Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications
(AutomotiveUI 2010), November 11-12, 2010, Pittsburgh, Pennsylvania, USA

Spoken Tasks for Human-Human Experiments: Towards In-Car Speech User Interfaces for Multi-Threaded Dialogue

Andrew L. Kun[1], Alexander Shyrokov[1], Peter A. Heeman[2]

[1] University of New Hampshire
Electrical and Computer Engineering Department
Kingsbury Hall
Durham, NH 03824, USA
603.862.1357
{andrew.kun, shirokov}@unh.edu

[2] Center for Spoken Language Understanding
Oregon Health & Science University
Beaverton OR 97006, USA
503.748.1381
heemanp@ohsu.edu

ABSTRACT
We report on the design of spoken tasks for a study that explored how people manage spoken multi-threaded dialogues while one of the conversants is operating a simulated vehicle. Based on a series of preliminary studies we propose a set of considerations that researchers should take into account when designing such tasks. Using these considerations, we discuss two spoken tasks, the *parallel twenty questions game* and the *last letter game*, and discuss the successful utilization of these tasks in a study exploring human-human dialogue behavior.

Categories and Subject Descriptors
H5.2. User Interfaces: Evaluation/methodology.

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Speech user interfaces, multi-threaded dialogue, spoken task, driving simulator.

1. INTRODUCTION
When people engage in a manual-visual task, such as driving, they still want to interact with a computer to accomplish various tasks. In the case of driving, these tasks include getting navigation information, selecting music, or reading email or text messages. Interacting with a graphical user interface to accomplish these tasks can be dangerous, as it requires taking your hands and eyes away from the manual-visual task. Speech interaction with a computer does not require the use of the hands and eyes, with the possible exception of operating a push-to-talk button. Thus, speech is a viable human-computer interaction mode when the human is also engaged in a manual-visual task.

However, using speech as an interaction mode while engaged in a manual-visual task needs to be done with care. The research literature provides ample evidence that this is true for the manual-visual task of driving. A number of researchers found that conversing on a mobile phone degrades driving performance, even when the phone is used in hands-free mode [1]. In our own work we found evidence that certain characteristics of a speech user interface (e.g. low recognition rate [2]), and of human-human spoken dialogues (e.g. having to switch from one task to another [3]) can negatively influence driving performance. Speech interfaces have been used for performing a single task at a time, where the user finishes with one task before moving on to the next. However, real-time tasks might require the user's interactions on different tasks to overlap in time. For instance, a police officer might need to be alerted to a nearby accident while accessing a database during a traffic stop; or a driver might need driving instructions while reviewing appointments in a calendar. We refer to the speech interaction about each individual task as a dialogue thread and say that together they constitute a multi-threaded dialogue. The dialogue threads can overlap with each other in time. In this paper we focus on our work in laying the groundwork for in-car speech user interfaces capable of carrying out multi-threaded spoken dialogues with drivers.

Our long term goal is to build a spoken dialogue system that allows the user to complete task-oriented spoken dialogues without negatively impacting performance on the driving task. To build such as system, we need to know what types of dialogue behaviors the system should engage in. We propose that we should identify these behaviors by observing the way humans manage dialogues between drivers and one or more remote or co-present conversants. The behaviors can be characterized by different utterance types, pauses and speaking rates, as well as higher-level dialogue reasoning. We can also expect to find patterns in how conversants alter their speech as the driving difficulty changes. The hypothesis that we can use behaviors observed in human-human dialogues between a driver and a co-present conversant (that is a passenger) is supported by evidence that the presence of a passenger, and thus very likely conversing with a passenger, reduces the probability of an accident [4]. This is the basis for our broader hypothesis that human-human dialogues can serve as inspiration for designing appropriate behaviors for the computer to follow in human-computer spoken dialogues in cars.

2. BACKGROUND
In our previous work, we studied how people manage multi-threaded dialogues. As part of that work, we explored different verbal tasks to use. In our first experiments, we had a subject interact with an actual spoken dialogue system [5]. Due to the complexity involved in building a functional system, we had

AutomotiveUI'10, November 11-12, 2010, Pittsburgh, Pennsylvania
ACM 978-1-4503-0437-5

subjects complete simple tasks with the computer, including addition, circular rotation of number sequences, discovery of short letter sequences, and category-matching word detection. However, subjects did not find these tasks engaging, and the resulting dialogues did not seem to capture the complexity of behaviors we expected to see in more realistic tasks. In some of the pilot experiments, we tried to motivate subjects by telling them they were playing a game and their goal was to solve as many tasks as possible; however, this did not seem to help.

To make the tasks more engaging and realistic, we turned our attention to human-human dialogues. We gave conversants an ongoing task in which they had to work together to form a poker hand [6]. Each conversant had three cards, and they took turns drawing and discarding a fourth card. Conversants could not see each other, nor could they see the cards in each other's hands. They communicated via headsets and used speech to share what cards they have and what poker hand to try for. Periodically, one of the conversants was prompted to solve a real-time task, that of determining whether the other conversant has a certain picture displayed on her screen. The urgency of the real time task was an experimental variable: conversants were given either 10, 25, or 40 seconds to complete it. To make the task engaging, conversants received points for each completed poker hand and each picture task. We found that this setup elicited both rich collaboration for the card game [7] and interesting task management. The problem is that this setup, with the ongoing task having a minor manual-visual component, is not representative of the types of tasks that are of interest to us, in which the ongoing task is exclusively verbal and where the user is also engaged in the separate manual-visual task of driving.

For our next study we used a navigation problem as the ongoing task [8], inspired by the Map Task experiments [9]. One conversant (the driver) operated a simulated vehicle, and a second (the dispatcher) helped the driver navigate city streets. The conversants could not see each other and communicated via headsets. Unknown to the dispatcher, some of the city streets were blocked by construction barrels and so the driver was unable to follow some of the dispatcher's instructions. The conversants thus collaborated to find an alternate route. Periodically, the driver was prompted to initiate a short real-time task with the dispatcher. As in the poker-playing task, the prompt included information about the urgency of the real-time task. Although this setup elicited rich task management behavior, participants do not seem to build up discourse context as they converse. This is because the verbal component of the navigation task is more like a series of separate small real-time tasks.

3. CONSIDERATIONS FOR SPOKEN TASKS

3.1 Proposed Requirements

Based on our experiences with the studies described in section 2, we propose the following requirements for verbal tasks to be used in human-human experiments exploring in-car user interfaces:

• *Tasks are engaging.* Real tasks that users want to perform will undoubtedly be engaging for them. Such engaging verbal tasks have the potential to divert the user's attention from a manual-visual task [10], [11].

• *Tasks are relatively complex, and require both participants to participate.* Tasks that can be accomplished with one of the participants speaking little or not at all would provide little data to evaluate different dialogue behaviors in human-human spoken interaction.

• *Tasks allow scoring participant performance.* This will allow us to quantitatively evaluate participant performance under different driving task difficulty levels, and when testing the impact of different dialogue behaviors.

• *Tasks have identifiable discourse structure.* This will allow us to easily analyze how the conventions interact with discourse structure. In particular, tasks should give rise to adjacency pairs, such as question-answer pairs, as these are common in human-machine spoken interaction; therefore learning more about this particular type of interaction will be valuable.

3.2 On Interference with the Driving Task

In addition to the requirements above, researchers designing verbal tasks for human-human experiments need to realize how these tasks may interfere with the manual-visual task of driving. The four-dimensional multiple resource model proposed by Wickens [11] provides insight into how people utilize available mental resources when performing multiple tasks in parallel. This model is therefore useful when considering how a spoken task might interfere with the primary task in the vehicle, which is driving. In Wickens' model each of the four dimensions has two discrete levels. Two tasks that utilize resources pertaining to the same level of a given dimension will interfere with each other more than two tasks that require resources at different levels and/or dimensions of the model.

The dimensions in Wickens' model are: processing stages perceptual modalities, visual channels, and processing codes. The perceptual modality used for driving is primarily the visual modality, while for verbal tasks it is the auditory modality. Similarly, the resources used in processing the visual signals related to driving, and the auditory signals used in verbal tasks are different. The differences in the resources used along these two dimensions suggest that interference between the driving and verbal tasks may not be significant. Wickens also proposes spatial and verbal codes. Tracking and steering are spatial tasks, while speaking is of course a verbal task, again indicating that interference between driving and verbal tasks may not be significant. However, navigation can be accomplished using spoken directions, but it might utilize spatial resources, something we did not account for in our navigation experiment [8].

Overall, Wickens' model indicates that verbal tasks should not interfere with driving significantly. However, studies such as Strayer and Johnston's [10] clearly show that engaging verbal tasks have the potential to divert the user's attention from the driving task. This effect presents a challenge to Wickens' multiple resource model [11]. Hence, as a fifth consideration, experimenters need to consider the extent of interference between the driving task and the verbal tasks.

4. CONSIDERATIONS APPLIED: THE MULTI-THREADED DIALOGUES STUDY

We applied the considerations from Section 3 in our study on multi-threaded dialogues [3]. As shown in Figure 1, in our experiment, pairs of subjects were engaged in two spoken tasks and one of the subjects (the driver) also operated a simulated vehicle. One spoken task was the ongoing task and it was

Figure 1 Driver and dispatcher.

periodically interrupted by another spoken task. The interruptions forced subjects to switch between different dialogue threads. We tracked the pupil diameter-based physiological measures and the driving performance measures of the driver's cognitive load.

4.1 Equipment

The driver operated a high-fidelity driving simulator (DriveSafety DS-600c) with a 180º field of view, realistic sounds and vibrations, a full-width car cab and a tilting motion platform that simulates acceleration and braking effects. We recorded pupil diameter data using a Seeing Machines faceLab 4.6 stereoscopic eye tracker mounted on the dashboard.

The two subjects communicated using headphones and microphones and we recorded their dialogues. Their communication was supervised by the experimenter to enforce time limits on tasks.

4.2 Driving (primary) and spoken tasks

The primary task of the drivers was to follow a vehicle while driving responsibly. They drove on two-lane, 7.2 m wide roads in daylight. The lead vehicle traveled at 89 km/h (55mph) and it was positioned 20 meters in front of the subject. There was also a vehicle 20 meters behind the subject's car. No other traffic was present on the road. The roads consisted of six straight and six curvy road segments with straight and curvy segments alternating.

The difficulty of the driving task is influenced by factors such as road type, traffic volume, visibility, visual demand of route following, etc. In this research, we only manipulated visual demand of route following. This can be accomplished by controlling the radius of curves [12]. Sharper curves correspond to increased visual demand. As mentioned above, we created routes with a mixture of straight segments and curved segments. Each segment was long enough for participants to complete at least one real-time task, allowing us to evaluate task switching behavior for the given visual demand level.

Our ongoing spoken task was a parallel version of twenty questions (TQ). In TQ, the questioner tries to guess a word the answerer has in mind. The questioner can only ask yes/no questions, until she is ready to guess the word. In our version, the two conversants switch roles after each question-answer pair is completed. The commonality between their roles allows us to contrast their behaviors. Words to guess were limited to a list of household items (hair dryer, TV, etc.). In order to minimize learning effects and to make the dialogue pace more realistic, we trained participants to use a question tree to guess the objects. For example, each item was in one of three rooms, thus the first fact to be established was the room where the object was located. The words to be guessed were presented to the subjects visually. We showed words to the driver just above the dashboard which minimizes interference with driving. We told subjects that there was a time limit to finish a game, and we enforced this time limit.

Our interrupting task was a version of the last letter word game (LL). In our version of this game a participant utters a word that starts with the last vowel or consonant of the word uttered by the other participant. For example, the first participant might say, "page" and the second says "earn" or "gear." Subjects had 30 seconds to name three words each. After completing this task they resumed the TQ game. Subjects played one TQ game and were interrupted by one LL game per curvy and straight road segment.

4.3 Results

The experiment was completed by 16 pairs of participants (32 participants) 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. We recorded ongoing and interrupting task dialogues for 16 pairs x 2 subjects/pair x 12 games/subject = 384 games. This translates into 9.3 hours of speech interactions with synchronized simulator and eye tracker data. The driving and eye-tracker data were collected over 800 km traveled. During the experiments 25% of the time the subjects were saying something to each other.

The tasks described in section 4.2 followed the proposed requirements presented in section 3.1.

Engaging. The interactions were engaging, as demonstrated by the example from the corpus shown in Table 1. In this example the two conversants successfully complete both the parallel TQ games and the interrupting LL game. The example also illustrates that sometimes subjects negotiated that the dispatcher's first question will not only serve to determine the location (room) of the item the driver has in mind, but will also to indicate to the driver the location of the item the dispatcher has in mind. Thus, in U1 the dispatcher's "Is it in the kitchen?" is equivalent to "The item I have in mind is in the kitchen, is yours also in the kitchen?" In our study three of the 16 subject pairs used this approach. The negotiation happened during the training period. Such negotiations are an indication that participants took the games seriously and engaged in them. These negotiations are also an example of unexpected ways in which subjects can affect experiments and researchers should be on the lookout for them.

Both participants speak. By design the tasks required both participants to take turns speaking.

Scoring. The verbal tasks allowed scoring participant performance in multiple ways. The simplest one was evaluating if participants successfully completed individual TQ and LL games. For example, a total of 296 games (77%) of the 384 twenty questions games resulted in a successful completion. This indicates that the difficulty of the ongoing task 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.

Another way to score performance was to evaluate dialogue timing characteristics, such as the length of pauses preceding a participant asking a question in the TQ game. We compared these pause lengths for curvy and straight road segments and found no statistically significant differences, indicating that driving task difficulty did not influence this aspect of the TQ game.

Discourse structure. As shown in Table 1, both games had identifiable discourse structure in terms of question-answer pairs. This allowed us to analyze how interruptions affected the conversant when they were at different points in the question-answer pairs.

Code	Speaker	Utterance	Task
U1	Disp.	Is it in the kitchen?	TQ
U2	Driver	No.	TQ
U3	Driver	Does it have sharp edges?	TQ
U4	Disp.	No.	TQ
U5	Disp.	Is it in the bathroom?	TQ
U6	Driver	No.	TQ
U7	Driver	Does it produce heat?	TQ
U8	Disp.	No.	TQ
U9	Disp.	Is it on the ceiling?	TQ
U10	Driver	No.	TQ
U11	Disp.	Letter, word beginning with B	LL
U12	Driver	Ball.	LL
U13	Disp.	Like.	LL
U14	Driver	Kite.	LL
U15	Disp.	Time.	LL
U16	Driver	Move.	LL
U17	Disp.	Voice.	LL
U18	Driver	Okay.	Switch
U19	Disp.	Your turn to ask.	Switch
U20	Driver	Does it have a door?	TQ
U21	Disp.	Yes.	TQ
…			TQ
U28	Driver	Is it the refrigerator?	TQ
U29	Disp.	Yes.	TQ
U30	Disp.	Is it the TV?	TQ
U31	Driver	Yes	TQ

Table 1 Example ongoing and interrupting tasks.

Overall, our tasks elicited a variety of dialogue behaviors between the conversants. We have been able to use our qualitative and quantitative evaluations of these behaviors to provide suggestions for the development of in-car speech user interfaces [3], as well as to apply pupil diameter-based physiological measures to explore how different dialogues might affect cognitive load [13].

5. CONCLUSION

We successfully created spoken tasks that allowed us to explore human-human multi-threaded dialogues in cars. However, the behaviors we observed were primarily related to the timing characteristics of the dialogue. We expect that in real-world dialogues in vehicles, human conversants employ a much wider array of behaviors in order to adapt to the ever-changing requirements of the road. We believe that the reason these behaviors were not observed in our study is that the options at the conversants' disposal were limited by the relative simplicity of the tasks. After all, there is only so much one can do to change the flow of the last letter game! As indicated by Drews et al. [14], many other studies have a similar limitation. This observation points to the need to introduce more complex dialogue tasks in future studies with the expectation that they will allow a wide range of different dialogue behaviors to be exhibited.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Horrey, W. J., and Wickens, C. D. 2006. Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques. *Human Factors*, 48, 1, 196-205.

[2] Kun, A. L., Paek, T., and Medenica, Z. 2007. "The Effect of Speech Interface Accuracy on Driving Performance," In *Proceedings of Interspeech*. Antwerp, Belgium.

[3] Shyrokov, A. 2010. *Human-human multi-threaded spoken dialogs in the presence of driving*. Doctoral Dissertation. University of New Hampshire.

[4] Rueda-Domingo, T., Lardelli-Claret, P., Luna-del-Castillo, J. de, D., Jimenez-Moleon, J. J., Garcia-Martin, M., and Bueno-Cavanillas, A. 2004. The influence of passengers on the risk of the driver causing a car collision in Spain: Analysis of collisions from 1990 to 1999. *Accident Analysis & Prevention*, 36, 481–489.

[5] Shyrokov, A. 2006. *Setting up experiments to test a multi-threaded speech user interface*. Technical Report ECE.P54.2006.1, University of New Hampshire.

[6] Heeman, P. A., Yang, F., Kun, A. L., and Shyrokov, A. 2005. Conventions in human-human multi-threaded dialogues: A preliminary study. In *Proceedings of the International Conference on Intelligent User Interfaces*. San Diego, CA.

[7] Toh, S. L., Yang, F. and Heeman, P. A. 2006. An annotation scheme for agreement analysis. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP-06)*. Pittsburgh PA.

[8] Shyrokov, A., Kun, A., and Heeman, P. 2007. Experimental modeling of human-human multi-threaded dialogues in the presence of a manual-visual task. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*. Antwerp Belgium.

[9] Anderson, A. H., Bader, M., Bard, E. C., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. 1991. The HCRC map task corpus. *Lang. and Speech*, 34, 4, 351–366.

[10] Strayer, D. L., and Johnston, W. A. 2001. Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone. *Psych. Sci.*, 12, 462–466.

[11] Wickens, C. D. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3, 2, 159–177.

[12] Tsimhoni, O., and Green, P. 1999. Visual demand of driving curves determined by visual occlusion. In *Proceedings of the Vision in Vehicles 8 Conference*. Boston, MA.

[13] Palinko, O. Kun, A.L., Shyrokov, A., Heeman, P. 2010. Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator. In *Proceedings of the Eye Tracking Research and Applications Conference*. Austin, TX.

[14] Drews, F. A., Pasupathi, M. and Strayer, D. L. 2008. Passenger and Cell Phone Conversations in Simulated Driving. *Journal of Exper. Psych.: Applied*. 14, 4, 392–400.