The effect of cognitive load on disfluencies during in-vehicle spoken dialogue

Anders Lindström¹, Jessica Villing², Staffan Larsson², Alexander Seward³, Nina Åberg⁴, Cecilia Holtelius⁵

¹Mobility Services R&D, TeliaSonera, Sweden
²Department of Linguistics, University of Gothenburg, Sweden
³Veridict AB, Sweden
⁴Volvo Technology AB, Sweden
⁵Volvo Car Corporation AB, Sweden

andres.x.lindstrom@teliasonera.com, jessica@ling.gu.se, a1@ling.gu.se,
alec@veridict.com, nina.aberg@volvo.com, cholteli@volvocars.com

Abstract

In-vehicle spoken dialogue systems are gaining increased interest by the automotive industry. They enable the driver to perform secondary tasks (i.e. tasks not related to driving the vehicle) without having to take her eyes off the road or her hands from the steering wheel. Dialogue systems also enable the driver to speak in a natural way, without having to memorize commands or navigate through a menu structure. It is however crucial to take the cognitive load of the driver into consideration, in order to be able to adapt the dialogue system accordingly. This paper presents a user study where spoken dialogues between drivers and passengers have been analysed to find out how spontaneous speech is affected by driving and carrying out other activities that increase the cognitive load of the user. The results indicate systematic changes in specific disfluency rates as the cognitive load increases.

Index Terms: spoken language dialogue, in-vehicle dialogue systems, spontaneous speech, disfluencies

1. Background

The study reported on in this paper was carried out within the DICO project (www.dicoproject.org). The overall purpose of this project is to demonstrate how state-of-the-art spoken language technology can enable access to communication, entertainment and information services as well as to environment control in vehicles. We intend to demonstrate this primarily by means of working prototypes which promote safety in driving while at the same time delivering ease-of-use in access to commercially viable sets of on-line as well as in-vehicle services. To this end, the project has developed a working prototype of a speech-based and multimodal dialogue system [1], which has previously been tested on real users both in simulator tests and while driving in real traffic. During these user tests some questions have arisen. One question concerns how to deal with the abundance of disfluencies, which are often typical of spontaneous spoken dialogue [2, 3]. Of specific interest is to study if and how these features of spoken interaction are affected by the cognitive demands of driving. The reason for this is that the use of dialogue systems in vehicles raises the problem of making sure that the spoken interaction does not distract the driver from the primary task of driving. Earlier studies have indicated that humans are apt at adapting their interaction, to accommodate the cognitive demands of the combined tasks of driving and interacting through spoken language [4]. Researchers within the fields of vehicle safety and ergonomics have furthermore proposed that in-vehicle spoken dialogue systems should adapt to the workload of the driver [5] and even that the dialogue behaviour should be designed in such a way that a “neutral, small talk-like interaction results” [6].

Several research initiatives have been made to address different aspects of this rather daunting goal. For instance, within the SENECA project [7], special attention was paid to generally provide flexible and context-dependent clarification dialogues. Similarly, the CHAT project [8] attempted to achieve a robust, wide-coverage, and cognitive load-sensitive spoken dialogue interface, addressing issues related to dynamic and attention-demanding environments such as driving. However, the CHAT system was not designed to monitor the driver’s cognitive load. Instead, different methods such as robust interpretation and dynamically pruned content information presentation were employed to decrease cognitive load more generally.

In order to adapt a spoken dialogue system to the driver’s cognitive load, more data is needed on how spontaneous speech is affected by driving and carrying out other activities that increase the cognitive demands on the user. To this end, dialogues between driver and passenger in real traffic were recorded and videotaped under controlled conditions, where the driver’s cognitive load was simultaneously measured by use of an indirect method. In the remainder of this paper, we will investigate some of the features of the human-human in-vehicle spoken interaction that was collected in this way. In particular, we will present results concerning the effect of variation in cognitive load on the frequency and distribution of disfluent speech.

Regarding the effect of increased cognitive load on the frequency of different disfluency types, there are at least three possible hypotheses:

- null hypothesis: all disfluency rates remain the same during low as well as high workload
- hypothesis 1: all disfluencies increase during high cognitive load
- hypothesis 2: some disfluencies increase while others remain the same or decrease

2. Method

The goal of the test setup was to elicit driver–passenger dialogue which would feature a substantial and measurable number of in-
instances of the different types of human speech-communicative strategies and linguistic devices known to be employed under cognitive load and other forms of driving-induced stress. One specific challenge was therefore how to make driver and passenger engage in natural dialogue and conversation of sufficient intensity that any additional distractions or increase in the cognitive load, due to driving or the surrounding traffic situation, would immediately compel the subjects to adapt their spoken language in ways which would be detectable from subsequent transcription of the conversation.

2.1. Subjects, tasks and test environment

Eight subjects (two female and six male) between the ages of 25 and 36 were recruited internally with one of the partners (Volvo Technology), and were divided into driver-passenger pairs. The subjects had no previous experience from using speech technology or dialogue systems. To meet the requirements mentioned above, the subjects were given two separate tasks, one navigation task and one memory task. In the navigation task the passenger simply had to instruct the driver where to drive. In the memory task, the driver and passenger were to interview each other regarding personal background and interests during the drive, after which their individual ability to recall this information was scored using a fill-out form. Subjects were informed that their joint score would be the basis for a competition, to further encourage interaction, collaboration and thereby conversation. All tests were performed under real and challenging conditions, in relatively dense city traffic in central Gothenburg.

A previously unknown driving route was given to the passenger at the start, together with the interview sheet. The passenger was told only to give verbal driving instructions, spanning no more than one intersection ahead. The driver was told to focus on the main tasks and on driving for safety reasons, but was told also to perform the best he or she could in a so-called Tactile Detection Task (TDT), requiring the driver to press a button at irregular intervals. Each team was free to manage and solve the interview task in any way they saw fit, except from note-taking or use any other memory aids. Within the teams, each subject acted both driver and passenger, since the subjects were instructed to switch roles halfway into the test, which lasted for 60 minutes in total.

The test car, a Volvo XC 90 (model year 2004), was equipped with a dual headset microphone setup, enabling recording of driver and passenger on separate channels. Two digital video cameras were mounted inside the vehicle, one capturing a close-up of the driver’s face, and the other capturing a wide-screen view of the road ahead. To measure driver workload, a system for performing TDT was utilized in the test. The system consists of a buzzer attached to the driver’s forearm and a response button attached to the index finger. At random intervals, the TDT issues a tactile stimulus to the driver and the driver is supposed to react as quickly as possible on each stimulus by pressing the response button. Driver distraction can then be measured dynamically in terms of user hit-rate and reaction latency [9]. The TDT also enables the measurement of driving-unrelated cognitive load, caused by other cognitive processes generated by the dialogue itself or by memory processing, even when car is not moving, e.g. at stoplights.

2.2. Transcription and coding

For the transcriptions, the transcription tool ELAN (http://www.lat.mpi.eu/tools/elan/) was used. The speech and video data were transcribed using an orthographic transcription. ELAN is able to handle both audio- and video resources, and it allows annotation along multiple tiers (i.e. an utterance can be annotated with several independent annotation schemas). The annotation schema was designed to enable analysis of utterances with respect to disfluencies. The schema uses notions of disfluencies according to [2]. The notion of “utterance” we are using here is approximately “maximal syntactic phrase not interrupted by a long silence”; what counts as a “long silence” varies with context and has not been further operationalized.

The disfluency tier is used for annotating the type of disfluency appearing in the utterance. The following labels are used in the disfluency tier (“*” indicates point of interruption):

- filled pause: the phrase contains a filled pause such as “err”
- repetition: a word or sequence of words is repeated, e.g. “Turn* turn left here”
- deletion: often referred to as “false start”, the speaker starts saying something, interrupts, and then starts again on a new sentence, e.g. “No it’s* do you see the sign over there”
- substitution: a word or sequence of words is changed, e.g. “Do you have a caravan* summer cottage”
- insertion: the speaker starts to say something, interrupts and restarts but with one or more words inserted, e.g. “I want to* I really want to…”
- articulation error: a correction of an articulation error, e.g. “turn lift* left here”

Cognitive load was derived from the TDT measurement:

- workload: high workload according to TDT yields an annotation stretch (typically encompassing several utterances)
- reliability: indicates whether the workload level is reliable or not (enabling discounting of sections where subjects clicked repeatedly instead of only when prompted).

High workload utterances could then be found by searching for annotations where workload and reliability are overlapping, and low workload utterances could similarly be found by searching for annotations where reliability is annotated, but workload is not.

The annotation schema has not been tested for inter-coder reliability, due to limited resources. Instead, annotators have discussed problematic examples and agreed on consensus decisions. While full reliability testing would have further strengthened the results presented here, we believe that our results are still useful as a basis for future implementation and experimental work.

3. Results

Of the four driver–passenger pairs, data from one pair had to be discarded because of a technical problem with the TDT equipment. Since the subjects swapped seats half-way into the test, there is data from all 6 remaining subjects, henceforth labelled A–F, acting as both driver and passenger. Each such session lasted approximately half an hour.

Disfluent speech is an area which still represents a formidable challenge to most spoken dialogue systems, and as expected, the spoken dialogue between our six subjects contained numerous examples of this type of speech. For each
Table 1: Difference between individual disfluency rate under high workload and low workload for 6 drivers. Negative numbers (in boldface) indicate a decrease in disfluency frequency during high workload, positive numbers indicate an increase.

<table>
<thead>
<tr>
<th>Subject</th>
<th>filled pause</th>
<th>repetition</th>
<th>deletion</th>
<th>substitution</th>
<th>insertion</th>
<th>articulation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.002</td>
<td>-0.00004</td>
<td>0.009</td>
<td>0.0</td>
<td>0.003</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>-0.006</td>
<td>-0.01</td>
<td>-0.001</td>
<td>0.001</td>
<td>0.005</td>
<td>-0.001</td>
</tr>
<tr>
<td>C</td>
<td>0.001</td>
<td>-0.002</td>
<td>0.004</td>
<td>-0.003</td>
<td>0.0</td>
<td>0.006</td>
</tr>
<tr>
<td>D</td>
<td>-0.001</td>
<td>0.001</td>
<td>0.02</td>
<td>0.0</td>
<td>-0.001</td>
<td>0.0</td>
</tr>
<tr>
<td>E</td>
<td>-0.008</td>
<td>-0.02</td>
<td>0.01</td>
<td>-0.007</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F</td>
<td>0.004</td>
<td>0.0001</td>
<td>0.020</td>
<td>0.004</td>
<td>0.0</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

Figure 1: Individual disfluency rates for 6 drivers (A–F) calculated as disfluencies/word during high workload (black bars) vs. disfluencies/word during low workload (grey bars).

Figure 2: Average disfluency rate difference for 6 drivers calculated as disfluencies/word during high workload minus disfluencies/word during low workload with standard deviations.

It is clear from Figure 2 that in general for the drivers, their rate of deletions increased under high workload. Other rates remained almost unchanged or decreased, with a more pronounced decrease in repetitions and to some extent filled pauses. On the passenger’s side, Figure 3 shows that on average, all types of disfluencies except articulation errors decreased.

An ANOVA test was carried out, showing that between high and low workload, the difference in repetitions was significant, and between driver and passenger, the difference in filled pauses was significant, both at $\alpha = 5\%$.

4. Discussion

As can be noticed, there is considerable individual variation both in the absolute frequencies and in the distribution across disfluency types. This is well in line with [2], who points out that there seem to be different strategies, with some individuals being typical “deleters”, and others being “repeaters”, but it is also underlined that factors such as speaking rate may play a role. Gender of speaker and listener is also a parameter generally affecting disfluency rates (but which remains to be investigated in the present material). This high degree of variation of course makes it more difficult to draw general conclusions and, from a technical point of view, it also makes it more difficult to design systems that handle disfluencies equally well for all users. The individual variation in disfluency frequencies and distribution might instead call for speaker-adaptive disfluency modelling.

However, our results indicate that the null hypothesis and
hypothesis 1 could be rejected for both driver and passenger. Under high workload, filled pauses and repetitions decreased for drivers while the other disfluencies increased or remained constant. One interpretation of our results is that deletions increase naturally in a more stressful and event-driven situation, calling for repeated changes of topics in mid-sentence, whereas there is unchanged or even reduced time for self-correction by re-phrasing of utterances where one word needs to be inserted or altered. Another, more speculative, interpretation is that filled pauses and repetitions can be seen as devices for turn-taking (floor holders), and the results could be taken to indicate that turn-taking devices become slightly less common during high cognitive load. Deletions, substitutions and insertions function more like self-corrections, where substitutions and insertions might be more cognitively demanding to perform. When making a substitution or insertion it is necessary to discover when the utterance fails, mentally go back and correct the incorrect word(s), and then resume the same sentence including the corrections. In the case of deletions, the speaker only has to discover when the utterance fails and then restart in any way, not having to care about the previous utterance. This might explain why deletions increase most of all disfluencies, while the others are almost constant.

All passenger disfluencies decrease or remain constant during the driver’s increased cognitive load. One, admittedly speculative, way of interpreting this is that passengers may have perceived the increased workload experienced by the drivers, and made extra efforts to be clear and concise.

5. Conclusions

This study has presented some tentative results regarding the distribution of disfluencies in relation to cognitive load in in-vehicle dialogue

- during high cognitive load, all disfluencies on the passenger’s side decrease
- there is considerable individual variation in the rate of different kinds of disfluencies

The investigation presented here only covers some aspects of in-vehicle spoken interaction under cognitive load, and the data collected will be further analysed by the project group. It has already been used to investigate different patterns of sequencing and domain shifts [10], and will be used in the next round of implementation within the DICO project for the design of the system’s dialogue behaviour on all levels.

6. Acknowledgements

The study presented here was carried out within DICO, a joint project between Volvo Technology, Volvo Car Corporation, TeliaSonera, Gothenburg University and the Royal Institute of Technology (KTH) with funding from the Swedish Governmental Agency for Innovation Systems, VINNOVA (project P28536-1). The authors wish to thank Prof. Anders Eriksson, Gothenburg University, for help with the statistical analysis, as well as Johan Jarlengrip, Volvo Technology AB.

7. References