The Effect of Cognitive Load on a Statistical Dialogue System

M. Gašić*, P. Tsiakoulis*, M. Henderson*, B. Thomson*, K. Yu*, E. Tzirkel and S. Young***
*Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK
{mg436, pt344, mh521, brmt2, ky219, sjy}@eng.cam.ac.uk
**General Motors Advanced Technical Centre, Israel
eli.tzirkel@gm.com

Abstract

In recent years statistical dialogue systems have gained significant attention due to their potential to be more robust to speech recognition errors. However, these systems must also be robust to changes in user behaviour caused by cognitive loading. In this paper, a statistical dialogue system providing restaurant information is evaluated in a set-up where the subjects used a driving simulator whilst talking to the system. The influences of cognitive loading were investigated and some clear differences in behaviour were discovered. In particular, it was found that users chose to respond to different system questions and use different speaking styles, which indicate the need for an incremental dialogue approach.

1 Introduction

A spoken dialogue system enables a user to obtain information while using their hands to perform some other task, which in many cases is the user's primary task. A typical example is an in-car spoken dialogue system where the spoken interaction is secondary to the main task of driving the car (Weng et al., 2004). This domain is particularly challenging since it involves dealing with the errors caused by the varying noise levels and changes in user behaviour caused by the cognitive load.

A statistical approach to dialogue modelling has been proposed as a means of automatically optimising dialogue policies. In particular, the partially observable Markov decision process (POMDP) model for dialogue provides a representation of varying levels of uncertainty of the user input, yielding more robust dialogue policies (Williams and Young, 2007; Thomson and Young, 2010; Young et al., 2010).

Another thread of research deals with speech interfaces for in-car applications, see (Baron and Green, 2006) for a review. Past research has investigated the extent to which speaking is cognitively less demanding than typing (Gartner et al., 2001; Tsimhoni et al., 2004; Kun et al., 2007). In addition, considerable research has examined how driving safety is influenced by a dialogue system (Lai et al., 2001; Lee et al., 2001; Nielsen et al., 2008). However, to the best of our knowledge, little work has been done to investigate the effect of the cognitive load when interacting with a real conversational spoken dialogue system. The work presented in (Mishra et al., 2004) suggests that the user speech is more disfluent when the user is performing another task. However, this work is based on a Wizard of Oz framework, where a human provides the system's responses. Also, a push-to-talk button was used for every utterance which will have affected the natural flow of the dialogue. It is important to know if the change of cognitive load has an effect on the speaking style and whether the system can alter its behaviour to accommodate for this.

In this paper we try to answer these questions by examining dialogues where users drove a car simulator and talked to an open-microphone fully automated spoken dialogue system at the same time.

The rest of the paper is organised as follows. Section 2 provides an overview of the dialogue system used and section 3 describes the evaluation set-up. The analysis of the results is given in Section 4. Section 5 concludes the paper.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 74–78,
Seoul, South Korea, 5-6 July 2012.

Table 1: Example dialogue task

You are looking for a cheap restaurant and it should be in the east part of town. Make sure you get the address of the venue.

2 System overview

The user speaks to the system, and the acoustic signal is converted by the speech recogniser into a set of sentence hypotheses, which represents a probability distribution over all possible things that the user might have said. The sentence hypotheses are converted into an N-best list of dialogue acts by a semantic decoder. Since the dialogue state cannot be directly observed it maintains a probability distribution over all states, which is called the belief state. The belief state is updated at every user turn using Bayesian inference treating the input dialogue acts as evidence. Based on belief state, the optimal system act is selected using a policy and which is trained automatically using reinforcement learning. The abstract system dialogue act is converted to an appropriate utterance by a natural language generator and then converted to speech by an HMM-based speech synthesiser. To enable in-car speech interaction via mobile phone, a VoIP interface is implemented. The domain is Cambridge restaurant information with a database of about 150 venues and 7 slots that users can query.

3 Evaluation set-up

Our goal is to understand system performance when driving. However, due to the safety restrictions, performance was tested using a driving simulator. The following sections explain the set-up.

3.1 Car simulator

The car simulator used in the evaluation was the same as in (Davies and Robinson, 2011). It consists of a seat, a steering wheel and pedals, which give a realistic cab-like environment for the participants. There is also a projection screen which largely fills the visual field of the driver. The simulation software is a modified version of Rockstar Games' Grand Theft Auto: San Andreas, with over 500 km of roads. For the purpose of the evaluation, the subjects were asked to drive on the main motorway, to keep the lane and not to drive over 70mph.

3.2 Subjects

For the study 28 subjects were recruited, 22 where native speakers. Each subject had to complete three scenarios: (1) to drive the car simulator for 10 minutes, (2) to talk to the system for 7 dialogues and (3) to talk to the system for 7 dialogues while driving. The scenarios were in counter-balanced order.

While they were driving, the speed and the road position were recorded. If the scenario involved talking to the system, the instructor read out the dialogue task (see an example in Table 1) and dialled the phone number. In addition, the subject had the dialogue task displayed on a small screen next to the driving wheel. The subject talked to the system using loud speaker mode on the mobile phone.

4 Results

To examine the influence of cognitive load, the following examinations were performed. First, we investigate if the subjects felt any change in the cognitive load (Section 4.1). Then, in Section 4.2, we examine how the driving was influenced by the subjects talking to the system. Finally, we investigate how successfully the subjects were able to complete the dialogue tasks while driving (Section 4.3). This is followed with an examination of the conversational patterns that occurred when the subjects were driving whilst talking to the system (Section 4.4).

4.1 Cognitive load

After each scenario the subjects were asked to answer five questions based on the NASA-TLX self-reporting scheme for workload measurement. They answered by providing a rating from 1 (very easy) to 5 (very hard). The averaged results are given in Table 2. We performed a Kruskal test, followed by pairwise comparisons for every scenario for each answer and all differences are statistically significant ($p < 0.03$) apart from the differences in the frustration, the stress and the pace between talking and talking and driving. This means that they were clearly able to feel the change in cognitive load.

Table 2: Subjective evaluation of the cognitive load

Driving	Talking	Talking&Driving
How mentally demanding was the scenario?		
1.61	2.21	2.89
How hurried was the pace of the scenario?		
1.21	1.71	1.89
How hard did you have to work?		
1.5	2.32	2.96
How frustrated did you feel during the task?		
1.29	2.61	2.61
How stressed did you feel during the task?		
1.29	2.0	2.32

Table 3: Analysis of driving speed to determine which measures are larger for Talking&Driving than Driving

Measure	Percentage of users	Confidence interval
Higher speed	8%	$[1\%, 25\%]$
Larger std.dev	77%	$[56\%, 91\%]$
Larger entropy	85%	$[65\%, 95\%]$

4.2 Driving performance

For 26 subjects we recorded position on the road and the speed. Since these measurements vary significantly across the subjects, for each subject we calculated the average speed, the standard deviation and the entropy and similarly for the average position in the lane. For the speed, we computed how many subjects had a higher average speed when they were talking and driving versus when they were just talking and similarly for the standard deviation and the entropy. The results are given in Table 3. It can be seen that the user's speed is lower when they are driving and talking, however, the increase in the standard deviation and the entropy suggest that their driving is more erratic. No significant differences were observed for the road position.

4.3 Dialogue task completion

Each participant performed 14 dialogues, 7 for each scenario. In total, there were 196 dialogues per scenario. After each dialogue they told the instructor if they thought the dialogue was successful, and this information was used to compute the subjective completion rate. In addition, all dialogues were transcribed and analysed to see if the system provided information the user asked for and hence calculate an objective completion rate. The results are given in Table 4. These differences are not statistically significant due to the small sample size. However, it can be seen that the trend is that the dialogues where the subject was not performing another task at the same time were more successful. Also, it is interesting that the subjective scores are higher than the objective ones. This can be explained by the fact that the dialogue tasks were predefined and the subjects do not always pay sufficient attention to their task descriptions.

Table 4: Subjective and Objective Task completion (196 Dialogues per scenario)

	Talking	Talking&Driving
Subjective	78.6%	74.0%
Objective	68.4%	64.8%

Table 5: Percentage of turns that are in line with the predefined task

	Talking	Talking&Driving
Percentage of turns that follow the task	98.3%	96.79%
Number of turns	1354	1388

4.4 Conversational patterns

Given that the subjects felt the change of cognitive load when they were talking to the system and operating the car simulator at the same time, we were interested to see if there are any changes in the dialogues which might suggest this.

First, we examine how well they follow the given task on a turn-to-turn basis. For example, if the task is to find a cheap restaurant and if at some point in the dialogue the user says *I'd like an expensive restaurant* that turn is not consistent with the task. The results are given in Table 5 and they are statistically significant ($p < 0.01$).

We then examine the number of contradictions on a turn-to-turn basis. For example, if the user says *I'd like a cheap restaurant* and later on they say *I'd like*

Table 6: User obedience to system questions

1. system requests or confirms and requests		
	Samples	Obedience
Talking	392	67.6%
Talking&Driving	390	63.9%
2. system confirms		
	Samples	Obedience
Talking	91	73.6%
Talking&Driving	92	81.5%

an expensive restaurant the latter turn is clearly a contradiction. The percentage of contradicting turns is less than 1% and the difference between the scenarios is not statistically significant. This suggests that while users tend to forget the task they are given when they are driving, they still act rationally despite the increase in the cognitive load.

The next analysis concerns the user obedience, i.e. the extent to which subjects answer the system questions. We grouped the system questions in two classes. The first class represents the questions where the system requests a value for a particular slot, for instance *What part of town are you looking for?* and the questions where the system confirms and requests at the same time, for instance *You are looking for a cheap restaurant. What part of town are you looking for?* The second class correspond to system confirmations, for example *Did you say you are looking for a cheap restaurant?* The percentage of the obedient user turns per class is given in Table 6. Due to the small sample size these results are not statistically significant. Still, it is interesting to see that when driving the subjects appear to be more obedient to the system confirmations than when they are just talking. When the system makes a confirmation, the user can answer with simple yes or no, whereas when the system requests the value of a particular slot, the user needs to think more to provide an answer.

The number of barge-ins, the number of filler words and the average speech intensity vary considerably among the subjects. Therefore, we average these statistics per user and examine the number of users for which the particular measure is greater for the scenario where they talked to the system and drove the simulator at the same time. The results (Table 7) show that the number of barge-ins and the number of fillers is significantly greater for the scenario when they are talking and driving and the intensity on average tend to be greater.

Table 7: Analysis of measures related to the speaking style which values are larger for Talking&Driving than Talking

Measure	% of users	Conf. interval
More barge-ins	87%	[69%, 96%]
More fillers	73%	[54%, 88%]
Higher intensity	67%	[47%, 83%]

5 Conclusion and Future work

There are several important observations arising from this study. Firstly, dialogues with cognitively loaded users tend to be less successful. This suggests that the system should alter its behaviour to match user behaviour and alleviate the cognitive load in order to maintain the level of performance. This necessitates rapid on-line adaptation of dialogue policies.

The second observation is that cognitively loaded users tend to respond to some types of system questions more than others. This indicates that the user model within a POMDP dialogue system should be conditioned on a measure of cognitive load.

Finally, this study has found that users barge-in and use filler words significantly more often when they are cognitively loaded. This suggests the need for a much richer turn-taking model which allows the system to use back-channels and barge-in when the user hesitates. An obvious candidate is the incremental approach (Schlangen and Skantze, 2009; DeVault et al., 2009) which allows the system to process partial user inputs, back-channels, predict short term user input and interrupt the user during hesitations. While incremental dialogue is a growing area of study, it has not so far been examined in the context of dialogue for secondary tasks. We signpost this as an important area for future work.

Acknowledgments

We would like to thank to Peter Robinson and Ian Davies for their help with the experiments.

References

A Baron and P Green. 2006. Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review. Technical Report UMTRI-2006-5.

I Davies and P Robinson. 2011. Emotional investment in naturalistic data collection. In *International Conference on Affective Computing and Intelligent Interaction*.

D DeVault, K Sagae, and DR Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *10th Annual SIGDIAL meeting on Discourse and Dialogue*.

U Gartner, W Konig, and T Wittig. 2001. Evaluation of Manual vs. Speech Input When Using a Driver Information System in Real Traffic. In *International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design*.

A Kun, T Paek, and Ž Medenica. 2007. The effect of speech interface accuracy on driving performance. In *Interspeech*.

J Lai, K Cheng, P Green, and O Tsimhoni. 2001. On the Road and on the Web? Comprehension of synthetic and human speech while driving. In *SIGCHI*.

JD Lee, B Caven, S Haake, and TL Brown. 2001. Speech-based Interaction with In-vehicle Computers: The Effect of Speech-based E-mail on Drivers' Attention to the Roadway. *Human Factors*, 43:631–640.

R Mishra, E Shriberg, S Upson, J Chen, F Weng, S Peters, L Cavedon, J Niekrasz, H Cheng, and H Bratt. 2004. A wizard of Oz framework for collecting spoken human-computer dialogs. In *Interspeech*.

BS Nielsen, B Harsham, B Raj, and C Forlines. 2008. Speech-Based UI Design for the Automobile. *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, pages 237–252.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 710–718.

B Thomson and S Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588.

O Tsimhoni, D Smith, and P Green. 2004. Address Entry While Driving: Speech Recognition Versus a Touch-Screen Keyboard. *Human Factors*, 46:600–610.

F Weng, L Cavedon, B Raghunathan, D Mirkovic, H Cheng, H Schmidt, H Bratt, R Mishra, S Peters, L Zhao, S Upson, E Shriberg, and C Bergmann. 2004. Developing a conversational dialogue system for cognitively overloaded users. In *Proceedings of the International Congress on Intelligent Transportation Systems*.

JD Williams and SJ Young. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):393–422.

SJ Young, M Gašić, S Keizer, F Mairesse, J Schatzmann, B Thomson, and K Yu. 2010. The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.