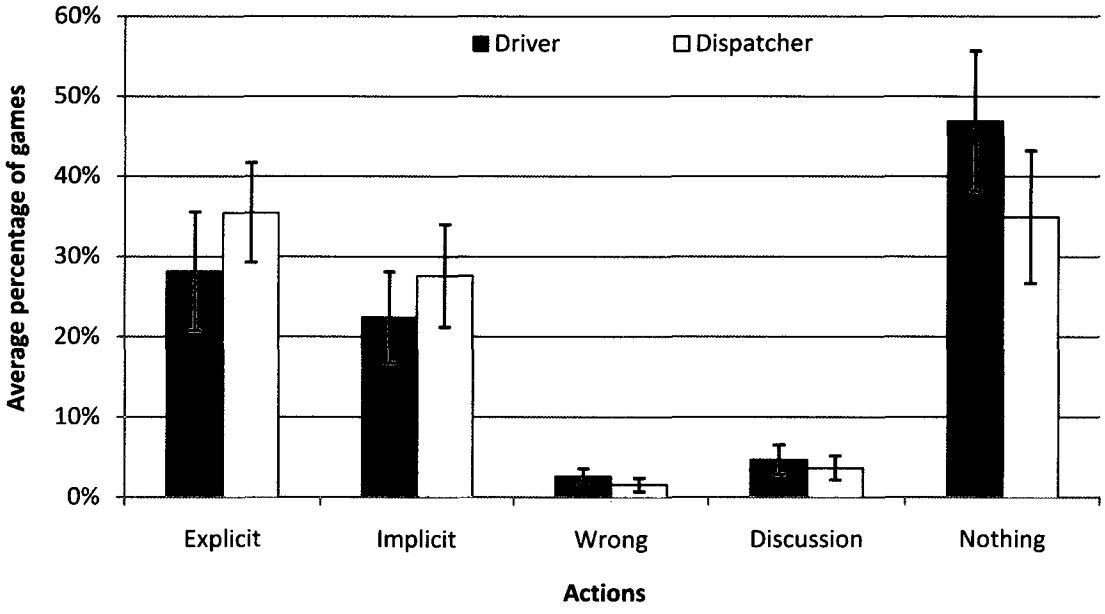


Figure 5.37 Average percentage of games with different types of finishing LLG.

To test this suggestion, we compared different types of finishing LLG for the drivers and the dispatchers on curvy roads only. Figure 5.38 is similar to Figure 5.37, but only the data from the games done when the driver was driving on curvy roads is used. Even though the plot suggests that the drivers used less explicit signaling, statistical analysis showed that there is no significant difference ($t(1,15)<1.218, p>0.242$). The statistical analysis did not support our expectation that the driving difficulty affects how a person handles multi-threaded dialogs. We treat this as a support for our previous observations that driving difficulty did not influence the interrupting task (section 5.5, pg. 126).

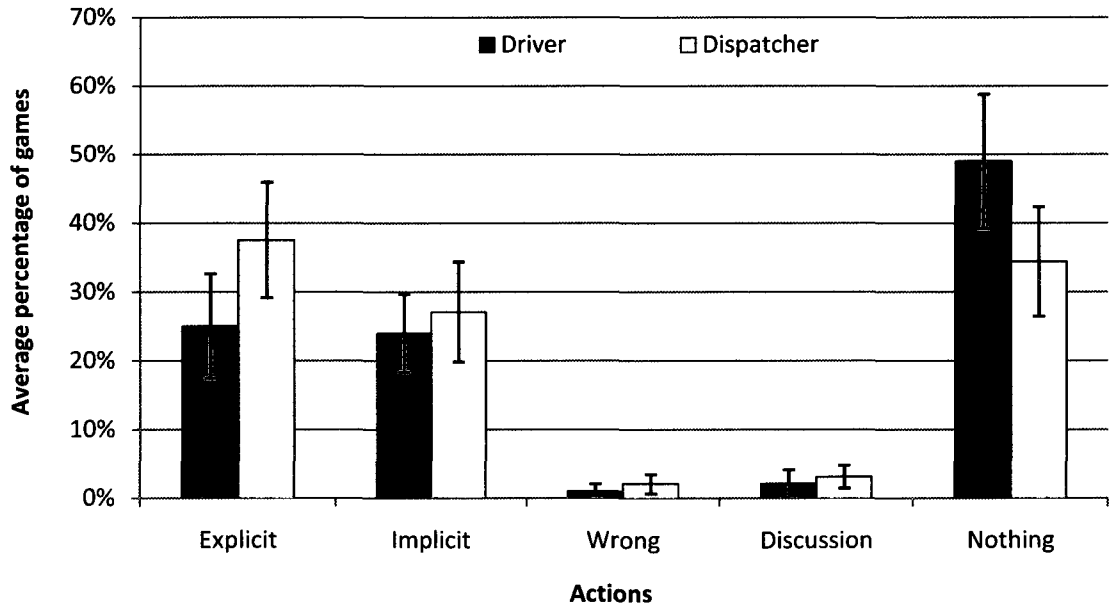


Figure 5.38 Average percentage of games with different types of finishing LLG using data from curvy roads only.

To further investigate the situation we compared how the drivers signal finishing of LLG during curvy and straight roads. Figure 5.39 shows how the drivers choose to finish LLG on curvy and straight roads. There was no significant difference between how the drivers handled finishing of LLG on curvy and straight roads [Explicit ($F(1,15)=0.19, p=0.19$; Implicit ($F(1,15)=0.319, p=0.58$); Wrong ($F(1,15)=1.9, p=0.188$); Discussion ($F(1,15)=3.151, p=0.096$); Nothing ($F(1,15)=1, p=0.33$)]. We suggest that the road difficulty was not chosen properly to show differences between behaviors on curvy and straight roads. This suggestion is also supported by the data in section 5.5 (pg. 126).

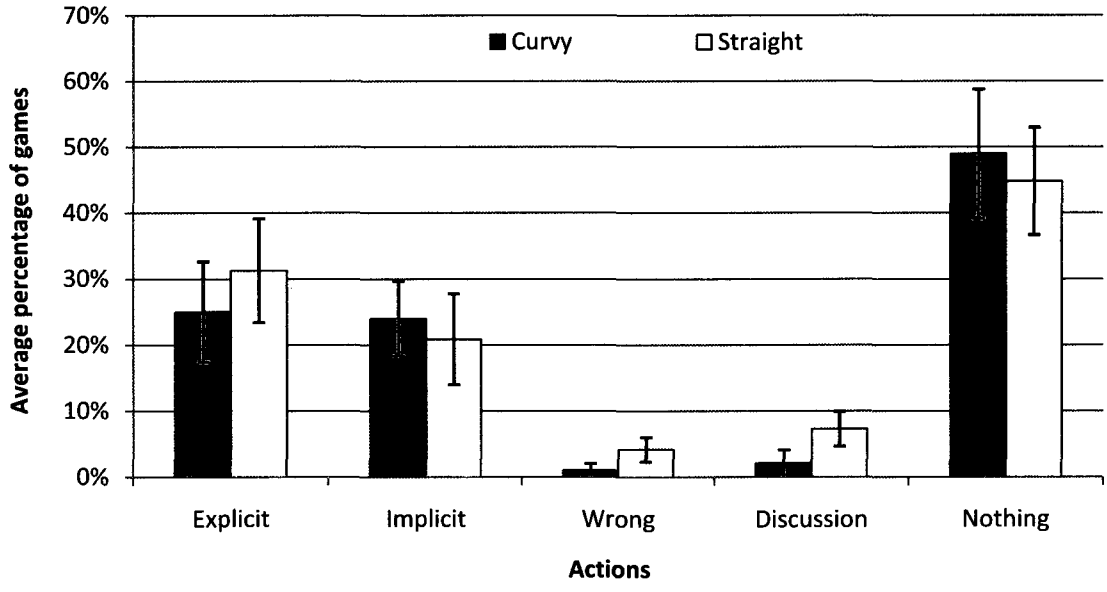


Figure 5.39 Average percentage of games with different types of finishing LLG on different road types.

Figure 5.38 hints that the drivers used less explicit signaling on curvy roads than the dispatchers. This could be explained by the additional workload caused by the driving task. If the driving task would be harder, then the difference could be more pronounced. The fact that we did not find statistical difference between the signaling behavior of the drivers and the dispatchers can be attributed to the insufficient road difficulty as explained earlier.

There are two other possible explanations to why the drivers might change their behavior. It could be that the drivers chose to speak less, so they can focus on driving. This indirectly implies that the drivers are aware of the increased workload and chose their priorities accordingly. It also could be that the dispatchers chose to provide more signaling to help the driver. We consider this case to be very unlikely because the dispatchers did not have information about the driving difficulty.

To further look into different types of finishing LLG we split the data according to the interruption timing as we did in section 5.3.1 (pg. 107). Figure 5.40 shows the average percentage of games with different types of finishing LLG for early, middle, and late interruption timings for the dispatchers. Figure 5.41 shows the same information for the drivers. Because types labeled "Wrong" and "Discussion" lack sufficient data for analysis we focused on explicit, implicit and no signaling types. Statistical analysis showed that the interruption timing, the type of signaling, and the interaction between these two factors do not have significant effects on the dispatchers ($F(4,11)<1.59, p>0.24$) or the drivers ($F(4,11)<2.14, p>0.143$).

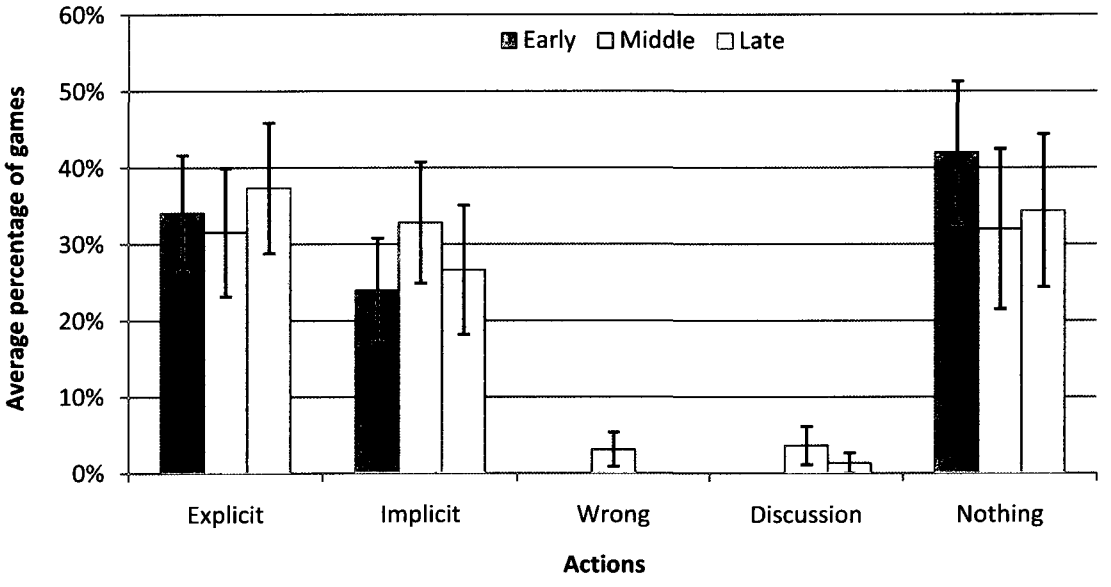


Figure 5.40 Average percentage of games with different types of finishing LLG for different interruption timings for the dispatchers.

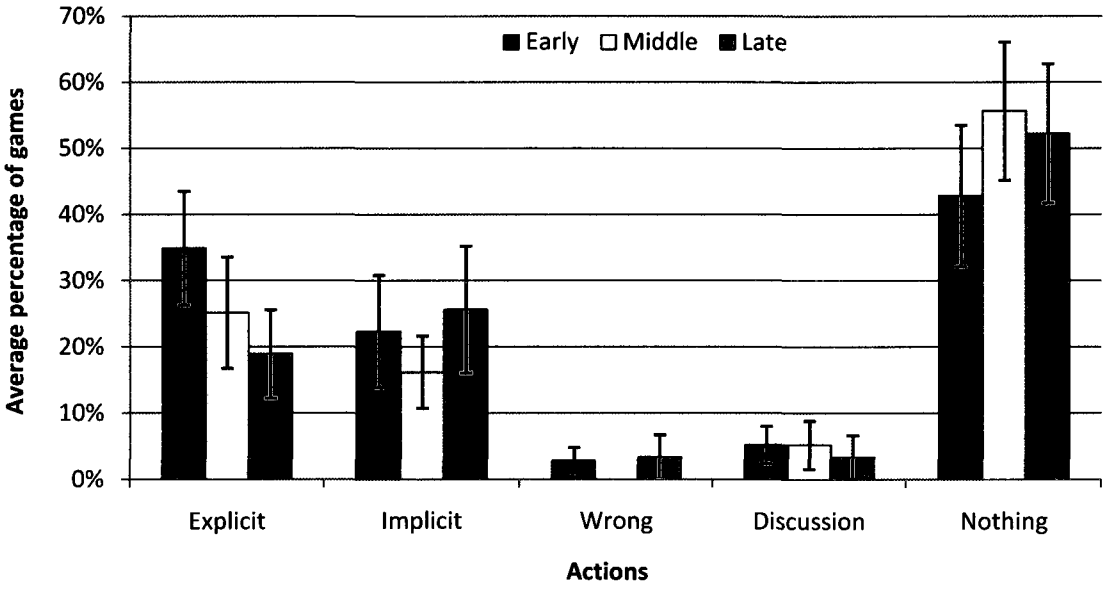


Figure 5.41 Average percentage of games with different types of finishing LLG for different interruption timings for the drivers.

As expected, Figure 5.41 shows the same trend as Figure 5.37, which indicates that the drivers chose to not signal finishing of LLG more often as compared to other types of signaling or as compared to the dispatchers. We performed similar analysis using timing of interruptions according to the level of a turn instead of the number of a turn as explained in section 5.3.2 (pg. 113). The results were similar for both types of interruption timings for the dispatchers and the drivers. This indicates that the timing of interruptions did not influence how the drivers or the dispatchers chose to finish LLG. We suggested (section 5.3.1, pg. 107) that *middle* and *late* interruption timings had a higher perceived urgency. Given that the subject did not change how they finish LLG in those cases might indicate that types of finishing LLG are not affected by the task urgency. This can be explained by the fact that the signaling itself does not take much time (the signaling utterances are short), and therefore, the subjects did not have to change their behavior.

The switch back to the ongoing task might require people to restore their previous state. Table 5.7 shows possible state restoration techniques (section 4.8.3, pg. 82). Figure 5.42 shows the average percentage of games that utilized these techniques. The plot demonstrates that the drivers and the dispatchers utilized each of these techniques equally often. Statistical analysis showed that there is a statistical difference between different types of state restoration [($F(1,15)=84, p<0.001$) for the drivers and ($F(1,15)=96, p<0.001$) for the dispatchers], but post hoc analysis revealed that only "Nothing" is different from all other types ($p<0.001$), but the other types do not differ significantly between each other ($p>0.06$). The fact that both the drivers and the dispatchers did not use any context restoration in more than 70% of the time indicates that the interrupting task did not create enough interference with the ongoing task to require context restoration. On the other hand, the fact that both the drivers and the dispatchers used different techniques the same way could indicate that they matched each other behavior. "Summary" has a significant correlation ($r(190)=0.205, p=0.004$) and "Nothing" has significant correlation ($r(190)=0.325, p<0.001$) for the drivers and the dispatchers, while resumptions and reminders are not highly correlated. It is important to notice that the small number of data points for "Summary" can be responsible for the obtained significance of the correlation. Similarly, the large number of data points for "Nothing" resulted in high significance of the correlation.

Action	Example
Summary	Mine had sharp edges.
Question	Was mine used for heating?
Reminder	You were in the living room.
Nothing	

Table 5.7 Types of state restoration techniques.

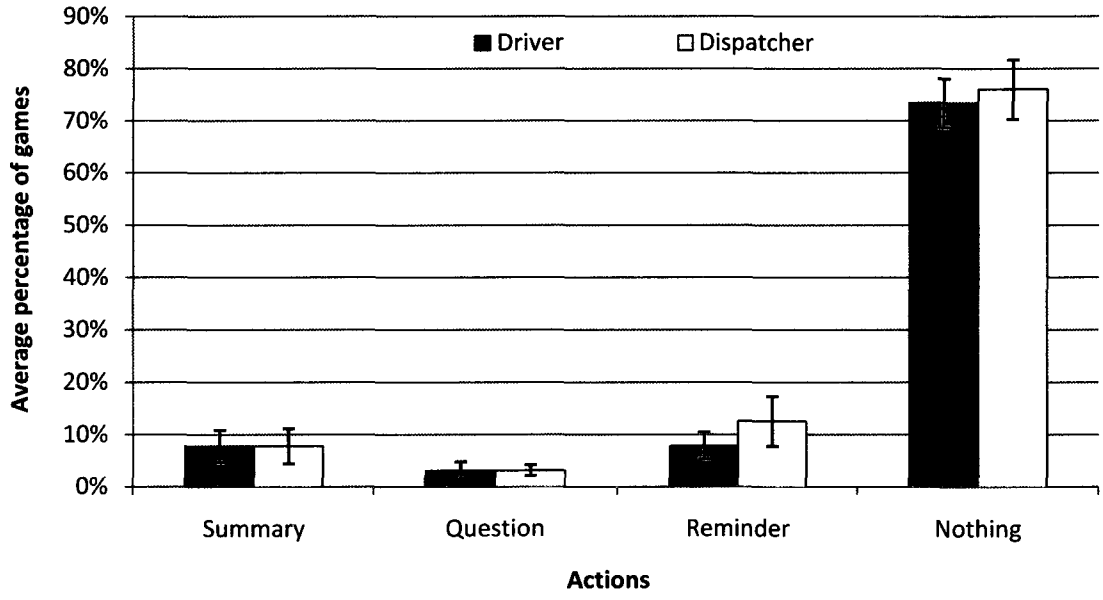


Figure 5.42 Type of the state restoration for TQG.

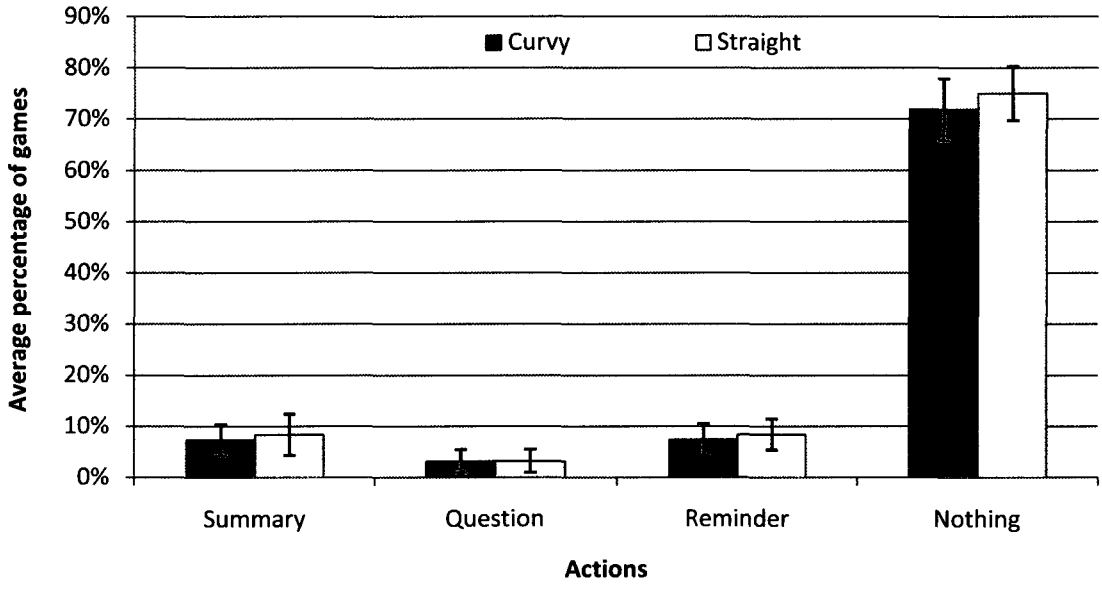


Figure 5.43 Effect of driving difficulty on state restoration for TQG for drivers.

Figure 5.43 shows how often different resumption methods were used on different road types. The plot demonstrates that driving difficulty did not affect how the drivers resumed the ongoing task. It could be that the actions the drivers take to switch back to the ongoing task are not influenced by the driving difficulty. On the other hand, it

could be that the difference in the road difficulties between curvy and straight roads was not enough to show a difference in the drivers' behavior. Results shown in section 5.5 (pg. 126) also support this explanation.

The lack of data for different types of state restoration (less than 10% for individual types, see Figure 5.42) does not allow us to investigate how interruption timing according to the turn number or the turn level changes the behavior of the drivers and the dispatchers.

5.6.3 Driving performance

In addition, we also investigated the interaction between driving performance and the switching behavior of the subjects. This investigation was not part of our initial hypotheses, because we did not want to assume that the distribution of different types of behaviors would allow us to investigate driving performance. Our data show that such an assumption would be wrong for different types of state restoration for TQG, because there are not enough data points (Figure 5.43). On the other hand, the number of data points for different interruption initiations and different types of finishing LLG allows us to look at the interaction between driving performance measures and switching behavior.

None of the driving performance measures showed a significant difference between games with different interruption initiations. Similarly, none of the driving performance measures showed a significant difference between games with different types of finishing LLG. This suggests that the timing of an interruption initiation or the type of finishing LLG did not influence overall driving performance. This can be explained by the fact that the initiation or finishing LLG happens in a short period of time

as compared to the duration of the ongoing and interrupting tasks (average time from interruption presentation to interruption initiation is 2.5 seconds). Another confounding factor is that driving performance data (section 5.4, pg. 116) suggests that the drivers neglected driving during the interrupting task. Therefore, any decrements in driving performance due to the different types of interruption initiation were masked by general driving performance degradation during the interrupting task. The same explanation holds true for the different types of finishing LLG. In addition, we hypothesize that the driving performance after an interrupting task is finished is affected by the perceived urgency of the ongoing task, because subjects could run out of time before finishing the ongoing task (see section 5.4, pg. 116 for more explanations). This also might mask the changes in driving performance due to the changes in the switching behavior.

5.7 Self assessment

All subjects were administered a questionnaire after the experiment (Appendix B). They had to rate their agreement with given statements using Likert scale from 0 to 4 (0 - strongly disagree, 1 – disagree, 2 – undecided, 3 – agree, 4 – strongly agree). There were two questions that show how subjects perceived difficulty of the spoken tasks: "Twenty Questions game was difficult" (the ongoing task) and "Last letter word game was difficult" (the interrupting task). Figure 5.44 shows how the drivers rated the tasks, while Figure 5.45 shows how the dispatchers rated the tasks. Figure 5.46 presents the same ratings as histograms.

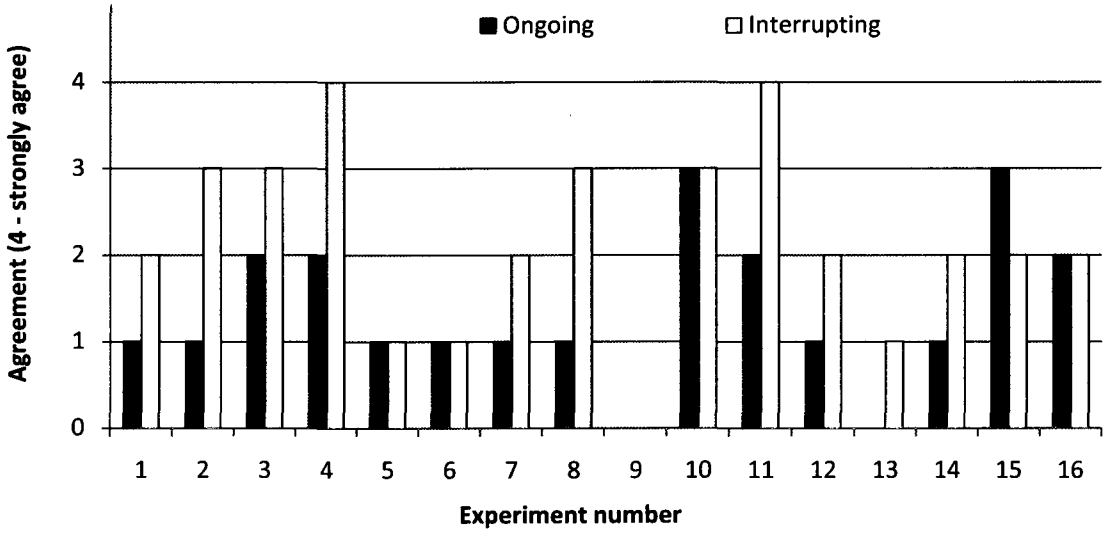


Figure 5.44 Task difficulty rated by the drivers (subject pair nine gave zero score).

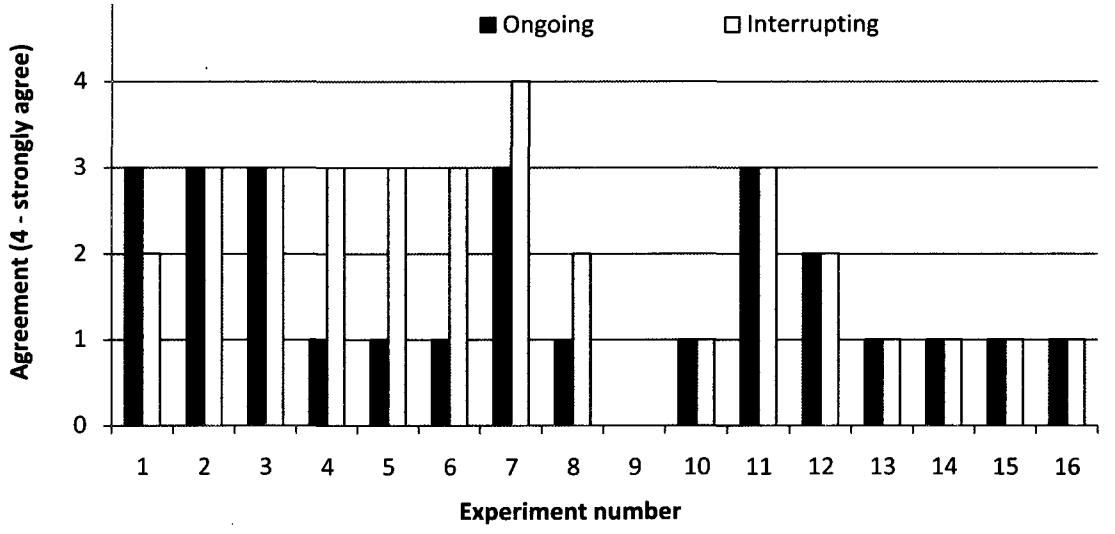


Figure 5.45 Task difficulty rated by the dispatchers (subject pair nine gave zero score).

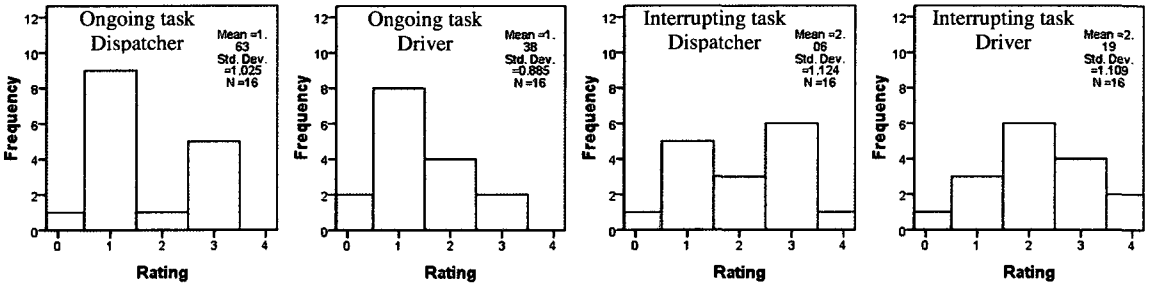


Figure 5.46 Task difficulty rating presented as histograms.

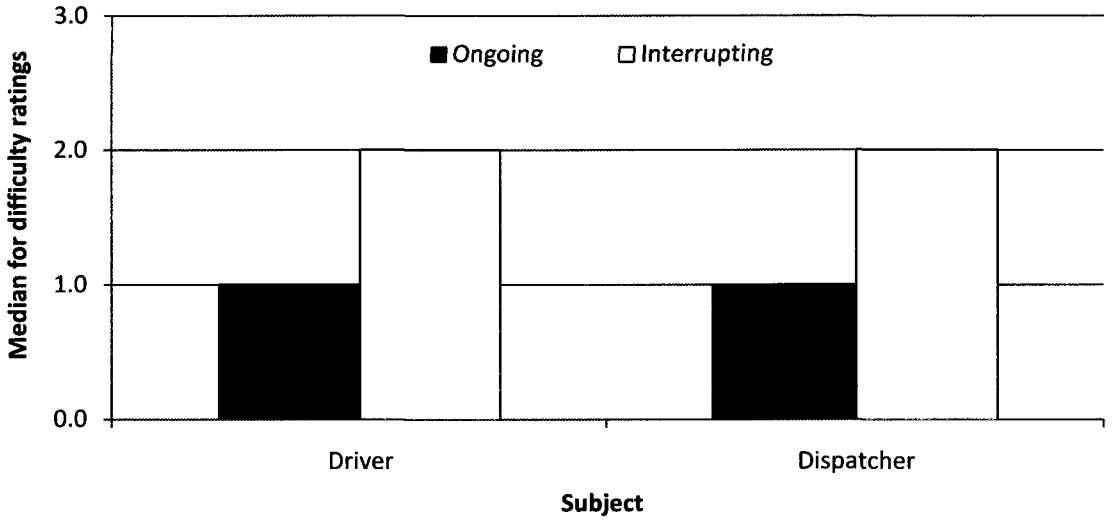


Figure 5.47 Median for difficulty ratings for the ongoing and the interrupting tasks for the drivers and the dispatchers.

Statistical analysis showed that the drivers and the dispatchers rated the interrupting task as significantly more difficult than the ongoing task ($F(1,15)=6.25, p=0.002$). It is important to understand that ANOVA analysis might not be applicable to the data from Likert scales [88], but the same conclusion is supported by the median values. Figure 5.47 shows the median difficulty ratings for the drivers and the dispatchers. This demonstrates that the subjects realized that the tasks had different difficulties, which is consistent with the performance measures. The same conclusion is confirmed by inspecting the histograms of the ratings in Figure 5.46. This conclusion implies that the subjects expected the interrupting task to be more difficult and, therefore, could prepare themselves to pay extra attention to it. For the drivers this could be the cause of the decreased driving performance during the interrupting task as shown in section 5.4 (pg. 116).

5.8 Observations

The current experiment setup was not designed to make conclusions about some trends observed in the data. We still felt compelled to share our observations, because they could contribute to future research, which we describe in Chapter 7.

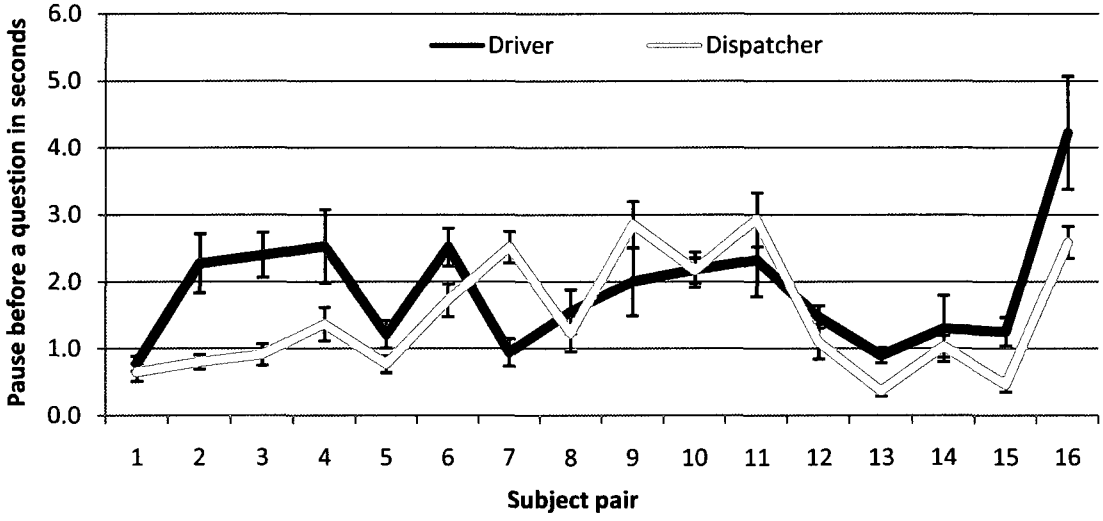


Figure 5.48 Pause before questions for different subject pairs.

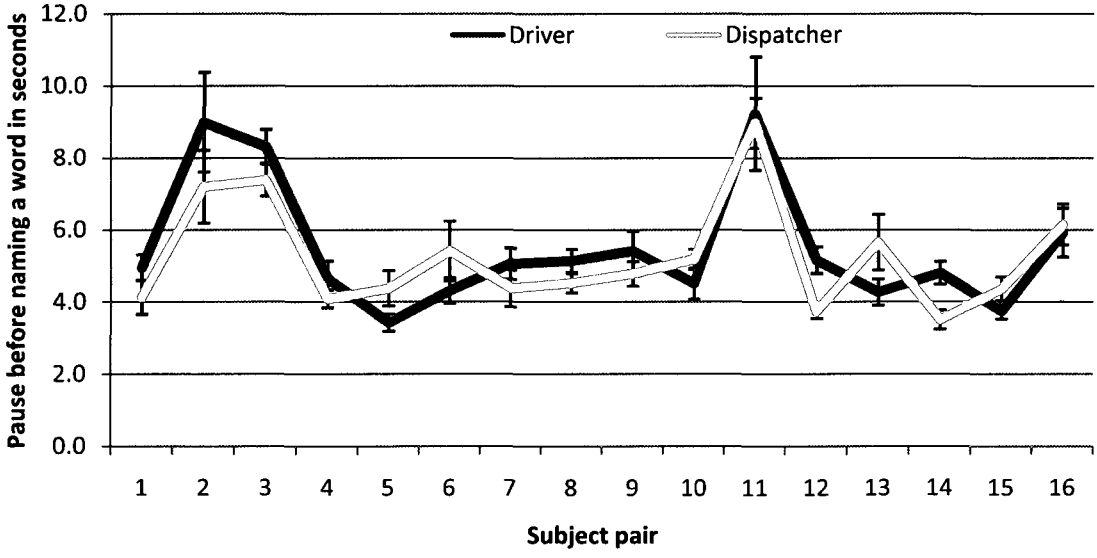


Figure 5.49 Pause before naming a word for different subject pairs.

Figure 5.48 and Figure 5.49 show the average duration of a pause before asking a question or naming a word for all subject pairs. These plots suggest that subjects adapted their speech to each other, which is consistent with the findings of Oviatt et al. [89]. Even though it is clear that different subjects have different pause durations there is a significant correlation between the subjects for the ongoing task ($r(16)=0.502, p=0.048$) and the interrupting task ($r(16)=0.840, p<0.001$). Figure 5.50 shows the speaking rate for the drivers and the dispatchers for different subject pairs during the ongoing task. Figure 5.51 shows the speaking rate during the interrupting task. The correlation between the drivers and the dispatchers is not significant ($r(16)=0.322, p=.224$) during the ongoing task, but it is significant ($r(16)=0.821, p<0.001$) for the interrupting tasks. It seems that the subjects are adapting to each other more during the interrupting task then during the ongoing task. For our research it means that the performance measures for the spoken tasks could be affected not just by driving, but also by the behavior of the dispatchers. For example, a driver might slow down in verbal response not because of the difficulty of the driving task, but because he is adapting to the slow pace of his dispatcher.

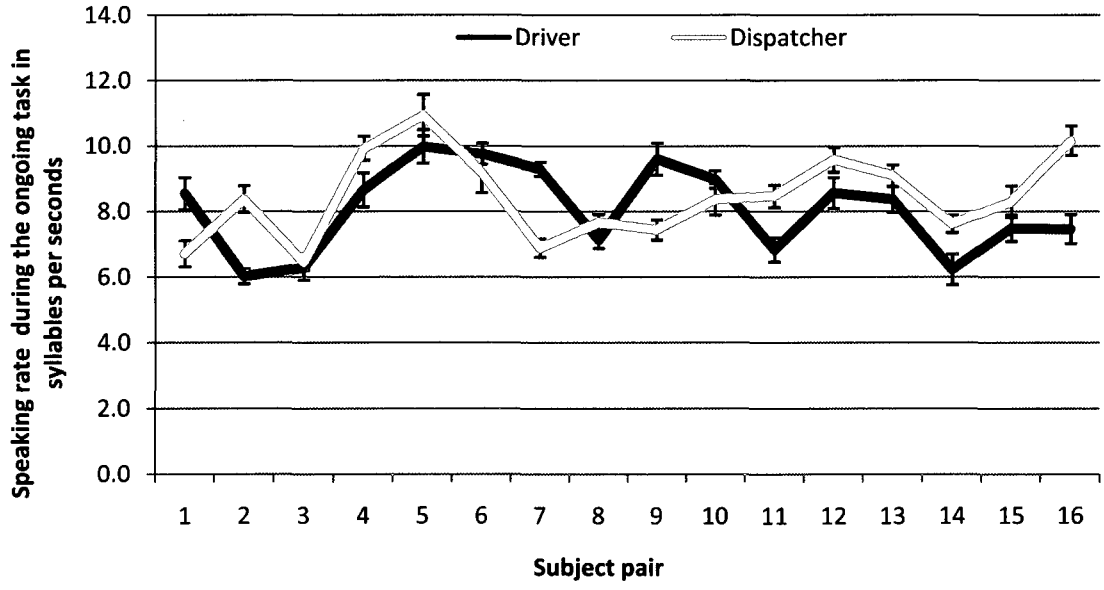


Figure 5.50 Speaking rate during the ongoing task.

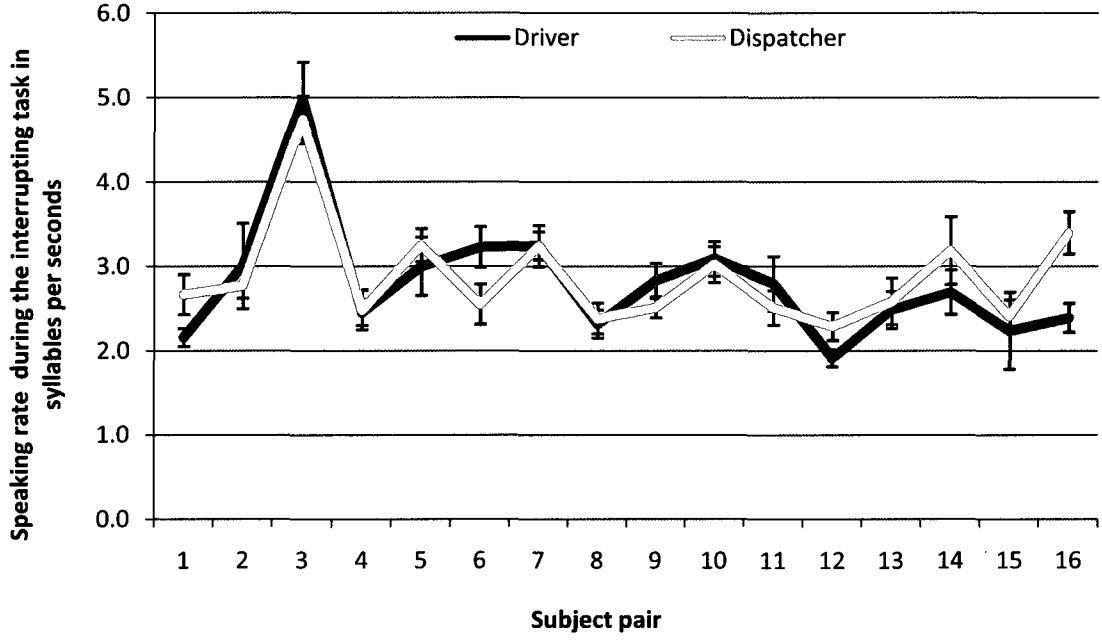


Figure 5.51 Speaking rate during the interrupting task.

Figure 5.6 and Figure 5.7 (section 5.2, pg. 96) show that both subjects learn during the duration of the experiment, but the plots do not exhibit a gradual adaptation. Figure 5.52 shows speaking rate during different games (averaged for all experiments).

This plot also does not exhibit a gradual adaptation between the subjects. Overall, we were not able to find that the subjects adapt to each other more as the experiments progressed, which could imply that the adaptation, if any, happens quickly.

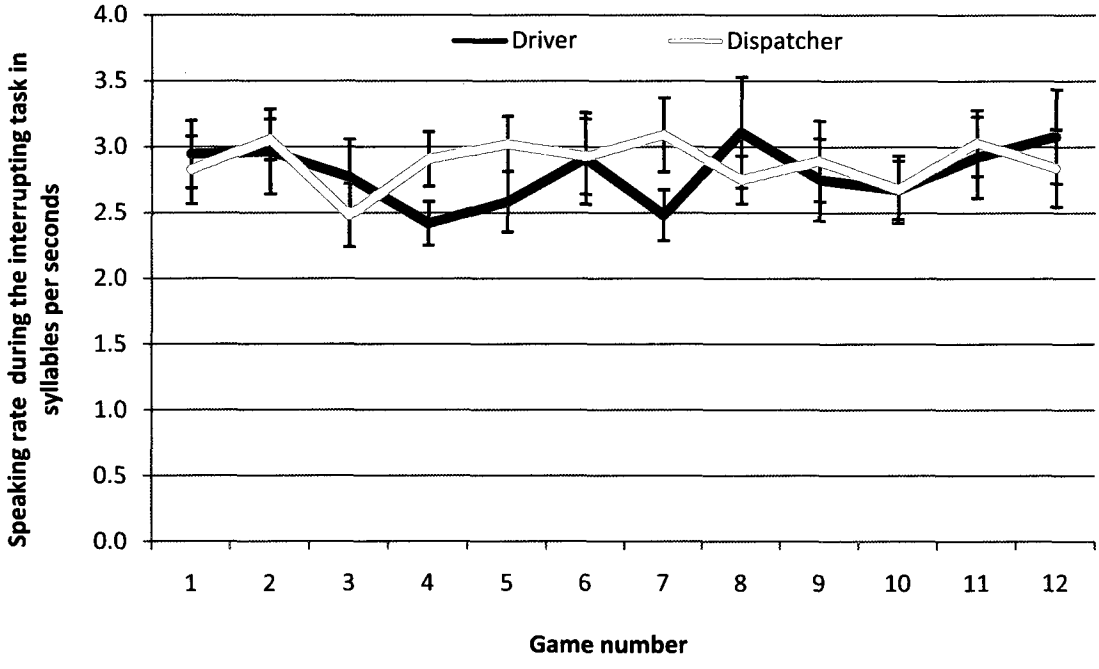


Figure 5.52 Speaking rate during the interrupting task for different games.

Driving performance measures (section 5.4, pg. 116) suggest that the drivers allocated more attention to driving during the ongoing task. It could be that because more attention was given to the interrupting task, the subjects adapted better to each other during the interrupting task. In order to test this hypothesis we would need to switch drivers and dispatchers between different pairs. The data do not show who is adapting to whom. It seems logical to assume that because the driver has to drive the dispatcher has more resources to adapt. On the other hand, the adaptation could be subconscious and both the drivers and the dispatchers change their behavior. We leave further elaboration on the subject to future research.

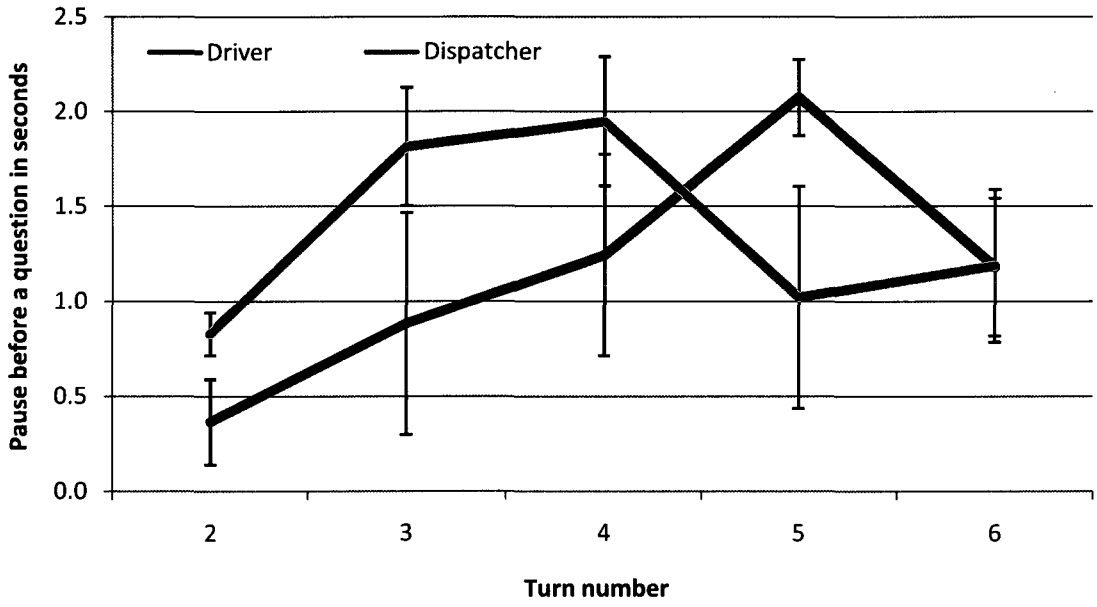


Figure 5.53 Pause before a question as a function of the turn number.

Figure 5.53 shows that the pause before asking a question depends on the turn number. To build this graph we removed all unsuccessful games and focused on the games that had exactly 6 turns. For example, if the driver finished the ongoing task in five and less turns, or seven and more turns, or the driver failed the game, then we would exclude this game from the analysis. In other words, we used the data only from the games that had 6 complete turns for the twenty questions game, which is the largest subset of games (27% or 103 games as shown in Figure 5.1). The shape of the curves in Figure 5.53 is consistent with the predictions of Art-R models [24], which state that the more items a person must recall the longer it takes to recall them. For the ongoing task the very first question is simple, because there are only three rooms to choose from. The very last question is simple because by this time it is clear what the object is. On the other hand, the measure for every turn might be biased by the presence of an interruption. Figure 5.54 shows the pause before a question for games that were not interrupted. These games were completed before an interruption happened. Given that we had only 15

games (4%) that were not interrupted we cannot make a strong conclusion and, hence, defer elaboration on the subject to future research.

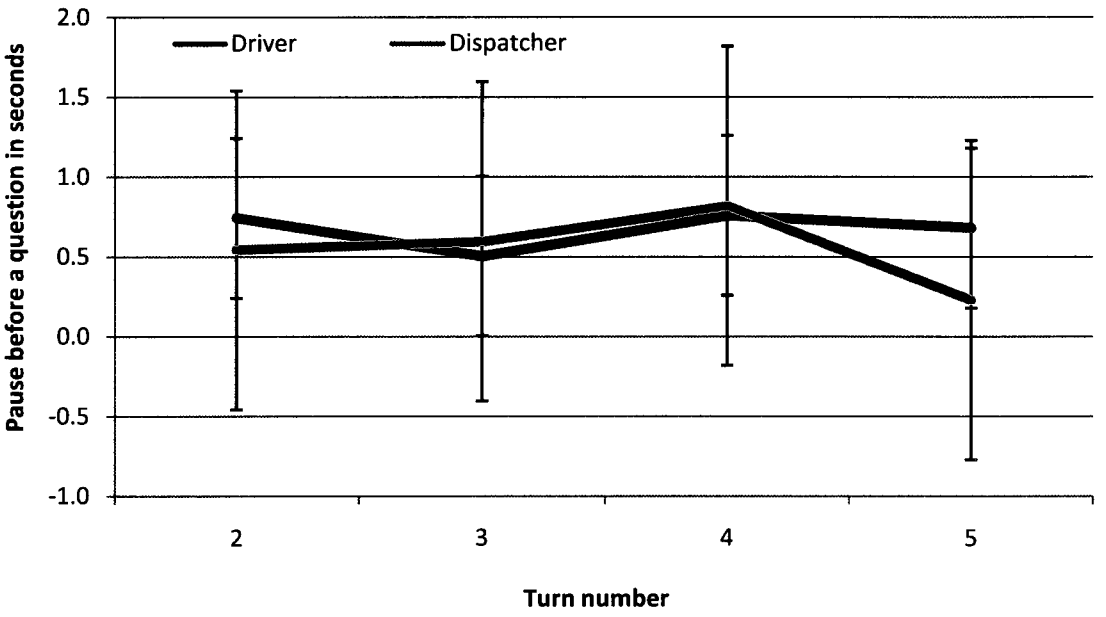


Figure 5.54 Pause before a question as a function of the turn number for uninterrupted games.

CHAPTER 6

CONCLUSION

The problem we are addressing in this work is the lack of knowledge about the interaction between multi-threaded dialogs and a manual-visual task. We designed experiments that utilized driving as an example of a manual visual task, and two spoken tasks as a basis for our multi-threaded dialog. Our goals were to look at the interaction between the performance measures in driving and the spoken tasks, and how people manage multi-threaded spoken dialogs while driving. We designed and ran the experiments. We analyzed the collected data, and in our conclusion we will go over our findings and summarize our contributions.

6.1 Spoken task performance while driving

Hypothesis 1 stated that a spoken task performance degrades in the presence of driving. We found indications that driving influenced the twenty questions game, because drivers made more wrong guesses than the dispatchers (section 5.3, pg. 104), but this difference was not significant. We hypothesize that increasing the difficulty of the ongoing task by increasing the number of participating objects (as explained in section 4.3, pg. 58) will result in a larger impact of driving on the ongoing task. We did not find

indications that driving affected any performance measure of the last letter word game. We hypothesize that this difference between the ongoing task and the interrupting task is caused by the difference in perceived urgency or difficulty of the tasks. This means if the last letter word game would not be perceived as urgent, then we would see degradation of the task performance in the presence of driving. It is important to notice that for certain interruption timings we did observe the impact of driving on both spoken tasks, as discussed below in section 6.3.

We also predicted that more demanding driving conditions will negatively influence the spoken tasks. The data (section 5.5, pg. 126) did not show that driving difficulty influenced our spoken tasks. This might be due to the fact that the difference between driving difficulties for straight and curvy roads were not big enough to produce noticeable changes in the spoken tasks. In other words, our assumption about the difficulty of the road curvature as compared to the straight road was not correct (section 4.6, pg. 71).

6.2 Spoken tasks affect driving performance

Hypothesis 2 stated that the spoken tasks affect driving performance. Our data testify that two different spoken tasks affected driving differently. The last letter word game affected driving more than the twenty questions game. For example, the lane position variance increases during the last letter word game as compared to the lane position variance before the interruption (section 5.4, pg. 116). This finding is consistent with the results found by Strayer and Johnston [52]. Wickens acknowledges that the multiple resource model cannot properly explain this difference [1]. The multiple

resource model states that if the tasks are separated in all dimensions from each other, there should be no performance decrements in either task, because no resources are shared. In our experiment different spoken tasks affected driving differently, which cannot be explained using multiple resource model. We suggest that the urgency associated with the last letter word game caused the driver to focus more on the last letter word game, which resulted in the neglect of the driving task. It also could be that the expected difficulty of the task changed how the task was handled (section 5.7, pg. 146).

6.3 Timing of a switch influences spoken tasks

Hypothesis 3 stated that there is an interaction between the time when a second dialog thread interrupts the first dialog thread and the performance associated with the spoken tasks. We found that the timing of an interruption affects the drivers and the dispatchers differently. The drivers were affected by the timing of interruptions, while the dispatchers were not. For example, for turn based interruption timings the drivers had a longer pause before asking a question during early interruptions as compared to the dispatchers, or when comparing drivers' pauses between early and middle interruptions (section 5.3.1, pg. 107). Similarly, we found that according to level based interruption timings the drivers had longer pause before naming a word during early interruptions when comparing to the pauses for middle and late interruptions (section 5.3.2, pg. 113). It seems that the additional load imposed by driving resulted in such an effect. This implies that dialog management has increased importance for drivers, because a driver can be affected by poor dialog management performance more than a person not engaged in a manual-visual task.

We did not find an interaction between driving difficulty and timing of interruptions which might be expected given the conclusion above. We hypothesize that this might be due to our wrong assumption about driving difficulty as explained in section 6.1.

6.4 Switching behavior

Hypothesis 4 stated that people utilize a number of switching behaviors during their interactions. We found indications that the drivers and the dispatchers might use different switching behavior, but the trend was not significant. We suggest that the trend was not significant because the levels of the road difficulty were not properly chosen (section 4.6, pg. 71). Still, the drivers seem to use signaling for finishing the interrupting task less often as compared to the dispatchers (section 5.6.2, pg. 135). This could be explained by the additional workload caused by the driving task. Another possibility could be that the drivers chose to speak less in order to focus on the driving task. This would mean that the drivers are aware of the increased workload and attempt to maintain the driving performance. We also found that, in relation to the adjacency pairs, the drivers and the dispatchers introduce interruptions similarly. This could imply that the process of decision making of when to interrupt was not affected by the driving task or that the drivers did not allow the driving to affect their decision making process.

6.5 Urgency of the interrupting task

Hypothesis 5 stated that more urgent interrupting tasks will be dealt with more quickly. The data from the navigation experiment described in Chapter 3 did not show that the urgency of the interrupting task changed how the drivers reacted to the task

(section 3.12, pg. 43). It could be that the subjects choose to react as quickly as possible, because they were given instructions to complete the tasks quickly. It also could be that the difference in the levels of urgency was too small to encourage a changed behavior. The data from our twenty questions experiment described in Chapter 4 suggested that the urgency of the task might influence how the tasks are performed if we compare a task that have urgency associated with it and a task that does not (section 5.4, pg. 116).

6.6 Goal 1

Our first goal was to investigate the interaction between multi-threaded dialogs formed by two spoken tasks and driving. The data collected from our experiments (described in Chapter 3 and Chapter 4) showed that there is, indeed, an interaction. Moreover, the spoken tasks influence the driving performance, and driving influences the spoken tasks. We also found that this interaction was different for our spoken tasks as explained above.

It could be that the urgency associated with a task allows the shift of attention from one task to another, resulting in a degraded performance on the tasks that are perceived less urgent. On the other hand, the perceived difficulty of the tasks could create the same situation. In either case, the fact that the driving performance decreased during the interrupting task suggests that even through the driving task and the interrupting tasks must use different resources according to the multiple resource model [1], there is a shared resource between them, which can be allocated to one task or another (Vergauwe et al. [90] arrived to a similar conclusion using data from their own experiments). This means that the perceived urgency or the perceived difficulty of a task must be controlled

or else the driving performance will suffer. This is an important consideration for the design of the human-computer interactions.

We propose that resources shared by tasks are shared based on attention, which can shift any given resource to any given task, while ignoring the demands of the other task. For example, in case of the interrupting task, the driver focused on the interrupting task (allocated more attention to this task), and thus the driving performance suffered. With the ongoing task, the driver focused more on driving, and, as a result, the performance of the ongoing task suffered. This interpretation is consistent with the previous research [52,56]. MacDonald and Hoffmann [56] also concluded that a driver's strategy of attention allocation would affect the driving performance measures.

6.7 Goal 2

Our second goal was to investigate how people manage multi-threaded dialogs when one participant is driving a vehicle. Our experiment setup did not produce a range of different behaviors for the drivers and the dispatchers. We attribute this to the experiment setup, which allowed subjects to complete the tasks without using different behaviors. This implies that in some cases (as in this research) manual-visual task does not require a change in subject's behavior in order to complete required spoken tasks.

On the other hand, the data provided an interesting insight that, on average, the drivers and the dispatchers used the same number of turns in their games, but the drivers were slower than the dispatchers. We hypothesize that the drivers sometime have slower responses in the ongoing task due to the increased workload caused by the presence of the driving task. Theoretically, the drivers could have used a different strategy to cope

with the increased workload: instead of thinking longer about a question, they could have asked more questions while thinking less about each question (our data show that the drivers did not use this strategy). It is possible that we do not observe such a behavior because the drivers are unable to ask questions faster, which would indicate a limit caused by the cognitive load. We do not know whether the drivers chose to think longer or had to think longer. In either case, the exhibited behavior is an indicator that people might prefer a slower but more precise response from the computer rather than a faster but less precise response. This is based on the fact that the drivers had longer pauses during the games with early interruptions.

Collecting data about how different types of spoken tasks interfere with driving is an important step for understanding the connection between the cognitive load imposed by the different spoken and manual-visual tasks. This research provided data for an improvement of our understanding of how drivers can use speech to safely interact with proliferating in-car electronic devices.

6.8 Contributions

The first contribution is finding a spoken task (section 4.3, pg. 58) that satisfies the constraints (section 4.1, pg. 49) imposed by the presence of a manual visual task. We showed how this task can be used with another spoken task (section 4.4, pg. 66) to enable subjects to participate in a multi-threaded spoken dialog. We created an experiment setup (Chapter 4) that can be used to investigate the interaction between spoken tasks in a multi-threaded dialog and manual-visual tasks. We used driving as a manual-visual task,

but we envision the applicability of this experiment setup to research that uses other manual-visual tasks.

The second contribution of this dissertation is the corpora that we collected during our experiments (section 3.11, pg. 43 and section 5.1, pg. 89). The corpora allow researches in different disciplines (human-factors, computer science, linguistics, etc.) to select their assumptions for future research. For example, the data show how learning affects the twenty questions game (section 5.2, pg. 96), which might be a starting point for research on how learning in the twenty questions game is effected by different driving conditions. The corpora also contain data channels that were not used for analysis in this dissertation and these data channels are available for future investigations. For instance, Palinko et al. [87] use our eye-tracker data to estimate the cognitive load of the drivers based on the recorded pupil size.

The third contribution is the data analysis. We showed our findings about the interaction between the spoken tasks and driving, as well as, investigation of different behavior exhibited by the drivers and the dispatchers. We also showed observed trends in the data, such as indications for accommodation between the subjects. We found that driving affects the spoken tasks and the spoken tasks affect driving. The data collected in this research suggests that this interaction between driving and the spoken tasks cannot be explained by the multiple resource model [1]. The multiple resource model states that if the tasks are separated in all dimensions from each other, there should be no performance decrements in either task, because no resources are shared. The data do show that driving affects the spoken tasks, even though, according to the multiple resource model, they are not sharing the same resources. We suggest that subjects allocate different resources to

different tasks based on the perceived urgency or difficulty of the tasks. This implies that designers of systems that can handle multi-threaded dialogs in a vehicle should consider how the tasks urgency or difficulty is perceived by the users.

The following chapter outlines opportunities for future research that can utilize our contributions to further our understanding of interaction between multi-threaded dialogs and manual-visual tasks.

CHAPTER 7

FUTURE WORK

Our conclusions discussed in Chapter 6, as well as the trends visible in the data provided us with the ideas for future research. In this chapter we outline a few suggestions for future work based on our conclusions and results. Some of the suggestions will be aimed to improve the current experiment setup, while others will require completely new experiment setups.

7.1 Spoken task performance while driving

We found that driving influenced the ongoing task (twenty questions game), but did not influence the interrupting task (last letter word game). We hypothesize that the perceived urgency of tasks, and not the tasks themselves, is the cause. In order to test this hypothesis one can use the last letter word game as the ongoing task, and the twenty questions game as the interrupting task. If the new experiment setup shows the same trends for the ongoing and the interrupting tasks as in this research, then the difference in how driving influenced the spoken tasks cannot be attributed to the tasks. It is also important to ask the participants how they perceived the urgency of the tasks. This could help us to assess if, indeed, subjects perceive one task as more urgent than the other.

The difference in driving difficulties between straight and curvy roads did not allow us to see the influence of the road difficulty on the spoken tasks. We suggest that using turns with a smaller radius should create more difficulty difference between straight and curvy roads [20]. Introducing crosswinds along the road is another possibility [91] that could increase the driving difficulty. Increased difference in driving difficulty would allow us to see the nature of the interaction between driving difficulty and the spoken tasks.

7.2 Spoken tasks affect driving performance

We found that the ongoing task did not influence the driving as much as the interrupting task did. Similarly to the suggestions in section 7.1, switching the ongoing and the interrupting tasks might show the source of this difference. We hypothesize that the perceived urgency is the source of this situation. Alternatively, it is possible to instruct the drivers to treat the driving task as a priority, regardless of the current spoken task. This approach could force the drivers to maintain the driving performance and as a result one might see more degradation in the spoken task performance and less degradation in driving performance. On the other hand, if one knows that the drivers make their best effort to focus on the driving, then one can judge how much the spoken tasks interfere with driving.

Increasing the driving difficulty, as suggested in section 7.1, can highlight the effects that the spoken tasks have on driving performance. Increasing the spoken tasks difficulty also might create more interference with driving. Our data suggest that there is

a relationship between driving difficulty and the spoken tasks difficulty. Manipulating these difficulties in an experiment would allow one to investigate this relationship.

7.3 Timing of a switch influences spoken tasks

We only observed the effect of early interruptions on the spoken tasks. We did not monitor the emotional state of the participants, which according to the previous research [38] might be affected by the timing of interruptions. It could be beneficial to use physiological measurements [43,44,46] to track the emotional state of the subjects. These measurements can also be used to estimate of the cognitive load for the participants. The cognitive load estimation should also help with computational approach for multiple resource model as described by Horrey and Wickens [50].

Horrey and Wickens [50] developed a computational model for the multiple-resource model. Current experiment design did not produce large variability in performance measures. If our experiment design is modified to produce more variability in task performance measures, then it will be possible to compare the measured values with predictions of the computational model. Introducing more variation into spoken tasks or driving difficulty should produce more changes in the performance measures.

Currently, Palinko et al. [87] use data from our experiments to estimate the cognitive load of the drivers based on the recorded pupil size. New information about what cognitive load is experienced by the drivers could allow one to investigate the relationship between the spoken tasks and driving from a new prospective.

7.4 Switching behavior

The small number of the resumption activities in our latest experiment setup (section 5.6.2, pg. 135) calls for the increased difficulty of the interrupting task. This can be accomplished by increasing the number of words a person must name during the interrupting task, or by providing additional restrictions on words that can be used. For example, subjects could be restricted to name only food items that have only four letters. On the other hand, increasing the difficulty of the twenty questions game by increasing the number of participating objects might also result in an increase of the resumption activities.

7.5 Urgency of the interrupting task

Data from the navigation experiment (section 3.12, pg. 43) did not show that the subjects were affected by the urgency level. We hypothesize that the lack of differentiation between two urgency levels was the cause. One can use a longer time delay for non-urgent interruptions to make it clear to the subjects that the urgency levels are different.

For the twenty questions experiment (Chapter 4), similar to our suggestions in section 7.2, explicitly specifying task priorities for the tasks could allow us to see if the perceived urgency effected the performance measurements. It is also possible to use the same tasks, but remove the time limit for the interrupting task. By comparing the new data to the data from the current experiment (Chapter 5) one could find if the urgency associated with the tasks had an effect on the performance measures.

7.6 More suggestions

Adding more events to the simulation scenario, such as sudden brakes of the leading vehicle, could allow one to measure brake reaction times of a driver [54,85]. These measures could provide more information about the driver's attention to the road on a small time scale. For instance, by timing the stimulus for braking for particular subtasks in the ongoing spoken task (asking a question, answering a question, etc.) it would be possible to compare how the driver's attention changes during these subtasks. This information would allow one to locate the parts of the spoken tasks that create the most interference with the driving task.

The current experiment design did not test the effects of the interruptions on the ongoing task. By having a baseline by allowing the subjects to perform the ongoing task without any interruptions, one can compare the subject's performance on the ongoing task before and after interruptions. This comparison with the baseline could show how long the interruptions disrupt the ongoing task and driving.

The data from the current experiment suggest that humans exhibit adaptive behavior, which is in agreement with the previous work by Oviatt et al. [89]. There is also research on convergence during conversational interactions, which suggest that people adapt their speech to match each other. For example, Pardo [92] found phonetic convergence during spoken interaction in Map Task corpus [69]. We hypothesize that the dispatchers are more likely to adapt to the drivers than vice versa. This can be tested by pairing different drivers and dispatchers to see how they adapt to each other. When doing this, one must be careful to manage learning effects of the participants. In addition, it is possible to correlate variables from different pairs of subjects by randomly pairing

dispatchers and drivers [93]. This also might test if drivers and dispatchers adapt to each other.

The "Wizard of Oz" approach [10] can also be used to manipulate how the system responds to the user, to see how subjects adjust to these changes. In the "Wizard of Oz" approach, the drivers will think that they are playing the games with a computer, while in fact, there is a person controlling the computer. It will be possible to compare the results between the new setup and this research to see if drivers use the same methods when performing tasks with another person or when doing these tasks with a computer.

LIST OF REFERENCES

[1] C. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, 2002, pp. 159–177.

[2] S. Lowe, *Many Workers Have Long Commutes to Work*, US Census Bureau Press Release, CB05-CR.01, 2005.

[3] P. Green, "Crashes Induced by Driver Information Systems and What Can Be Done to Reduce Them," Society of Automotive Engineers, 2000, pp. 26-36.

[4] J. McCarley, M. Vais, H. Pringle, A. Kramer, D. Irwin, and D. Strayer, "Conversation disrupts visual scanning of traffic scenes," *The 9th Vision in Vehicles Conference*, Australia, 2001.

[5] J. Lee, B. Caven, S. Haake, and T. Brown, "Speech-Based Interaction with In-Vehicle Computers: The Effect of Speech-Based E-Mail on Drivers' Attention to the Roadway," *Human Factors*, vol. 43, 2001, pp. 631-640.

[6] N. Memarovic, "The influence of personal navigation devices on drivers' visual attention on the road ahead and driving performance," MS Thesis, University of New Hampshire, 2009.

[7] Governors Highway Safety Association, "Cell Phone Driving Laws," http://www.ghsa.org/html/stateinfo/laws/cellphone_laws.html (Last accessed January 2010).

[8] A. Barón and P. Green, *Safety and Usability of Speech Interfaces for In-vehicle Tasks While Driving: A Brief Literature Review*, Technical Report UMTRI-2006-5, University of Michigan, Transportation Research Institute, 2006.

[9] O. Tsimhoni, P. Green, and J. Lai, "Listening to Natural and Synthesized Speech while Driving: Effects on User Performance," *International Journal of Speech Technology*, vol. 4, April 2001, pp. 155-169.

[10] Z. Medenica and A. Kun, "Comparing the influence of two user interfaces for mobile radios on driving performance," *Driving Assessment*, Stevenson, WA, 2007.

[11] A. Shyrokov, A. Kun, and P. Heeman, "Experimental Modeling of Human-Human Multi-Threaded Dialogues in the Presence of a Manual-Visual Task," *The 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 2007.

[12] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, The MIT Press, 2005.

[13] H. Clark and S. Brennan, "Grounding in communication," *Perspectives on socially shared cognition*, 1991, pp. 127–149.

[14] P. Kenne and M. O'Kane, "Topic change and local perplexity in spoken legal dialogue," *The 4th International Conference on Spoken Language*, 1996, pp. 721-724.

[15] M. O'Kane and P. Kenne, "Changing the topic: how long does it take?," *The 4th International Conference on Spoken Language*, 1996, pp. 1041-1044.

[16] D. Byron and P. Heeman, "Discourse Marker Use in Task-Oriented Spoken Dialog," *The 5th European Conference on Speech Communication and Technology*, 1997, pp. 2223-2226.

[17] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, 1986, pp. 175–204.

[18] F. Yang, P. Heeman, and A. Kun, "Switching to real-time tasks in multi-tasking dialogue," *The 22nd International Conference on Computational Linguistics*, Manchester, United Kingdom, 2008, pp. 1025-1032.

[19] DriveSafety, Inc., "DS-600c Research Simulator," http://www.drivesafety.com/LinkClick.aspx?fileticket=T0KPJ0BlXRM%3d&tabid=97&mid=536, (Last accessed January 2010).

[20] S. Yee, L. Nguyen, and P. Green, *Visual, Auditory, Cognitive, and Psychomotor Demands of Real In-Vehicle Tasks*, Technical Report UMTRI-2006-20, University of Michigan Transportation Research Institute, 2007.

[21] J. Bellegarda and K. Silverman, "Natural Language Spoken Interface Control Using Data-Driven Semantic Inference," *Transactions on Speech and Audio Processing*, vol. 11, April 2003, pp. 267-277.

[22] A. Kun, W. Miller, and W. Lenharth, "Project54: Standardizing electronic device integration in police cruisers," *IEEE Intelligent Systems*, vol. 18, October 2003, pp. 10-13.

[23] E. Schegloff and H. Sacks, "Opening up closings," Semiotica VIII, 1973, pp. 290-327.

[24] J. Anderson and C. Lebiere, *The Atomic Components of Thought*, Lawrence Erlbaum, 1998.

[25] F. Yang and P. Heeman, "Context restoration in multi-tasking dialogue," *The 13th international conference on Intelligent user interfaces*, Sanibel Island, Florida, USA, Association for Computing Machinery, 2009, pp. 373-378.

[26] R. Slick, E. Kim, D. Evans, and J. Steele, "Using Simulators to Train Novice Teen Drivers: Assessing Psychological Fidelity as a Precursor of Transfer of Training," *Asian Conference on Driving Simulation*, 2006.

[27] C. Rosé, B. Eugenio, L. Levin, and C. Ess-Dykema, "Discourse processing of dialogues with multiple threads," *The 33rd annual meeting on Association for Computational Linguistics*, Cambridge, Massachusetts, Association for Computational Linguistics, 1995, pp. 31-38.

[28] O. Lemon, A. Gruenstein, A. Battle, and S. Peters, "Multi-tasking and Collaborative Activities in Dialogue Systems," *The 3rd SIGdial Workshop on Discourse and Dialogue*, 2002.

[29] O. Lemon and A. Gruenstein, "Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments," *ACM Transactions on Computer-Human Interaction*, vol. 11, 2004, pp. 241-267.

[30] P. Heeman, F. Yang, A. Kun, and A. Shyrokov, "Conventions in Human-Human Multi-Threaded Dialogues: A Preliminary Study," *The 10th international conference on Intelligent user interfaces*, New York, NY, USA, Association for Computing Machinery, 2005, pp. 293-295.

[31] A. Oh, H. Fox, M.V. Kleek, A. Adler, K. Gajos, L. Morency, and T. Darrell, "Evaluating look-to-talk: a gaze-aware interface in a collaborative environment," *Conference on Human Factors in Computing Systems*, Minneapolis, Minnesota, USA, Association for Computing Machinery, 2002, pp. 650-651.

[32] I. McCowan, D. Gatica-Perez, S. Bengio, D. Moore, and H. Bourlard, "Towards Computer Understanding of Human Interactions," *Machine Learning for Multimodal Interaction*, 2005, pp. 56-75.

[33] A. Paivio, *Imagery and Verbal Processes*, New York, Holt, Rinehart and Winston, 1971.

[34] E. Arroyo, T. Selker, and A. Stouffs, "Interruptions as multimodal outputs: Which are the less disruptive?," The 4th IEEE International Conference on Multimodal Interfaces, 2002, pp. 479-483.

[35] T. Gillie and D. Broadbent, "What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity," *Psychological Research*, vol. 50, 1988, pp. 243-250.

[36] Y. Miyata and D. Norman, "Psychological issues in support of multiple activities," *User Centered Systems Design: New Perspectives on Human-Computer Interaction*, Hillsdale, NJ, USA, L. Erlbaum Associates Inc., 1986, pp. 265-284.

[37] B. Bailey, J. Konstan, and J. Carlis, "Adjusting windows: Balancing information awareness with intrusion," *The 6th Conference on Human Factors and the Web*, 2000.

[38] P. Adamczyk and B. Bailey, "If not now, when?: The effects of interruption at different moments within task execution," *Conference on Human Factors in Computing System*, New York, NY, USA, Association for Computing Machinery, 2004, pp. 271-278.

[39] D. McFarlane, "Coordinating the Interruption of People in Human-Computer Interaction," *The 13th International Conference on Human-Computer Interaction*, 1999, pp. 295-303.

[40] S. Rubio, E. Diaz, J. Martin, and J.M. Puente, "Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods," *Applied Psychology*, vol. 53, 2004, p. 61.

[41] B. Bailey and S. Iqbal, "Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management," *Transactions on Computer Human Interaction*, 2008.

[42] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," *The 2008 symposium on Eye tracking research & applications*, 2008, pp. 69–72.

[43] K. Brookhuis, "Psychophysiological methods," *Handbook of Human Factors and Ergonomics Methods*, Boca Raton, FL, CRC Press, 2004.

[44] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (GSR) as an index of cognitive load," *Conference on Human Factors in Computing Systems*, New York, NY, USA, Association for Computing Machinery, 2007.

[45] S.G. Hart and L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, 1988, pp. 139–183.

[46] J. Healey and R. Picard, "Smartcar: Detecting driver stress," *International conference on pattern recognition*, 2000, pp. 218–221.

[47] M. Recarte and L. Nunes, "Effects of Verbal and Spatial-Imagery Tasks on Eye Fixations While Driving," *Journal of Experimental Psychology Applied*, vol. 6, 2000, pp. 31-43.

[48] W. Horrey, C. Wickens, and K. Consalus, "Modeling Drivers' Visual Attention Allocation While Interacting With In-Vehicle Technologies," *Journal of experimental psychology applied*, vol. 12, 2006, p. 67.

[49] C. Wickens, "Processing resources and attention," *Multiple-task performance*, 1991, pp. 3–34.

[50] W. Horrey and C. Wickens, "Multiple resource modeling of task interference in vehicle control, hazard awareness and in-vehicle task performance," *The 2nd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 2003, pp. 7–12.

[51] D. Strayer, F. Drews, R. Albert, and W. Johnston, "Cell phone induced perceptual impairments during simulated driving," *Driving Assessment*, 2001.

[52] D. Strayer and W. Johnston, "Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone," *Psychological Science*, vol. 12, November 2001, pp. 462-466.

[53] V. Neale, T. Dingus, S. Klauer, J. Sudweeks, and M. Goodman, "An Overview of the 100-car Naturalistic Study and Findings," *The 19th International Technical Conference on Enhanced Safety of Vehicles*, 2005.

[54] P. Green, "How Long Does It Take to Stop?" Methodological Analysis of Driver Perception-Brake Times," *Transportation Human Factors*, vol. 2, 2000, pp. 195-216.

[55] H. Jamson and N. Merat, "Can low cost road engineering measures combat driver fatigue? A driving simulator investigation," *The 5th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2009.

[56] W. MacDonald and E. Hoffmann, "Review of relationships between steering wheel reversal rate and driving task demand," *Human Factors*, vol. 22, 1980, pp. 733-739.

[57] O. Tsimhoni and P. Green, "Visual Demand of Driving Curves Determined by Visual Occlusion," *Vision in Vehicles*, vol. 8, 1999, pp. 22-25.

[58] P. Green, *Visual and task demands of driver information systems*, Technical report, UMTRI-98-16, Transportation Research Institute, 1999.

[59] D. Strayer, F. Drews, and W. Johnston, "Cell Phone-Induced Failures of Visual Attention During Simulated Driving," *Journal of experimental psychology applied*, vol. 9, 2003, pp. 23-32.

[60] S. Chisholm, J. Caird, J. Lockhart, L. Fern, and E. Teteris, "Driving Performance while Engaged in MP-3 Player Interaction: Effects of Practice and Task Difficulty on PRT and Eye Movements," *The 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 2007, pp. 9–12.

[61] D. Lamble, M. Laakso, and H. Summala, "Detection thresholds in car following situations and peripheral vision: implications for positioning of visually demanding in-car displays," *Ergonomics*, vol. 42, 1999, pp. 807-815.

[62] O. Tsimhoni, P. Green, and H. Watanabe, "Detecting and Reading Text on HUDs: Effects of Driving Workload and Message Location," *The 11th ITS meeting. ITS connecting the Americas*, Miami Beach, FL, Intelligent Transportation Society of America, 2001.

[63] H. Lew, J. Poole, E. Lee, D. Jaffe, H. Huang, and E. Brodd, "Predictive validity of driving-simulator assessments following traumatic brain injury: a preliminary study," *Brain Injury*, vol. 19, 2005, pp. 177-188.

[64] A. Kemeny and F. Panerai, "Evaluating perception in driving simulation experiments," *Trends in Cognitive Sciences*, vol. 7, January 2003, pp. 31-37.

[65] R. Mourant and T. Thattacherry, "Simulator Sickness in a Virtual Environments Driving Simulator," *Annual meeting*, Human factors and ergonomics society, 2000, pp. 534-537.

[66] M. Vollrath, "Speech and driving-solution or problem?," *Intelligent Transport Systems*, vol. 1, 2007, pp. 89–94.

[67] J. Villing, C. Holtelius, S. Larsson, A. Lindstrom, A. Seward, and N. Aberg, "Interruption, resumption and domain switching in in-vehicle dialogue," *Lecture Notes in Computer Science*, vol. 5221, 2008, pp. 488–499.

[68] A. Lindstrom, J. Villing, and S. Larsson, "The effect of cognitive load on disfluencies during in-vehicle spoken dialogue," *Interspeech*, 2008.

[69] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Garrod, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, 1991, pp. 351-366.

[70] A. Shyrokov, *Setting up experiments to test a multi-threaded speech user interface*, Technical report, ECE.P54.2006.1, University of New Hamsphire, ECE, Project54, 2006.

[71] S. Nakajima and J. Allen, "A study on prosody and discourse structure in cooperative dialogues," *Phonetica*, vol. 50, 1993, pp. 197–210.

[72] W. Horrey and C. Wickens, "Focal and ambient visual contributions and driver visual scanning in lane keeping and hazard detection," *The 48th Human Factors and Ergonomics Society annual meeting*, 2004, pp. 2325–2329.

[73] D.E. Knuth, "Optimum binary search trees," *Acta Informatica*, vol. 1, May 1971, pp. 14-25.

[74] J. Kowtko, S. Isard, and G. Doherty, *Conversational games within dialogue*, Technical Report HCRC/RP-31, Human Communication Research Centre, University of Edinburgh, 1992.

[75] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. Kowtko, and A. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, 1997, pp. 13–31.

[76] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, 1972, pp. 283-292.

[77] O. Palinko, "Effects of different push-to-talk solutions on driving performance," MS Thesis, University of New Hampshire, 2008.

[78] A. Shyrokov, *Experiment Wizard 1.0*, http://expwiz.sourceforge.net/.

[79] C. Schmitz, *LimeSurvey 1.86*, http://www.limesurvey.org/.

[80] R. Kirk, *Experimental Design: Procedures for Behavioral Sciences*, Wadsworth Publishing, 1994.

[81] W. Stine, *Personal communication*, Associate Professor of Psychology, Department of Psychology, University of New Hampshire.

[82] T. Paek, *Personal communication*, Researcher, Microsoft Research.

[83] A. Kun, T. Paek, Ž. Medenica, N. Memarović, and O. Palinko, "Glancing at personal navigation devices can affect driving: experimental results and design implications," *The 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Essen, Germany, Association for Computing Machinery, 2009, pp. 129-136.

[84] J. McLean and E. Hoffmann, "Steering reversals as a measure of driver performance and steering task difficulty," *Human Factors*, vol. 17, 1975, pp. 248-256.

[85] W. Consiglio, P. Driscoll, M. Witte, and W. Berg, "Effect of cellular telephone conversations and other potential interference on reaction time in a braking response," *Accident Analysis & Prevention*, vol. 35, July 2003, pp. 495-500.

[86] S. Amado and P. Ulupinar, "The effects of conversation on attention and peripheral detection: Is talking with a passenger and talking on the cell phone different?," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, November 2005, pp. 383-395.

[87] O. Palinko, A. Kun, A. Shyrokov, and P. Heeman, "Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator," *Eye Tracking Research and Applications*, 2010.

[88] S. Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, 1946, pp. 677-680.

[89] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *Transactions on Computer-Human Interaction*, vol. 11, 2004, pp. 300-328.

[90] E. Vergauwe, P. Barrouillet, and V. Camos, "Do Mental Processes Share a Domain-General Resource?," *Psychological Science*, 2010.

[91] J. Klasson, "A generalised crosswind model for vehicle simulation purposes," *The 17th IAVSD Symposium, The Dynamics of Vehicles on Roads and on Tracks*, 2003, pp. 350–359.

[92] J. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, April 2006, pp. 2382-2393.

[93] P. Heeman, *Personal communication*, Associate Research Professor, Department of Science and Engineering, Oregon Health and Science University.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVALS

University *of* New Hampshire

Research Conduct and Compliance Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

28-Feb-2007

Kun, Andrew
Electrical & Computer Eng Dept
Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Approval Expiration Date: 23-Jul-2007
Modification Approval Date: 28-Feb-2007
Modification: Addition of surveys

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your modification to this study, as indicated above. Further changes in your study must be submitted to the IRB for review and approval prior to implementation.

Approval for this protocol expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects in this study. If your study is still active, you may request an extension of IRB approval.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

University *of* New Hampshire

Research Conduct and Compliance Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

18-Jul-2007

Kun, Andrew
Electrical & Computer Eng Dept
Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Review Level: Expedited
Approval Expiration Date: 23-Jul-2008

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your request for time extension for this study. Approval for this study expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects. If your study is still active, you may apply for extension of IRB approval through this office.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

University *of* New Hampshire

Research Conduct and Compliance Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

01-Nov-2007

Kun, Andrew
Electrical & Computer Eng Dept
Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Approval Expiration Date: 23-Jul-2008
Modification Approval Date: 31-Oct-2007
Modification: Collection of additional data (e.g. physiological measures) per 10/22/2007 email

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your modification to this study, as indicated above. Further changes in your study must be submitted to the IRB for review and approval prior to implementation.

Approval for this protocol expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects in this study. If your study is still active, you may request an extension of IRB approval.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F Simpson
Manager

University *of* New Hampshire

Research Conduct and Compliance Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

24-Jun-2008

Kun, Andrew L
Electrical & Computer Eng Dept, Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Review Level: Expedited
Approval Expiration Date: 23-Jul-2009

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your request for time extension for this study. Approval for this study expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects. If your study is still active, you may apply for extension of IRB approval through this office.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

University *of* New Hampshire

Research Integrity Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

16-Mar-2009

Kun, Andrew L
Electrical & Computer Eng, Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Approval Expiration Date: 23-Jul-2009
Modification Approval Date: 13-Mar-2009
Modification: Changes to consent form per 3/11/09 email

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your modification to this study, as indicated above. Further changes in your study must be submitted to the IRB for review and approval prior to implementation.

Approval for this protocol expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects in this study. If your study is still active, you may request an extension of IRB approval.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

University *of* New Hampshire

Research Integrity Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

09-Jul-2009

Kun, Andrew L
Electrical & Computer Eng Dept
Kingsbury Hall
Durham, NH 03824

IRB #: 2993
Study: ITR: Multi-Threaded Dialogues for Real-Time Applications
Review Level: Expedited
Approval Expiration Date: 23-Jul-2010

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved your request for time extension for this study. Approval for this study expires on the date indicated above. At the end of the approval period you will be asked to submit a report with regard to the involvement of human subjects. If your study is still active, you may apply for extension of IRB approval through this office.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the document, *Responsibilities of Directors of Research Studies Involving Human Subjects.* This document is available at http://www.unh.edu/osr/compliance/irb.html or from me.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

APPENDIX B

QUESTIONNAIRES

B.1 For navigation experiment

Personal information questionnaire

Subject ID: _____ Date: _____ Time: _____

Gender:
Female ___ Male ___

Age: ___

Are you a student?
No ___ Undergraduate ___ Graduate ___

If not a student, what is your highest education level?
High school ___ College ___ Graduate ___

Is English your native language?
Yes ___ No ___ but I've been speaking English for ___ years.

Are you left-handed or right-handed?
Left-handed ___ Right-handed ___

If you have a valid driver's license, what year you got it?
Exactly in _____ Approximately in _____
I do not remember _____ No driver's license _____

Approximately how often do you drive?
Never ___ A few times a month ___ A few times a week ___ Daily ___

Have you been in a driving simulator before? Check all that apply.
Never ___ Once or twice ___ Many times ___ At UNH ___

How well do you know your partner for the experiment?
We never met before _____ We never talked _____
We talk occasionally _____ We are friends _____

Approximately how often do you play video games?
Never ___ Once a month ___ Once a week ___ Daily ___

Experiment questionnaire

Subject ID: _____ Date: _____ Time: _____

Please indicate the level of agreement with each of the 14 statements belo.

The instructions at the beginning of the experiment were clear.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I understood what I had to do in the navigation task.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I understood what I had to do when a warning message appeared on the screen.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

Communication with the other person worked well.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

Training was sufficient.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The experiment was interesting.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The experiment was very short.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The experiment was very long.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The on-screen messages were frustrating.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

Car breakdowns were frustrating.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was satisfied with the team performance.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

Please use the space below to provide comments and suggestions about the study.

Questions for Police Officer

I gave driving a higher priority than reacting to on-screen messages.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The simulated road was difficult to drive on.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was comfortable driving in the simulator.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The dispatcher successfully guided me to my destination points.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was waiting until the intersection to provide information about an interruption.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was waiting for a straight part of a road to provide information about an interruption.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I did not need to provide feedback to the dispatcher, because he knew where I was.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was lost and dispatcher did not know where I was.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I learned the layout of the city and could navigate it by myself.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I responded to interruptions as quickly as I could.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

Questions for dispatcher

The police officer provided me with enough feedback.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

The police officer followed my directions well.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I knew where in the city the car was at all times.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

I was frustrated with the map.
Strongly Agree ___ / Agree ___ / Undecided ___ / Disagree ___ / Strongly Disagree ___

B.2 Twenty questions experiment

All questionnaires were presented using in a computerized form [77] and are presented here for completeness. All surveys were automatically marked with the proper experiment code and subject role.

Before experiment questionnaire

What is your gender? Please choose *only one* of the following: Female / Male

What is your age? _____

What is your level of education? Please choose *only one* of the following:
Freshman / Sophomore / Junior / Senior / 1st year graduate / 2nd year graduate / 3rd year graduate / More than 3 years of graduate school

Is English your native language? Please choose *only one* of the following:
Yes / No

Only answer this question if you answered No to the previous question

How many years are you using English for spoken communication? Please choose *only one* of the following:
1 / 2 / 3 / 4 / 5 / 5 to 10 / more than 10

For how many years have you been driving? Please choose *only one* of the following:
1 / 2 / 3 / 4 / 5 / 5 to 10 / more than 10

Indicate level of your agreement with the following statements:

I have a seasonal sickness (flue, cold, etc.).
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

I am in my usual state of fitness.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Did you participate in a driving simulator study before? Yes / No

Did you use UNH simulator before? Please choose *only one* of the following: Yes / No

How often do you play computer games (not counting card and puzzle games)? Solitaire and minesweeper do not count. Please choose *only one* of the following:
Every day / A few times a week / Once a week / A few times a month / Rarely / Never

How well do you know the other person participating in this experiment? Please choose *only one* of the following:
We never met before
We talked once or twice before
We talk occasionally
We talk regularly
We know each other very well

After experiment questionnaire

Please indicate the level of agreement with each of the statements below. Please choose the appropriate response for each item:

The instructions at the beginning of the experiment were clear.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

I understood what I had to do for the twenty questions game.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Training was sufficient.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

It was difficult to remember the questions to ask about the objects.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

List of objects was too long.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The tasks were very easy.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The experiment was interesting.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The experiment was very short.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The experiment was very long.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Communication with the other person worked well.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

I was satisfied with the team performance.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The other person responded very slowly.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Please indicate the level of agreement with each of the statements below
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Last Letter game was difficult
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

It was difficult to come up with new words for Last Letter game
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Only answer the following questions if you are a driverI understood what I had to do in the driving task.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The simulated road was difficult to drive on.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

I was comfortable driving in the simulator.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

I responded to interruptions as quickly as I could.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The on-screen messages were interfering with driving.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

The on-screen messages were obstructing my view.
Strongly disagree / Disagree / Undecided / Agree / Strongly agree

Simulator Sickness Questionnaire

Are you talking this survey before the experiment or after? Please choose *only one* of the following: Before the experiment / After the experiment

Please, provide information about how the following symptoms are affecting you right now. Please choose the appropriate response for each item:
General discomfort: None / Slight / Moderate / Severe

Fatigue: None / Slight / Moderate / Severe

Drowsiness: None / Slight / Moderate / Severe

Sweating: None / Slight / Moderate / Severe

Difficulty concentrating: None / Slight / Moderate / Severe

Mental depression: None / Slight / Moderate / Severe

Visual flashbacks (visual illusion of movement or false sensations of movement, when NOT in a simulator, car, or aircraft):
None / Slight / Moderate / Severe

Faintness: None / Slight / Moderate / Severe / Aware of breathing: None / Slight / Moderate / Severe

Confusion: None / Slight / Moderate / Severe

Eyestrain: None / Slight/ Moderate/ Severe

Difficulty focusing: None / Slight/ Moderate/ Severe

Blurred vision: None / Slight/ Moderate/ Severe

Headache: None / Slight/ Moderate/ Severe

Fullness of the head: None / Slight/ Moderate/ Severe

Dizziness with eyes open: None / Slight/ Moderate/ Severe

Dizziness with eyes closed: None / Slight/ Moderate/ Severe

Vertigo (Vertigo is experienced as loss of orientation with respect to vertical upright): None / Slight/ Moderate/ Severe

Nausea: None / Slight/ Moderate/ Severe

Stomach awareness (Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea): None / Slight/ Moderate/ Severe

Loss of appetite: None / Slight/ Moderate/ Severe

Increased appetite: None / Slight/ Moderate/ Severe

Desire to move bowels: None / Slight/ Moderate/ Severe

Burping: No / One time/ 2 Times/ 3 Times/ Less than 5 times/ less than 10 times/ A lot

Vomiting: No / One time/ 2 Times/ 3 Times/ Less than 5 times/ less than 10 times/ A lot

Please specify what other symptoms you are experiencing and what their severity is.

APPENDIX C

GAME INFORMATION DOCUMENTS

C.1 Navigation experiment

Procedure

1. Read and sign IRB consent form (5 minutes)
2. Read instructions (5 minutes)
3. Training session (20 minutes)
4. Experiment (35 minutes)
5. Fill out questionnaire (5 minutes)

Police officer

You are taking the role of a police officer. You were sent into an unfamiliar part of your city. Your goal is to follow directions from a dispatcher using radio communication. The dispatcher has a map of the city, but because of construction, some parts of the map could be out of date. You should provide the dispatcher with landmarks, such as description of buildings and billboards. Your goal is to go through all destination points as fast as possible, but it's not allowed to go over 30 mph and you must stop at every stop sign.

You must not go past the construction barrels that are placed across some streets.

The car has a built in engine failure detection system. This system has the ability to fix the engine if it has information about how to do the fix. The dispatcher can send this information to your car. When you see a message "Check engine" on the screen, your car is about to break down. You should inform the dispatcher about this message, so he can send required information to your car.

Your radio system also detects the loss of connection strength of the data link between the car and the dispatcher office. When you see a message "Check link", you also must inform the dispatcher. If you fail to do so the car will stop until the data link is established again.

You will see "estimated time to failure" progress bar under the warning messages. The car will break or stop once the progress bar is at 100%.

Thank you very much for your participation.

Dispatcher

You are taking the role of a dispatcher in the police headquarters. There is a police officer who needs your assistance. Your goal is to navigate this officer from his current

location to the points marked on your map. There are three points marked 1, 2, and 3 respectively. You should communicate with the officer to discover where he is on the map and after that you provide directions to point 1. Once the officer reached point 1 you provide him with the directions to point 2. And from point 2, the officer should go to point 3.

There was recent construction in the city and some parts of the map could be out of date: some roads could be closed, and some roads could be opened. You should work with the officer to detect what parts of the map are out of date. There are red rectangles on the map that denote construction barrels and the officer is not allowed to go past them. Try all the streets leading to the destination one by one. Eventually one of them will be free of construction.

If the officer informs you that there is a "Check engine" sign, you should ask what the speed the vehicle is. This will provide enough information for the system to fix the car.

If the officer informs you that there is a "Check link" sign, you should ask how far the car is from the next road intersection (a block away, half a block away, third of a block away). This will provide enough information for the system to fix the data link.

Thank you very much for your participation.

C.2 Twenty questions experiment

The following text was given to all the subjects prior to the twenty question experiment, described in Chapter 4.

Team scoring

You and your partner have a goal to finish as many games of *Twenty Questions* (described below) as possible during the experiment, while completing all the *Last Letter* games (described below). Games will happen in parallel. There will be a limited time for each game. You will receive a point for each completed game and naming task. A point will be taken from you for every incomplete game or word naming task. If you finish game after the time ran out you will receive half a point. Depending on your performance you will receive a prize at the end of the experiment. You will receive $5 bonus if you will perform well.

The game of Twenty Questions

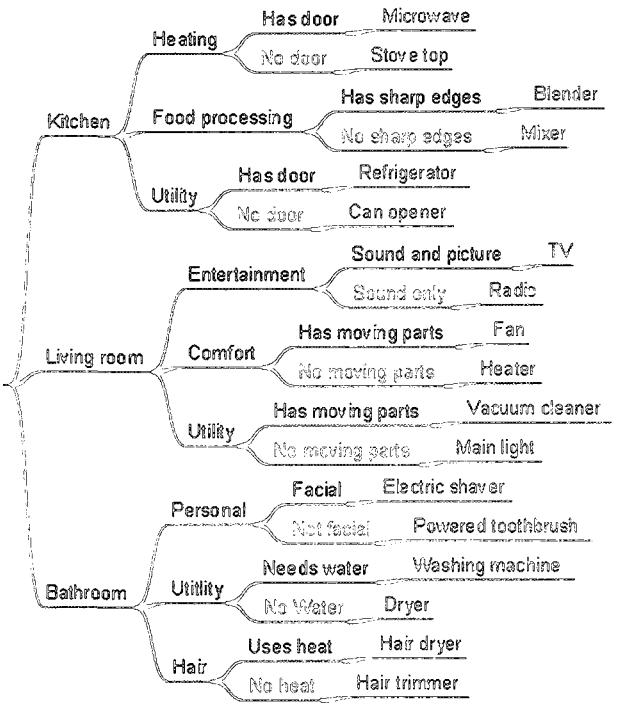
You are going to a play a variation of a game called *Twenty Questions*. Two people play this game. One person is the Answerer and the other is the Questioner. The Answerer is given a word or a phrase, and the goal of the Questioner is to discover that word or phrase in the shortest period of time (the smallest number of questions). The object that the word or phrase represents is always a home appliance. Figure bellow shows all appliances that will be used in the game as well as possible classification of them.

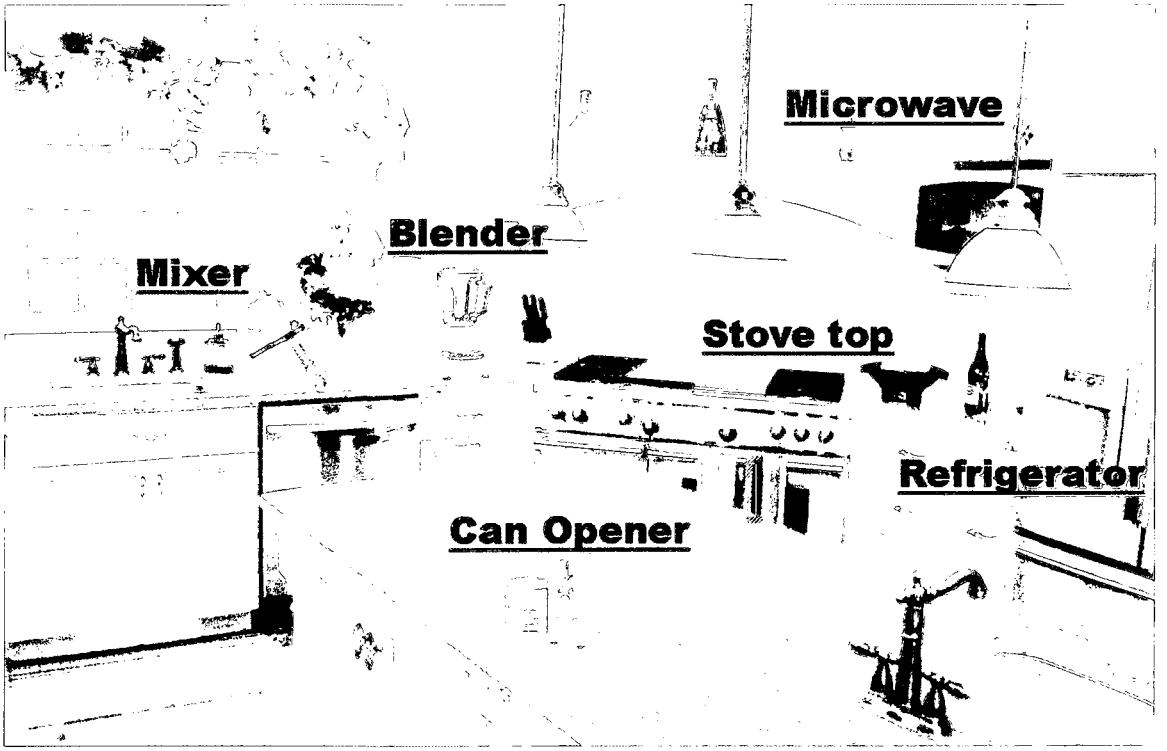
The Questioner can only ask questions that can be answered with yes or no. The goal of the Answerer is to help the Questioner, but the Answerer can only say: *yes, no* or *cannot say* (meaning that any answer would be ambiguous, or is simply not known by the Answerer). For your team to receive a point, the Questioner has to correctly identify the word that the Answerer was given at the beginning of the game. The Questioner has only one chance to name the appliance, so make sure you ask all the relevant questions. There should be no guessing.

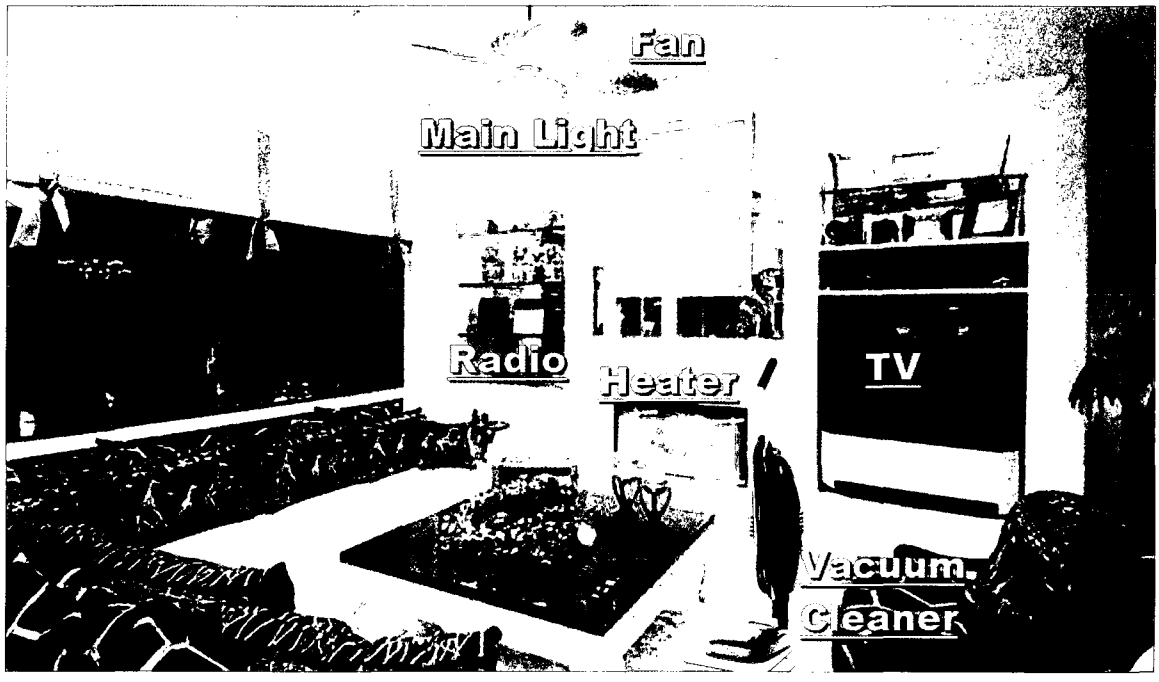
In this experiment both participants will be playing two games in parallel, performing a different role in each game. The person who is not driving starts asking

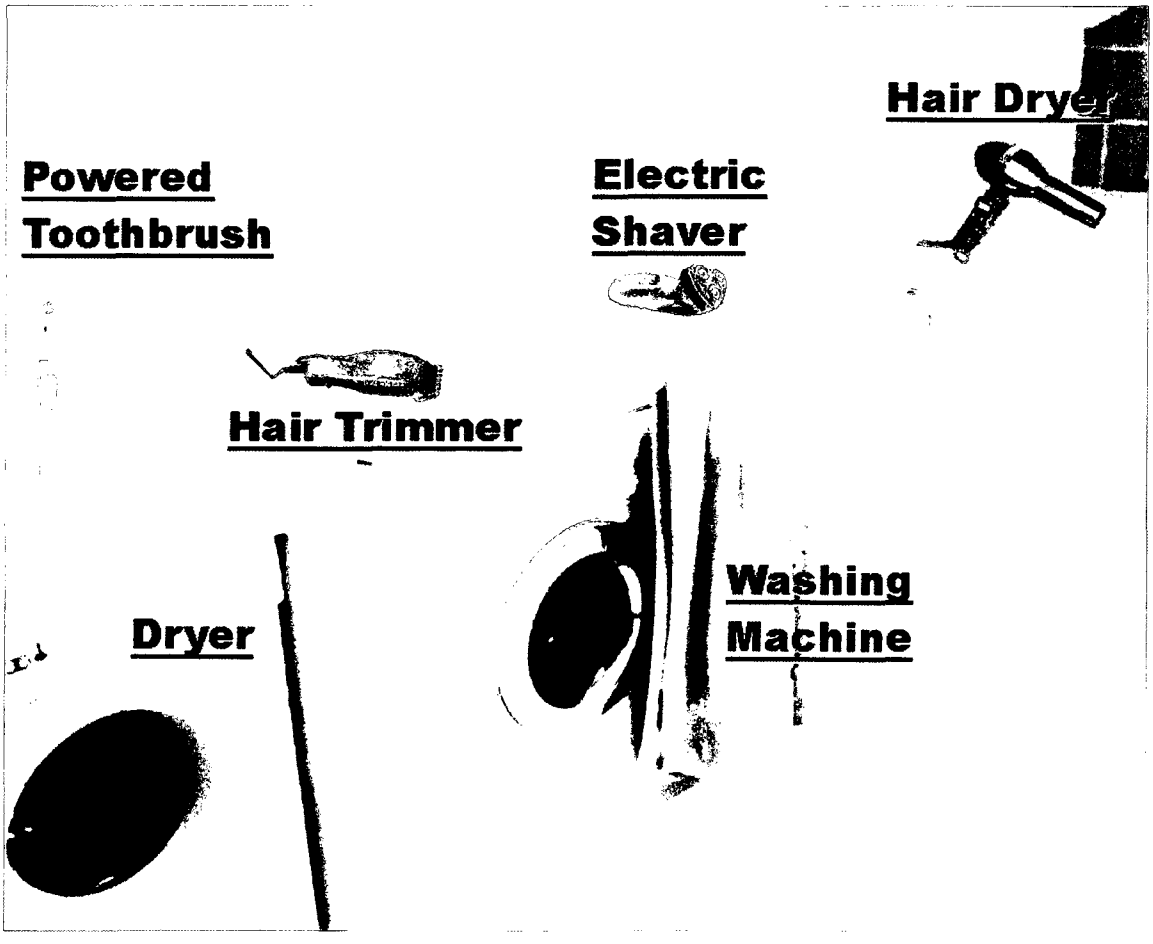
questions first when starting a new game. Here is an example game of two parallel Twenty Questions games between you and your partner. In game one (g1) you are the Questioner (Q) and in game two (g2) you are the Answerer (A). As the Answerer you are given object "Main Light." Your partner is given object "Blender":

You (Q g1):	Is it in the bathroom?
Partner (A g1):	*No*
Partner (Q g2):	Is it in the kitchen?
You (A g2):	*Yes*
You (Q g1):	Is it in the living room?
Partner (A g1):	*Yes*
Partner (Q g2):	Is it used for heating?
You (A g2):	*No*
You (Q g1):	Is it a utility item?
Partner (A g1):	*Yes*
Partner (Q g2):	Is it used for food processing?
You (A g2):	*Yes*
You (Q g1):	Does it have moving parts
Partner (A g1):	*No*
Partner (Q g2):	Does it have sharp edges?
You (A g2):	*Yes*
You (Q g1):	Is it a Main Light
Partner (A g1):	*Yes*
Partner (Q g2):	Is it a blender?
You (A g2):	*Yes*


Kitchen
Heating
Has door
Microwave
No door
Stove top
Food processing
Has sharp edges
Blender
No sharp edges
Mixer
Utility
Has door
Refrigerator
No door
Can opener
Living room
Entertainment
Sound and picture
TV
Sound only
Radio
Comfort
Has moving parts
Fan
No moving parts
Heater
Utility
Has moving parts
Vacuum cleaner
No moving parts
Main light
Bathroom
Personal
Facial
Electric shaver
Not facial
Powered toothbrush
Utility
Needs water
Washing machine
No Water
Dryer
Hair
Uses heat
Hair dryer
No heat
Hair trimmer


Microwave
Blender
Mixer
Stove top
Refrigerator
Can Opener


Fan
Main Light
Radio
Heater
TV
Vacuum
Cleaner


Hair Dryer
Powered
Toothbrush
Electric
Shaver
Hair Trimmer
Dryer
Washing
Machine

Examples of good questions

Is it usually found in a kitchen?	Is it used for entertainment?
Is it usually found in a living room?	Is it used for comfort?
Is it usually found in a bathroom?	Does it show pictures?
Is it used for heating food?	Does it play sounds?
Is it used for food processing?	Does it have moving parts?
Does it have a door?	Does it touch face when used?
Is it used directly on food?	Does it require water to work?

Last Letter game

You will be given a task of naming a word that starts with a given letter and is 4 or 5 letters long. For example, when you see a message that says "S" with a progress bar, you need to interrupt the ongoing Twenty Questions game and initiate the Last Letter game. You can do this by saying:

Name a 4 or 5 letter word that starts with S.

Your partner might say *Soda*. Now you have to name a 4 or 5 letter word that starts with the last letter of the word created by your partner. In this example, you may use a 4 or 5 letter word that starts with A (*arch* or *apple*, for instance). Now it's your partner's turn to name a word that starts with the last letter of your word. You repeat this 3 times. Overall, each of the participants names three words. Once you have named three words you can continue with the Twenty Questions game.

You have a time limit to complete a given Last Letter game. The message that informs you about this task will have a progress bar next to it. You must name three words before the progress bar reaches 100%. You cannot repeat words that you have already used. If it takes you too long to name a word which has a given number of letters you can name a word with any number of letters. If you use longer/shorter word or did not finish the game in time you will lose half a point. If you do not finish the game at all you will not get any points for it.

Playing games

You can play games when you see words shown on the screen. If there are no words shown, it means that you should stay silent. Once you see that words disappeared from the screen you should wrap up the current conversation and wait in silence until

words appear on the screen again. If words are visible on the screen and you already finished the game you may talk to each other or stay silent.

You always want to finish Last Letter game, but you stop playing twenty questions game as soon as the words disappear from your screen, even if you did not finish the game yet.

Driving (for driver only)

When driving, your goal is to follow the leading vehicle at a safe distance. You can ignore all speed limit signs. The vehicle in front of you will keep a constant speed of 55mph. You should make an attempt to stay with the leading vehicle. Please do not go past the leading vehicle, you should follow it. When the leading vehicle stops, you should stop as well.

Good-luck.