Investigating *negotiation for load-time* in the GetHomeSafe project

Jens Edlund, Catharine Oertel, Joakim Gustafson

KTH Speech, Music and Hearing, Stockholm, Sweden

{edlund, coertel, jocke}@speech.kth.se

Abstract

This paper describes ongoing work by KTH Speech, Music and Hearing in *GetHomeSafe*, a newly inaugurated EU project in collaboration with DFKI, Nuance, IBM and Daimler. Under the assumption that drivers *will* utilize technology while driving regardless of legislation, the project aims at finding out how to make the use of in-car technology as safe as possible rather than prohibiting it. We describe the project in general briefly and our role in some more detail, in particular one of our tasks: to build a system that can ask the driver if *now is a good time to speak about X?* in an unobtrusive manner; and that knows how to deal with rejection, for example by asking the driver to get back when it *is* a good time or to schedule a time that will be convenient.
ö

Index Terms: traffic safety, in-car systems, humanlikeness, proactive behaviour

1. Introduction

The EU funded project GetHomeSafe overall objective is to research, develop and evaluate extended multimodal search and communication systems for safe in-car applications. The project implicates speech as the key modality for these systems. KTH Speech, Music and Hearing collaborates with DFKI, Nuance, IBM and Daimler in the project, which was inaugurated in January 2011 and runs for three years.

KTH heads a work package on *humanlike proactive behaviour*, which lists three main tasks as its responsibility, all of which are outlined briefly in this paper. We also delve a little deeper into the plans and initial steps taken towards the first of these tasks - to unobtrusively negotiate with a driver for attention and time.

2. Background

In a government-commissioned survey from 2011, the Swedish National Road and Transport Research Institute reviews several hundred research publications on traffic safety and the use of mobile phones and other communication devices [1]. Amongst the most striking findings: although there is a broad consensus that visual-manual interactions (e.g. using social media or texting) with communication devices impair driving performance, bans have not had any measurable effects in terms of lowered accident rates or insurance claims. Ban compliance statistics show that typically, bans have an effect on driver behaviour the first year, after which drivers return to their former habits. The degree to which visual-manual use impairs driving varies with task completion time, task complexity, driver skill and traffic circumstances.

Allowing drivers to manage more tasks using speech, which does not occupy hands and eyes, would decrease the time spent in visual-manual interaction while driving, provided that the drivers can be persuaded to use the systems. With bans being virtually ineffective, solutions must be sought elsewhere. Clearly, the systems must work well - a large proportion of errors may well add cognitive load and put the driver at risk. It is also unlikely that drivers can be persuaded to use systems that do not work well. GetHomeSafe is well poised to succeed in this respect, as it takes state-of-the-art speech technology from its partners as a starting point. But using hand-free and eyes-free controls may not suffice. [1] notes that there is virtually no evidence that hands-free telephony is less risky than hand-held use, suggesting that the conversations in themselves may be a risk factor. Here, research is needed to find out when and how this may be the case, and how to adapt the technology to minimize such risks.

3. Humanlike proactive behaviour

The role of KTH in GetHomeSafe is to utilize the flexible speech input and output and the situation modelling developed by GetHomeSafe partners and develop safer dialogues through what we have termed humanlike proactive behaviour. The cover term captures the idea that we will take our cues from human behaviour in similar situations, and that we aim for a system that is proactive in the sense that it will not be content with handling dangerous situations when they occur, but rather will take preventive action in order to avoid these situations altogether. The following decision points serve as an example of the type of reasoning that will be added to the system:

1. Given that the system has information to share, initiate a dialogue only *if the situation is safe*.
2. Given that the system has information to share, initiate a dialogue only *if the driver is willing to talk*.
3. During conversation, be aware of any changes in the drivers attention and the traffic situation. Be prepared to change dialogue decisions at any point in time.
4. During conversation, use humanlike turntaking and grounding to reduce cognitive load.

The main innovation to come out of this work is that where a traditional spoken dialogue system bases its decisions largely on whether it has something to say, what the user has just said, and whether the user is speaking or is silent, a humanlike proactive system will also consider the (traffic) situation, the user's (driver's) estimated attention, and the urgency of the message/task at hand, and it will use this information to vary both its *timing* and the *manner* in which it performs its task.

We aim to enrich the systems with three specific abilities methods that allow us to handle these issues: *unobtrusive attention grabbing*, *user controlled pacing*, and *situation sensitive speech*.

3.1. Unobtrusive attention grabbing

The dialogue systems envisioned in GetHomeSafe are clearly mixed-initiative. They need to be able to inform the driver about changes as the need arises, and so must be allowed to initiate dialogues. Doing so at the wrong time, however, has obvious safety consequences. Keeping track of the driver's cognitive load and the traffic situation will allow us to avoid striking up conversations altogether at the most dangerous times, but times when the situation is safe and the driver seems relaxed may still not be convenient for the driver. A system that bluntly starts reading e-mails or providing traffic information at the first safe opportunity risks causing unnecessary annoyance and disturbing the driver.

The first ability to develop, then, is the ability to call for the driver's attention in a manner that is as unobtrusive and flexible as possible, and that leaves the driver in full control of the situation even though the system initiates the dialogue. A typical example scenario:

> *The system detects a new email from a sender that the driver normally pays attention to. The traffic situation is deemed safe, leaving the system free to initiate a dialogue with the driver. Instead of proceeding to read the email, the system opens with a brief initial dialogue to make sure that the time is appropriate for the driver, an if it is not, to decide when it will be.*

We discuss our initial work towards unobtrusive attention grabbing in sections 4 and 5.

3.2. User controlled pacing

Making sure that the driver has no objections to talking to the system is evidently relevant when initiating a new dialogue, but as the driver's situation may change, the driver should also be the one dictating how the dialogue proceeds, as exemplified in the continued scenario:

> *The system and driver agrees that now would be a good time for the system to read the email. The system proceeds to do so, but halfway through the reading, the driver gets distracted by a traffic event that has not been captured by the traffic monitoring system. The driver tells the system to hang on for a second. If the situation clears up quickly, the driver may tell the system to proceed; if it does not, the driver may briefly instruct the system to stop reading entirely and to give a reminder about the email at some later time.*

The second ability, then, is for the system to be able to understand the driver's instructions to pause ("Hang on", Wait a second"), resume ("Ok where were we?", "Go on"), or cease entirely ("We'll have to do this later, remind me tomorrow"), and to react appropriately. The system output must be able to handle changes mid-utterance and to cut itself short, as well as generating a new continuation that both makes sense in light of what has just been said and corresponds to a potentially new situation.

3.3. Situation sensitive speech

When speaking to each other, people also vary the manner in which they say things according to the situation, which is the third ability we want to endow the system with. The last part of the example scenario serves as an illustration:

> *The system detects a likely increase in the drivers load and in the traffic within half a minute, as the car approaches an area with dense traffic. At the leisurely pace the system is reading, it would not be done in time for this. In order to avoid having to cut itself shot, the system may inform the driver that it will have to speed up because of the upcoming traffic, and then increase its reading speed.*

In general, the system should be able to handle different speaking rates and also to generate more or less wordy (and detailed) utterances in order to speak in a manner that is appropriate for the situation. Naturally, it should also understand and respond to direct instructions of that sort, such as "Hurry up, I need to get going" or "Could you provide a little more detail?".

4. Negotiation for load-time

We now turn in somewhat more detail to our approach to the first task, unobtrusive attention grabbing. Conceptually, we can divide our system's obligations in two categories. The first one, which we will simply call the TASK, is the task at hand: to read an email, plan a trip, make a twitter entry, provide entertainment, or any other typical system task. Although we need to mention these tasks in our description - which would otherwise make little sense - how they are actually handled is both outside the scope of this paper and of our work in GetHomeSafe; our business is with the system's second duty. The second duty is come to an agreement with the driver on when and how to perform the task, and to remain sensitive to any changes in the driver's wishes in this respect.

To study this phenomenon, or language game, we would be helped if we could abstract away from the details of the TASK and keep only what is absolutely necessary to agree on the *when* and the *how*. We know that each task *will* take at least *some* time and effort from the driver. In our model, we simplify this and claim that for each moment the driver and the system is spending on the task, there is a certain effort spent by the driver - we call this the driver load, leading the mind to cognitive load (which is not coincidental). If we sum the loads for each time frame over the completion of the task we get the task load-time. For purposes of reasoning, we will assume that driver load is expressed in terms of the proportion of the driver's total capacity that is consumed by the task at a given moment. Following this, we refer to the second duty as *load-time negotiation* (LTN). In the mock dialogue below, the utterances that belong to LNT are in bold-face:

	Situation:	Car is parked, driver is about to take off. Driver is going for a trip by plane the next day. The system knows this, but lacks some details.
1	System:	**I need your travel details in order to plan the journey to the airport. Do you want to enter them on the keyboard right away, or will you tell me while driving?**
2	Driver:	**Lets do it while I drive.**
3	System:	OK, when does the plane take off?
4	Driver:	At 17.00 hours.
5	System:	Are you leaving from home or from the office?
	Situation:	The driver is distracted by a pedestrian who appears to be about to cross the road.
6	Driver:	**Hold on a second.**
	Situation:	The driver passes the pedestrian.
7	Driver:	**OK, where were we?** I'll leave from the office.
8	System:	OK, that's all I needed. Thanks.

We make the further simplifying assumption that we can keep the TASK and LNT dialogues untangled and deal with them more or less separately. In order to handle load-time negotiation, all we need to know about the task is the load and time it will cost to complete. If a task that will take five minutes to perform and will require most of the driver's attention, the system must negotiate for five minutes of the driver's time during which the car should be parked. A task that takes three minutes to perform at very low load requires the system to negotiate for three minutes of the driver's time, during which the car can be moving as long as the traffic situation is reasonable. By abstracting away from specific TASKS, we make it possible to use data from sources outside of the car to learn how to perform LNT, which is helpful as this type of negotiation ("Excuse me, do you have a minute to spare?") is common in other contexts as well. But the model has other, potentially quite strong, advantages as well.

If a task can be completed in more than one way - in the mock dialogue above, the driver can provide the required information either by typing it into a form or by speech - we can create an abstract representation of each of these using a *load-time profile*. In a load-time profile, we plot the drivers load for each time frame (these can be of arbitrary length) in a graph. The profile tells us what the duration of the task is, what the highest load required during its completion is and when it occurs, and so on. Using a keyboard and a form to provide structured information, for example, may well be the fastest way to do it, but it is combined with a high load as it requires both hands and eyes to complete. Using a speech interface to accomplish the same thing may take slightly longer, but at a lower load. Figure 1 exemplifies.

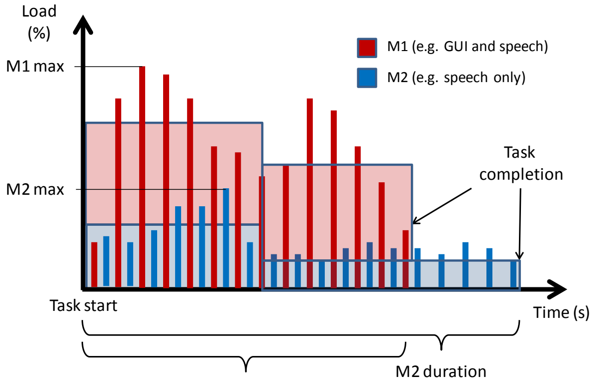

Fig. 1: Load-time profiles for two methods of input. The y-axis represents load as the proportion f the drivers full capacity, the x-axis represents time spent performing the task. Bars are frame-by-frame estimates of load, and the blocks are smoothed estimates over longer stretches of time. In this example, the method utilizing a GUI is faster, but requires more of the user.

How we acquire a load-time profile for a given means of performing a task is not specified in the model. It could be hand-coded from intuition or wild guesses, as is the case in the example in Figure 1, or it could be an automatic estimation captured when from a driver actually performing the task. Given some means of estimating the load handling the task places on a diver in each time frame, load-time profiles could be trained experimentally and honed over time. Note that this would also allow us to adapt the profiles. As a driver gets more proficient, load may decrease for a specific task. This could be captured by continuously decaying old observations and adding new ones to update the profile. Further improvements could be made. For example, the reliability of a profile can be gauged by tracking the variance of the data it is built on.

Before we arrive at load-time negotiation, we need to model one more factor. If we assume that the system can perform as varying tasks as warning about upcoming traffic conditions and delivering the latest tweets, it is clear that it needs the concept of *task urgency* to ensure that it behaves differently when the driver is about to pass the last exit for miles leading to a gas station with an all but empty tank than it does when Ashton Kutcher has added a new tweet.

Provided that our system has an idea of the urgency of a task, and knows the load-time profile of each available method to complete the task, it is ready to negotiate for load-time to complete the task. As with every action in the GetHomeSafe system, the decision of how to go about it is a combination of what the system knows about the current situation (e.g. traffic , driver load) and what it can glean from the driver. A lot of logic can be applied without asking the driver. If for example one of the available methods to perform an urgent task takes too long (i.e. the task has a latest completion time) and another does not, the system need not suggest the too-slow method at all, but can tell the driver that the other method is the only viable option. Likewise, if the system traffic situation is complex and one of the available profiles requires a lot from the user at some point (i.e. has a high max load), the system need not suggest that method, or may combine such a suggestion with the requirement that the driver first stop the car. Abstracting away from actual tasks to load-time profiles and task urgency allows us to compare and reuse different tasks with similar time and load requirements and to experiment with prototype systems for LTN without the need to know everything about future tasks, as we can initially make up mock task profiles.

5. Human-human data collections

We are currently preparing to collect data for the first of our main tasks: unobtrusive attention grabbing. This ask amounts by and large to handling exchanges like utterances 1 and 2 in the dialogue example above, or more simple initial utterances such as "Can you spare me a minute?". This language game - opening up a conversation with a question about the appropriateness of that very conversation - is common enough in real life (ask any phone sales person), but is not well-described in dialogue research. Our hope is to be unobtrusive by doing the familiar - by doing what humans would do - so the obvious starting point is to capture how humans negotiate for load-time. For safety concerns as well as feasibility reasons, we will not record people interrupting the driver in live traffic. Since our model abstracts away from the details of the task to be performed, we can allow ourselves to initially record the interaction in our normal working environment, which allows for much greater access to subjects and will result in a considerably lower cost per recorded dialogue.

The requirements for our data is that they be representative for how people behave when they negotiate for load-time under different circumstances - more specifically, for tasks that (a) require different levels of attention; that require (b) different amounts of time to complete, that (c) can be completed in different manners with different such requirements; that (d) are associated with varying urgency; and (e) that need be performed when he participants are in a variety of situations and under different loads.

We are currently performing an initial low-tech data collection in order to test the basic setup, and preparing for a second, more complex collection using a variants of the human-human manipulation methods we used in [2] and [3].

5.1. Initial data collection

The ongoing, initial data collection uses a simple setup: A number of persons from our research group act as subjects over a prolonged period of time - we expect the collection to go on for at least a week. The collection is arranged like a tag game, where the person who is "it" (the *tagged*) receives a wireless microphone and is told to check for email instructions. The experiment coordinator then selects one of a number of possible tasks from a list, a second person (the *accomplice*) with whom the tagged must perform the task, and a deadline for the task. The tagged receives this information in an email. When the tagged attempts to perform the task the wireless microphone must be worn and turned on. We will then get a recording of the tagged entering the office of the accomplice and asking for assistance. This occurs in real life on the subjects real work hours, so the accomplice may well be busy and reject the proposition. When the task is eventually completed, the accomplice becomes the tagged and awaits an email with task information from the coordinator.

There is room for a fair deal of control here from the coordinators side. We may for example select tasks that are lengthy at an inconvenient time, or simple tasks but with a very short deadline, or tasks that can be accomplished in a multitude of ways. Note that although we are not interested in the way the subjects actually perform the tasks, but rather how they negotiate for when and how to do them, they must have real tasks to negotiate about, or the negotiation dialogue would not reflect human behaviour since nothing would be at stake. In order to minimize the detrimental impact on the work performance of the participants during the data collection, most of the tasks chosen will be things that need doing in any case, such as emptying waste bins, changing the printer cassettes, planning a work trip, or discussing a common project.

The LTN parts of the captured dialogues will then be transcribed and analysed, whereas the dialogue that concern actual task completion will be annotated with a load-time estimation.

5.2. VoIP based human-human manipulation

The second, main data collection incorporates any fixes to the general "play tag" collection methodology that are deemed necessary after the initial collection results are evaluated. In addition, it differs in two major ways: the way the dialogues take place and the way the recordings are treated.

The dialogues are recorded over Voice over IP rather than using a wireless microphone to capture face-to-face dialogue. The reason is that the former setup differs from the situation we see in a car in several important ways. Most importantly the accomplice may well see the tagged coming well before the conversation starts in the former case - there might even be a knock on the door, which in practice already accomplishes the first steps of the LTN. But in a car, where both parties are already present - the system is seen as present in the car in the same way as a passenger is - knocking or ringing before speaking is not very intuitive. A passenger might instead speak straight out, or try a tentative throat-clearing, to get attention and start a conversation. In order to create an in-office situation that allows for this type of interaction, we will setup a system where any participant can speak out loud in any other participant's room without prior notice using simple loud-speakers and microphones connected over the office LAN. This is similar to any IP telephony system, except we eliminate the ring tones. Instead, participants press a button and speak, then wait for the other party to open the other channel so that the conversation can begin. Conversations are recorded like before, with several additions to the recording mechanism. The objective of these changes is to create a situation where the participants in effect build their own Wizard-of-Oz interfaces, which we can then use as a foundation for our system's behaviour.

1. During dialogues, speakers are requested to push a button while talking. Releasing the button does not mute the system; the button press start and end times are a means to get estimates of start and end times that agree with what the speaker intends. Alternatively, we may use a VAD for this purpose.
2. After each interaction, participants are required to classify their utterances as LTN, task oriented, or both.
3. After each interaction, participants are required to label their utterances using a simple tool. Labels can be reused from previous interactions or created.
4. Each participant is required to listen through a resynthesis of each LTN utterance to ensure that the utterance is viable for prosodic analysis, as in [4].
5. The labelled and verified utterances are used to create a prompt piano (5) that each participant is encouraged to use in dialogues whenever possible. The participant can chose freely to either speak or click on a button to replay a previously recorded and labelled utterance.

By taking these measures, each participant will slowly build a personal Wizard-of-Oz interface We measure this progress by continuously gauging the proportion of spoken versus replayed LTN utterances. By the end of the data collection, but before we have done any deeper analysis of the recorded data, we hope to have a fair grasp of what is needed to implement basic LTN dialogues from inspecting the labels and transcriptions in the prompt pianos alone.

6. Conclusion

We have described our role and initial efforts in the newly inaugurated GetHomeSafe project. The dialogue techniques we develop in the project have in common that they can be viewed as meta-dialogues; as dialogues about the dialogue. If the past decade has brought us great skills in how to design task-oriented dialogue, we may perhaps say that spoken dialogue design is now entering the era of meta-dialogue, in which control over the spoken dialogue system's own behaviour is gradually handed over to the user.

7. Acknowledgements

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

8. References

[1] Kircher, K., Patten, C., & Ahlström, C. (2011). *Mobile telephones and other communication devices and their impact on traffic safety: a review of the literature*. Technical Report VTI 729A, Stockholm.

[2] Edlund, J., & Beskow, J. (2009). MushyPeek - a framework for online investigation of audiovisual dialogue phenomena. Language and Speech, 52(2-3), 351-367.

[3] Gustafson, J., & Merkes, M. (2009). Eliciting interactional phenomena in human-human dialogues. In Proceedings of SigDial 2009..

[4] Gustafson, J., & Edlund, J. (2008). expros: a toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 293-296). Berlin/Heidelberg: Springer.

[5] Gustafson, J., Heldner, M., & Edlund, J. (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 240-251). Berlin/Heidelberg: Springer.