Exploring Natural Interaction in the Car

Niels Ole Bernsen
NISLab
University of Southern Denmark
+45 65 50 35 44
nob@nis.sdu.dk

Laila Dybkjær
NISLab
University of Southern Denmark
+45 65 50 35 53
laila@nis.sdu.dk

1 Abstract

The paper discusses the problems of selecting, and assigning appropriate roles for, the individual (unimodal) modalities which could be used for exchanging information between the car driver and a multiple-task virtual co-driver system. The system is currently being developed by the authors and partners in the VICO (Virtual Intelligent CO-driver) project. The strategy adopted for addressing those problems is to first apply modality theory and then use empirical Wizard-of-Oz (WOZ) simulation to decide issues which the theory does not cover in its present form. The paper present the results of applying modality theory followed by the results of an early WOZ experiment with the VICO system.

1.1 Keywords

Natural interaction, multimodality, in-car system, navigation assistance, modality selection.

2 Introduction

Car navigation systems are gaining ground in the market. They are probably still being used mainly by professional drivers such as truck drivers and taxi drivers but are increasingly being sold to private car owners as well. There are several brands to choose among, such as Blaupunkt TravelPilot, Philips's CARin, VDO Dayton, and Pioneer. They all use GPS (Global Positioning System) and have a CD-ROM with digital road maps which holds the information the system needs to provide navigation assistance. The systems are operated by the user in much the same way. A navigation system typically comes with a display and a remote control. The display may be small and without map information or it may be somewhat larger and display a map in addition to the textual and iconic information which is available on the small display. Route instructions are provided to the user by voice output, by an arrow on the display showing in which direction to turn next, and possibly by a map showing the present location and direction of the car. The output seems generally to work quite well and without overloading the driver, at least as long as the driver does not divulge in studying the map details. This, however, is not the case with the input.

Inputting a destination is not only cumbersome and unnatural but tends to absorb the driver's attention to a degree that may easily cause dangerous situations. Taxi drivers we have spoken to are usually aware of this and do not input information when they are in complex traffic which demands their full attention. Still, today's navigation systems interfaces remain highly questionable from a safety point of view even in low-traffic conditions and no matter who is driving. It cannot be assumed that all drivers will respect the system's advice not to input navigation instructions whilst driving. And even if they did, the growing functionality of the device (voice output control, map zooming control, populating the map with petrol stations, post offices and numerous other information items at will, etc.) makes it almost impossible for them to keep their fingers off the remote when in the traffic. The main problem is that the driver has to spend too much time looking at the screen and possibly also at the remote control. Thus, to input a destination, the driver first has to scan the options currently available on the display and choose one with the remote control. A click on the OK button on the remote leads to new information being displayed and the user must scan the screen again. Eventually, an alphanumeric table appears and now the user must spell the name of the destination by selecting letters one by one using the remote control, cf. Figure 1. The system provides some help by displaying, e.g., the city names which start with the letters selected so far. As this list cannot always be scrolled, however, the driver must continue to input letters until the desired city name appears on the display and can be selected. The driver then goes on to spell the street and possibly input the street number digit by digit.

Figure 1. Inputting a destination to an ordinary car navigation system.

It seems obvious that spoken input could make the in-car navigation input process much more efficient and user-friendly, and probably much safer as well. This has been realised by some navigation systems providers who offer simple voice command input. Even if this is an improvement it is far from achieving spoken natural interaction with a navigation system in which the user negotiates navigation goals in free-initiative spontaneous speech and not via hard-to-remember command keywords.

3 The VICO system

Natural interaction with an in-car system is being addressed in the European HLT-project VICO (Virtual Intelligent CO-driver) which began in March 2001 and has a duration of three years. VICO will build two prototypes of a natural interactive and multimodal in-car spoken dialogue system for English, German and Italian. The first prototype will enable navigation assistance, including streets and street numbers, parts of cities, cities, parts of country, petrol stations and hotels, as well as information about the VICO system itself. The second prototype will add hotel reservation over the web, scenic route planning including web-based access to information on touristic points of interest, such as castles and churches, car manual information, and spoken operation of in-car devices. Throughout its interaction with the driver, VICO will maintain some amount of situation awareness with respect to the car. For instance, VICO will stop speaking when the car brakes are being applied. The project partners are Robert Bosch GmbH, DaimlerChrysler AG, Istituto Trentino di Cultura, Phonetic Topographics N. V., and NISLab. NISLab will develop VICO's natural language understanding, dialogue management and response generation components.

VICO's two input modalities are speech and a (haptic modality) push-to-talk button for activating the speech recogniser to start a dialogue with the system. Output modalities include a graphics modality green/red light on the car screen which signals if the recogniser is active or not, speech, and additional graphics modalities on the car screen. The first prototype will provide graphics text output whereas the second prototype is envisaged to also display road maps and/or route icons. The prototypes will integrate an already existing navigation system from Bosch to the extent possible. The idea is that when it is clear where the driver wants to go, the route planning and driving instructions can be generated by an already existing system which works quite well.

4 Interaction with VICO

It is well-known by now that, generally speaking, it is far from simple to select and combine several different modalities for novel tasks in a way which ensures smooth and cooperative interaction with users. Wizard-of-Oz (WOZ) experiments [Bernsen et al. 1998] are often used to collect data on interaction adequacy and user satisfaction during the development of novel systems and interfaces. If done properly, WOZ simulations provide valuable data for evaluating the (fully or partially) simulated system and providing directions on how to improve it. However, WOZ is an expensive and time-consuming method. Moreover, it is not easy to simulate an in-car system in a way which makes the user believe to be interacting with a real system. This means that simulated in-car WOZ results may be less reliable than results collected with, e.g., a simulation of an over-the-phone spoken dialogue system where it is fairly easy to hide the fact that the system is (in part) simulated by a human. Still, WOZ would seem to be second only to useful theory for specifying novel multimodal interfaces. To the extent that theory can be applied, we do not need WOZ experimentation and, when theory cannot provide guidance any longer, WOZ can be used to explore the remaining questions of detail.

Modality theory analyses all possible unimodal modalities in the media of graphics, acoustics and haptics at different levels of abstraction [Bernsen 1994, Bernsen 2001]. The theory has been applied to the analysis of speech functionality, i.e. of when speech can (not) be used in a multimodal context [Bernsen 1997, Bernsen and Dybkjær 1999]. In particular, it was shown that a mere 25 modality properties were sufficient to evaluate a total of 273 speech functionality claims made in the literature 1985-1995. *Modality properties* are inherent properties of unimodal modalities, i.e. those elementary modalities which go into the creation of multimodal representations. Modality properties help determine the usefulness in context of a particular modality. Modality properties are in italics below.

The first VICO prototype will integrate input speech and haptics with output speech and graphics text. The main problem is the multimodal output integration of output speech and graphics text. We have applied modality theory to the problem of whether and how to combine speech and graphics text output in the car to see how far theory can help. Both speech and graphics text are linguistic modalities, so there is probably little to distinguish them in that regard for the tasks at hand.

The easier part is to justify the use of speech output in the car. We can ignore haptic modalities in what follows because there is no way that ordinary car drivers could handle haptic output from VICO for the tasks VICO has to solve. For the sake of completeness, the same applies to olfactory and gustatory output. This leaves us with the choice of acoustic and graphics modalities. The priority of the car driver is to steer the car safely through the traffic. Speech, being an acoustic modality, shares the property of acoustic modalities of *not requiring limb (including haptic) or visual activity*. Moreover, speech, being acoustic, is *omnidirectional*. Graphics, on the other hand, is *neither eyes-free and hands-free, nor is it*

omnidirectional: one has to look in a particular direction to receive graphics output and, for the time being, cars do not offer semi-transparent wind screens, or corresponding glasses for the driver, on which to display important textual information graphically. Finally, speech input/output modalities in native or known languages have *very high saliency*. This means that spoken output is eminently suited to catch the driver's attention whilst driving, even if the driver is listening to the radio or having conversation with passengers. In principle, this latter modality property of speech might distract the driver's attention from the traffic in situations where attention to the traffic has top priority. However, people already talk when driving together, and the saliency of VICO's output is hardly more of a liability than the conversation made by passengers. Moreover, the second VICO prototype will include an element of situation awareness which will make its speech synthesiser stop immediately when, e.g., the car brakes are being applied. Contrary to some passengers, therefore, VICO will not scream but go silent. In conclusion, there seem to be compelling theoretical reasons for preferring speech output over graphics output for tasks to be carried out by VICO, especially to the extent that these tasks require linguistic output.

The more difficult question is why in-car speech output should be complemented by graphics text output. In VICO, the original decision to include graphics text output was made not for compelling theoretical reasons but because it was agreed that in-car displays, which are now common even in low-price cars, are not likely to go away. In principle, therefore, we could just abandon the car display for conveying VICO information. However, a first, if not reason for using a display then, consideration is that no unimodal modality is perfect for all information representation purposes, users and environments. One drawback of using speech-only for conveying information in the car is that speech is a dynamic modality, and dynamic input/output modalities, *being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.* Freedom of perceptual inspection means that the user can spend as much time as desired (or as safe) on decoding the information presented in some modality. With speech, this is not possible: one either gets the message in real time or it is (normally) lost forever unless there is a way to get it repeated.

The car driver's attention is likely to shift over time primarily depending on the traffic but also depending on, e.g., conversation with passengers in the car. Conceivably, both factors might distract the driver from listening to VICO. It is not necessary to invoke stressful or even disaster scenarios to illustrate this point. All it takes is a heavy-traffic intersection which the driver wants to cross. In all such cases, having the most important VICO information presented on the car display might allow the driver to get the dialogue with VICO back on track, or remind the driver of what has already been agreed with VICO, once the distractions have gone away. The reason why the graphics display can do that is that its presentation of text is static: static graphic/haptic input/output modalities allow the *simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.*

Arguably, however, VICO might allow the driver to get any information repeated on demand when the traffic and/or fellow passenger distractions have gone away. Moreover, the theoretical arguments from modality properties above do not solve the more detailed question of which, and how much, information should be presented on the car display. If all spoken output is being presented on the display, we would seem to have reverted to the obvious danger of using text graphics rather than speech as discussed above. If only some spoken output is being presented on the display, which should it be? It should probably be the "most important" information, but what does that mean, exactly? Is this merely the final conclusion of a route dialogue, such as that the goal is X? Is it also a list of five cities with the same name but located in different parts of the country? Modality theory does not presently have the answer to these questions apart from the implication that the car display should be used sparsely during the spoken dialogue with VICO. To answer the questions, we have to resort to Wizard-of-Oz (WOZ) simulation.

5 Wizard-of-Oz setup

The purpose of the WOZ experiments presented here was to get a first approximate idea of the type and amount of textual information to be presented to the driver on the car display. To get reliable feedback, we needed to simulate the traffic distraction experienced when driving. Without some appropriate amount of distraction, the subjects would have ample time to scrutinise the car display, which is not realistic in a driving situation. The setup used was the following.

Subjects were equipped with a force-feedback steering wheel and pedals (accelerator and brakes) and seated in front of a 42" flat screen displaying a car computer game. They were asked to play the car game in the easiest possible mode, controlling the car with the steering wheel and pedals. Next to the large screen was a small portable computer simulating the car display. Textual system output was displayed on the screen. Subjects provided spoken input to the system via headset. Spoken system output was provided via the loudspeaker of the portable computer and not via the headset. The setup is shown in Figure 2. All sessions were video recorded for subsequent analysis, including the user's spoken input and the system' spoken output. The camera was operated by the experimenter.

In another office, a wizard typed the key contents of the user's input into a semantic frame acting as interface to

the dialogue manager. The dialogue manager processed the input and sent the result (a semantic representation) to the response generator which processed the frame and sent an output text string to the text-to-speech (TTS) component. The TTS component (Festival) then sent the spoken system output to the user. At the same time, the wizard sent text output to the car display using NetMeeting which was also running on the portable computer next to the user.

Figure 2: Wizard-of-Oz setup.

Each user was given three scenarios all of which were expressed graphically by highlighting the relevant city names, street names, etc. on two maps. One map would encircle the user's current position whereas the second map would encircle the user's destination. Graphics scenarios were used instead of textual scenarios in order to avoid priming the user's way of expressing the destination [Dybkjær et. al 1995]. Eight different scenarios were used.

Only three subjects were used in the experiments reported. All subjects were colleagues working in the same department a the VICO developers but on different projects. Thus, although the subjects had some idea of what the VICO project is about they were not familiar with any details. Two subjects had a background in computer science and one had a humanities degree.

6 Observations from the Wizard-of-Oz experiments

Each subject was briefly introduced to the scenarios and to the setup including the computer game, the spoken input/output and the text output. Subjects were told that the purpose of the experiments was to identify the amount of textual system output which would be appropriate for the VICO system. They were not encouraged to follow any particular strategy, such as to make sure they knew where they were supposed to go before they started "driving".

Three different text output versions were used, one for each scenario done by a particular subject. The versions were presented to each subject in a different order. One version presented an exact text copy of the spoken system output on the screen. A second version only displayed text output corresponding to each piece of key information provided by the user. For instance, the system would display the name of a city provided by the user, adding inferred information, such as the part of country in which the city is located, or the system would display a list of cities-cum-parts of country corresponding to the city name provided by the user. Once the user's intended destination had been uniquely identified, that destination was displayed in full. The third version only displayed the uniquely identified destination once it had been agreed upon between the user and the system.

Our hypothesis was that the second version would be the preferred one. Version one is too verbose and dangerous, as argued above. Version three, on the other hand, does not provide the kind of continuing external memory support, which the driver might need in case of distraction (cf. above).

The WOZ experiments provided a number of suggestions for improving the VICO system, the scenarios, and the way in which to run the experiments. Examples are mentioned in the following.

The implemented part of the dialogue manager was limited in functionality and only had access to a small test database of addresses. No additional functionality was simulated. For the next set of experiments it will be important to add, e.g., a repeat function since the subjects frequently tried to get VICO to repeat what it just said. There were two main reasons for asking for repetition. One was the quality of the synthetic output, the other was the distraction caused by the car game which meant that subjects sometimes stopped listening because they had to concentrate on handling a traffic situation.

Often subjects did not know the place they were supposed to go to and in some cases they were asked to clarify, e.g., which city or which street they had in mind if there were several of the same name. The problem was that the subject typically would not have a clue since the subject knew none of, e.g., the cities being offered and was not going there for real so the subject had no general geographical intuition about it either. Thus, usually subjects would look at the map to find some other name to provide the system with, hoping that that would be of any use in disambiguating the destination. Usually, it was not, and this was not least due to the very limited coverage of the test database. The next iteration of the scenarios will ensure that the destination is clearly marked and that the map provides sufficient information for subjects to have a general idea of where they are going.

All three subjects spent relative large amounts of time scrutinising the scenario map while they were "driving". This is, in most cases, not a very realistic situation. Thus, in the next experiments subjects will be explicitly asked to make themselves familiar with where they are going before starting the "car".

We did not manage to collect much information on the appropriateness of the text display versions we wanted to test. After interacting with the VICO system each subject was asked about his opinion on the different kinds of text feedback on the car display. However, none of the subjects had actually paid much attention to what was being displayed on the screen. They had all been busily occupied by "the car and the traffic". One subject explicitly stated that he found the spoken system output completely sufficient and did not bother looking at the display. Thus, an immediate result of the experiment is that the need for text output from VICO is very limited. This conclusion, however, is somewhat contradicted by the subjects' stated needs for VICO to repeat what it just said. So, another, equally possible conclusion is that subjects need to better master the car game before starting the experiment. Finally, we will have to study what happens when VICO will be able to orally repeat what it just said. We therefore need to run another set of experiments which include the improvements mentioned above and with more subjects. It is not unlikely that subjects will spend more time on the car display if they are not so busily occupied with the scenario maps and if they are more familiar with the car simulation.

REFERENCES

1. Bernsen, N. O.: Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers* 6, 4, 1994, 347-71.
2. Bernsen, N. O.: Towards a tool for predicting speech functionality. *Speech Communication* 23, 1997, 181-210.
3. Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. Chapter to appear in Granström, B. (Ed.): *Multimodality in Language and Speech Systems.* Dordrecht: Kluwer Academic Publishers 2001.
4. Bernsen, N. O. and Dybkjær, L.: A theory of speech in multimodal systems. *Proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems,* Irsee, Germany, 1999, 105-108.
5. Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: *Designing Interactive Speech Systems. From First Ideas to User Testing.* Springer Verlag 1998.
6. Dybkjær, L., Bernsen, N. O. and Dybkjær, H.: Scenario design for spoken language dialogue systems development. *Proceedings of the ESCA workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 1995, 93-96.