Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz Experiments

Petra Geutner, Frank Steffens and Dietrich Manstetten

Robert Bosch GmbH
Corporate Research and Development
P.O. Box 10 60 50
70049 Stuttgart, Germany
Petra.Geutner@de.bosch.com

Abstract

Steadily increasing dissemination of computer applications has resulted in an ever-growing functional complexity of electronic services and devices. Hence the utilization of natural language is highly desirable to facilitate their usage, especially in the automotive environment where safety is a mandatory requirement. User-friendly, comfortable and safe vocal interfaces that ensure natural interactivity are needed. VICO, the **V**irtual **I**ntelligent **Co**-Driver, aims at the development of an intelligent conversational agent enabling ubiquitous natural interaction between humans, and digital devices and services in the car. This paper gives an introduction into the key objectives and goals of the VICO project. It presents detailed information about design and experimental setup of the performed Wizard-of-Oz experiments to evaluate expectations of potential users early in the design and development process. The results of the conducted experiments are introduced. The subjective ratings of test persons towards the evaluated simulated prototype system were very high, and the speech-controlled approach considered as extremely easy-to-use. Finally, conclusions as well as consequences of the perceived results on design and development of the first prototype VICO system are described.

1. Introduction

Steadily increasing dissemination of computer applications has resulted in an ever-growing functional complexity. As a consequence, comfortable and user-friendly operation of such systems and services come more and more into focus recently. This especially applies to the automotive environment where safety is a mandatory requirement. Here, in recent years speech recognition has been used as an alternative and extension of the traditional tactile interfaces used in driver information systems.

First simple command-and-control applications have been studied in research projects, such as the European VODIS project (Geutner et al., 2000). As an outcome of these research activities simple command-and-control applications are already employed in current commercially available in-car systems. Although speech control of these systems obviously is advantageous over traditional tactile interfaces, the operation of a complex functionality, as e.g. a navigation request, can be a time-consuming and tedious task. Most of these command-and-control systems offer a strictly menu-based input structure guided by a very rigid dialogue, and requiring to enter each item of a complex user request separately. Hence, the utilization of natural language as means of operation is a logical consequence to facilitate the usage of in-car services and devices.

This paper introduces the VICO project, where the overall objective is the creation of a conversational speech interface allowing natural, user-friendly, safe and comfortable communication with a virtual co-driver under adverse conditions in an automotive environment. As part of the design process of the VICO spoken language dialogue system, the evaluation of user expectations by Wizard-of-Oz experiments are presented. To this account the experimental setup, realization as well as the results of the performed experiments are described.

2. The VICO Project

VICO[1], the **V**irtual **I**ntelligent **Co**-Driver, is a European project partly funded by the European Commission, with five participating partners: Robert Bosch GmbH (Germany) as initiating party and consortium leader, DaimlerChrysler AG (Germany), ITC-irst (Italy), the University of Southern Denmark (Denmark) and Phonetic Topographics N.V. (Belgium).

VICO aims at the development of an intelligent conversational agent enabling ubiquitous natural interaction between humans, and digital devices and services. Focusing on the automotive environment, the project will provide a user-friendly natural language interface for in-car devices and innovative services, such as navigation, on-the-fly route planning, interactive hotel reservation, customized sightseeing tours, an electronic car manual, and news services. Key objectives of the project focus on:

- robust speech technology
- user- and situation-adaptive intelligent dialogue strategies
- user-friendly, comfortable, efficient and safe-to-use vocal interfaces
- user acceptance and satisfaction

3. Motivation

User-friendliness and safety are crucial criteria within the design and development cycle of modern spoken language dialogue systems, especially in the car environment. The consideration of user expectations becomes more and more important for such systems and hence has great influence on their development. To ensure user satisfaction and acceptance as well as traffic safety, continuous system evaluation during the development

[1] http://www.vico-project.org

process is of key importance. In successive development phases each cycle includes the establishment of either a simulated test bed or a prototype demonstrator in various stages of a system, which is undergone a thorough evaluation both by experts as well as test subjects.

Within VICO the prototype demonstrator system will integrate all approaches developed during the project. User evaluations will be conducted in simulated as well as real driving situations all throughout the duration of the project to improve the design and usability of the resulting spoken language dialogue system.

To allow the integration of user feedback into the evolution process as early as possible as well as to ensure acceptance by potential users of a virtual co-driver, user evaluations have already been performed even before the establishment of a first prototype system. Prospective users have been involved in the system design process by conducting Wizard-of-Oz experiments integrated into a driving simulator. Several aspects were to be monitored through this process:

- How do drivers interact with a virtual co-driver when allowed to use natural language input?
- What are the user's reaction to such a system?
- What are the user expectations towards an in-car natural language system?
- What are the distraction effects on driving behaviour?

The benefit of a Wizard-of-Oz approach (Dahlbäck et al., 1993) is the possibility to test hypotheses about not yet implemented systems by simulating a non-existing spoken language system. The technical setup of the used simulated environment as well as the VICO functionalities taken over by the wizard will be described in the following section.

4. Experimental Setup

For the VICO system user expectations were to be evaluated even before a first working prototype had been technically realized, allowing to integrate the outcome of these user experiments into the design process already. To this account a human being, the wizard, had the task to simulate a working VICO prototype. A driving simulator was used to get as close as possible to a real driving situation without conducting the experiments in real traffic.

4.1. Driving Simulator

To run the experiments Bosch used a fixed-based driving simulator. Hardware and software are based on the commercial simulator STISIM 500W provided by Systems Technology Inc. It consists of a local network of four PCs. Pedals and steering wheel are integrated in a mock-up of a vehicle's front half. The instrumentation is attached to the simulator as well. The driving scenery is projected onto three front screens of 180x135 cm each, providing a total field of view of 135°. Rearview mirrors are integrated in the front image scenery. Figure 1 gives an impression of the used setup.

While performing the experimental tasks of the Wizard-of-Oz scenario the test subjects simultaneously had to drive a fixed route in this driving simulator (Manstetten et al., 2001).

Figure 1: View of the Bosch driving simulator

4.2. Environment of the Wizard

Main challenge of the experimental setup used for the Wizard-of-Oz experiments consisted in the development of the work environment and tools to be used by the wizard. The wizard was located in a separate room where the video image of the driver and scenery were transmitted including an acoustic channel for perceiving the driver's utterances. Figure 2 shows the video and driving information accessible to the wizard.

A software interface for the wizard had been designed offering a library of pre-defined system prompts to be triggered as vocal reactions to requests of the test subjects. Upon a user input the wizard would choose an adequate reaction by simply clicking on the respective library entry. The selected text was synthesized using text-to-speech software, and then transmitted to the driving simulator room with the test subject. This vocal response was perceived by the test subjects as the system reaction of VICO.

Figure 2: Video image recorded during experiments

Figure 3 shows a selection of prompts to be used by the wizard. If none of the listed answers was an appropriate reaction to a user request, the wizard was also able to type in a new text. The on-the-fly generated system reaction was then sent to the speech synthesis component of the setup. This procedure sometimes being time-consuming, the phrase "Einen Moment bitte" ("One moment please") could be used to bridge the time between user input and system response. The same phrase was also used when complex tasks, e.g. hotel reservation, had been requested by the user. As in a true technical system the reaction to a complex request would have taken some time to be fulfilled, this was also simulated in the performed Wizard-of-Oz experiments. Another option to react to unforeseen queries was the answer "Hierzu liegen keine Informationen vor." ("No information is available on this topic.").

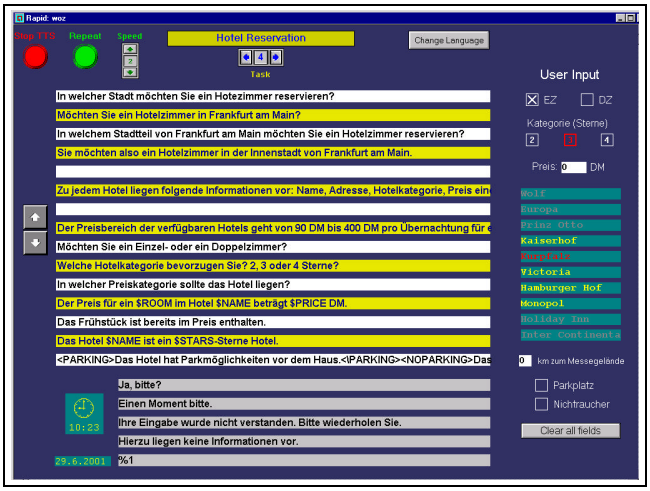

Figure 3: Software interface of the wizard

When trying to break down the task of the wizard into system components of a real VICO system, the wizard in our experiments was simulating the system components speech recognition, natural language understanding, dialogue modelling, and response generation (see also Figure 4 below).

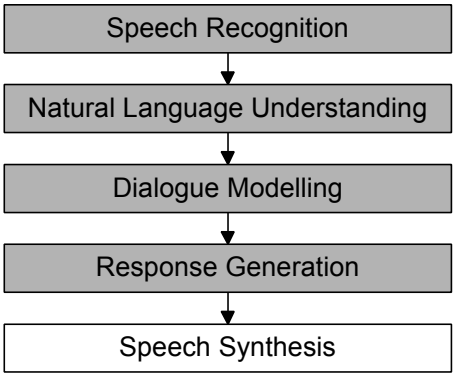

Figure 4: System components simulated by the wizard

5. Wizard-of-Oz Experiments

5.1. Test subjects

A cohort of ten subjects was selected to participate in the performed experiments. Even though the scope of this experiment was very small, special care was taken to achieve a balanced mixture of test subjects concerning age, gender and technical background. In the end, five male and five female subjects with the age varying from 24 to 60 took part in the simulated VICO experiments.

5.2. Task Description

Each of the test subjects had to execute seven tasks. Definition of these tasks focused on the application areas to be covered by the VICO project and consisted of:

- time information
- navigation
- tourist information
- car status
- car manual
- hotel reservation
- traffic information

The used scenarios are illustrated in Table 1 below. Whereas most tasks had to be initiated by the user, two of them (task 4 and 7) were prompted by VICO itself. In this case the wizard triggered a speech prompt of the simulated system expecting a reaction by the user. For the user-initiated tasks the task definitions from below were given to the test subjects as pre-recorded messages during their test drive.

No.	Task	Task Description
1	Current Time	"Ask the VICO system about the current time."
2	Navigation	"You are in the Gerlingen area. Use the navigation component of the VICO system to be directed to the trade fair in Frankfurt."
3	Tourism	"On your way there is some time left. Ask VICO about tourist attractions in Heidelberg."
4	Car Status	"Given the current driving conditions you will run out of fuel in 50 km."
5	Car Manual	"Ask the VICO system about the current driving range without re-fueling, and the average fuel consumption."
6	Hotel Reservation	"You are approaching Frankfurt and decide to stay over night. Use VICO to make a hotel reservation in the center of Frankfurt according to your preferences."
7	Traffic Information	"Due to an accident there is a traffic jam on your route. Should an alternative route be considered?"

Table 1: Experimental tasks of the test subjects

The route to be driven by the test subjects while performing the given tasks had a length of 53.4 km. It consisted of a mixture of highways, passing through two towns and a freeway section. These variations were intended to offer different impressions to the subjects

while driving, keeping the surrounding attractive and also trying to keep them busy to simulate real traffic situations as close as possible.

5.3. Test Procedure

All seven tasks were inserted in the scenario of the driving simulator so that they appeared at the same location within the driven route for all test drivers. The course of each experiment consisted of five parts:

1. First, the test subject had to take a 10-minute test drive to get familiar with the operation of the driving simulator.
2. After this trial run, the capabilities of the VICO system were briefly described by a supervisor. He also demonstrated a short example dialogue covering information on movie theatres, selection and ticket reservation.
3. Subsequently the test subject was left alone, and the actual Wizard-of-Oz experiment started. This included driving the given route while performing the seven given tasks. The driving scenario lasted about 30-35 minutes depending on the respective driver's speed.
4. Following the test drive a short interview was performed by the supervisor and a questionnaire had to be filled out by the test participant.
5. In the very end the true nature of the supposed prototype was revealed to the test driver.

Overall duration of the experiment lasted about one hour for each of the subjects.

The simulated VICO system was used as a speech-only system. User input had to be done by speech, and also VICO's output was presented to the user by voice only. No visual feedback was given on a display. User-initiated dialogues had to be started by activating VICO through uttering the keyword "Vico".

5.4. Example Dialogue

In the following a short example dialogue like the one demonstrated by the supervisor when introducing the features of VICO is presented:

User: VICO!
System: *Yes please.*
User: I'd like to see the movie "What women want" tonight.
System: *In which city would you like to see the movie?*
User: In Stuttgart.
System: *In Stuttgart "What women want" is showed at the movie theatres Europa and Gloria. Which one would you prefer?*
User: That depends on the starting times. I'd like to go at around 8 p.m.
System: *The Gloria starts at 7:45 p.m., the Europa at 8:15 p.m.*
User: Then I'd rather go to the Europa theatre.
System: *Do you want me to make a reservation for you?*
User: Yes, if this is possible. I need two tickets.
System: *One moment please ..... Two tickets will be reserved for you. They have to be picked up until 8 p.m.*

6. Results

To evaluate the monitored dialogues, all user utterances were transcribed using the video recordings. The recordings and transcriptions were then analyzed concerning dialogue duration and structure. Also driving behaviour of the test drivers while talking to VICO was studied. In addition, the subjective judgements of the test participants were examined.

6.1. Dialogue Duration

In general, the duration of a dialogue varied depending on the complexity of the task to be performed by the test subjects. Varying from very simple requests like asking for the time to very complex problems, such as the selection and reservation of a hotel room, dialogue duration ranged from 20 to 200 seconds.

The length of a dialogue was also influenced by certain individuals having clearly longer conversations with VICO than the other participants of the experiment.

6.2. Dialogue Structure

One of the expectations when designing the used Wizard-of Oz scenario had been that the library of pre-defined system prompts would not be sufficient for the various user requests. The possibility of entering a new system response by the wizard had been foreseen because we had anticipated a frequent usage of either new phrases or the "No information is available on this topic." response.

However after analyzing all recorded conversations with VICO, the dialogue structure of the individual tasks showed much less variation than had been expected. In most cases the library of pre-defined system prompts was sufficient for the wizard to answer the users' requests adequately. Very rarely a new answer had to be created or the wizard had to admit that no information on this topic was available.

6.3. Dialogue Evaluation

Although some of the tasks to be performed by the test subjects were very complex, all dialogues could be completed successfully.

Being able to respond to the majority of user input by choosing one of the pre-defined system prompts was due to the low variance in dialogue structure of all user requests as well as to the restricted application domain. Surprisingly the input structure of all individuals was very similar and during the conversation evolving, test subjects accepted some kind of system guidance willingly. However, the kind of information asked by the subjects was quite different. Some of them had clear ideas about the features of a hotel room they would like to reserve whereas others were already satisfied by giving the minimum information that was asked by VICO to fulfill their request. After being familiar with the system to some extent, some users initiated additional tasks without being explicitly prompted by a pre-recorded message. This included simple queries like "What is my current speed?" as well as clarification dialogues even after successful completion of a task. One example was task 4 (being low on gas) when some users asked for a different gas station than the one VICO had offered them.

Test Participants also handled very well mis-recognitions provoked by the wizard. For each of the subjects VICO responded to a user request once by "Ihre Eingabe wurde nicht verstanden. Bitte wiederholen Sie." ("Your input was not understood. Please repeat."). This simulated misrecognition of the system did not confuse the test drivers, but was accepted patiently by repeating the input phrase, sometimes rephrasing it to some extent.

6.4. Driving Behaviour

Not being the main objective of the performed experiments they still allowed to study the effects of distraction on driving behaviour that complex vocal tasks may have. Analyzing the files monitored by the simulator two aspects of driving were especially examined: lane-keeping and speed.

No accident occurred during the test drives. It turned out that exceeding the speed limit appeared much more frequently while not interacting with the system. Obviously, while involved in a conversation with VICO, subjects tended rather to drive slower than too fast. Looking at driving behaviour regarding lane-keeping, the measured error rate showed no considerable change when talking to VICO compared to not entering input to the system and just managing the driving task.

6.5. Unmasking the Wizard

All through performing any kind of Wizard-of-Oz experiments the most critical point is to be uncovered by a test subject. In our case all subjects were asked at the end of the experiment if they had suspected the system to be of true technical nature, or had been convinced to talk to a human being. Only one out of the ten participants claimed having noticed for sure that VICO was in fact simulated by a human being.

6.6. User Ratings

When interviewed by the supervisor after the experiments all test subjects pointed out to be satisfied with the VICO system. They especially liked the easy-to-use speech-controlled interface. Surprisingly, not having a display to also get visual feedback was not criticized as a major disadvantage.

The general conversation flow with VICO was judged to be very comfortable and pleasant. Most of the test subjects claimed not having been too much distracted while using the system. This observation is confirmed by the lack of severe driving errors monitored.

Few problems were reported concerning the interaction with VICO. Some of the users had difficulties understanding system responses by VICO. In two cases this was blamed on noise in the driving simulator environment and insufficient acoustic quality of the system's responses. Two other subjects felt VICO's responses as being overloaded with too much information from time to time, complicating to understand what had actually been said. One person felt the system had spoken too fast.

In general the presence of a virtual co-driver was felt as very pleasant (see also below).

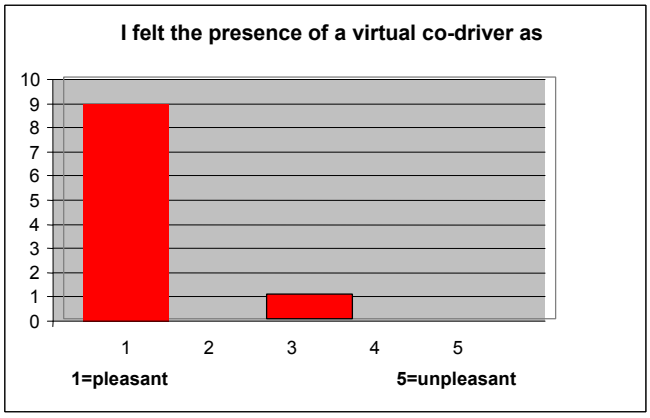

Figure 5: Subjective rating of VICO

7. Implications on System Design

The performed Wizard-of-Oz experiments had been initiated to answer the questions raised in section 3. Thus emphasis had been put on monitoring the dialogue interaction between user and system, user reactions towards the simulated prototype system, expectations of potential users, and distraction effects on driving behaviour.

7.1. Dialogue Interaction

The sample dialogues collected during our test drives provide a first model of interaction between drivers and a virtual co-driver when being allowed to use unrestricted natural language input. These dialogues, together with further research activities conducted into this direction, will serve as basis for a first dialogue model of the VICO prototype system to be built.

7.2. User Reactions and Expectations

Reactions of users to the simulated VICO system were overall very positive. VICO was considered to be easy-to-use, and speech was seen as a comfortable input modality. Test subjects in general liked the way in which they could converse with a driver information system in everyday language without being forced to remember a specific vocabulary or sentence structure. Being one of the main objectives of the VICO project this gives us an indication that the need for natural language interfaces and dialogues is indeed there, and system development has to be directed towards this goal. Special emphasis will have to be put on the development of robust speech technology and natural language understanding techniques supported by adaptive and flexible dialogue strategies.

Beside the conducted Wizard-of-Oz experiments inquiries have been performed regarding the desired functionalities a voice-operated in-car information system should cover. Traffic- and travel-related features were considered important, as these are the functionalities current users are familiar with. The originally envisioned functionalities for the first VICO prototype were strengthened by these inquiries, further focusing on navigation and tourist information issues for the future. Looking at the human-machine interface that should be

offered by an in-car information system, the following characteristics were favoured by potential users: no distraction from the primary driving task should be perceivable; speech input was preferred over a tactile interface; the usage of everyday language was favoured over a fixed command language. As speech input is one of the main objectives of the VICO project, these user expectations will be considered in the future development process.

7.3. Distraction Effects

Especially in the context of driver information systems, the issue of safety is absolutely mandatory and has to be considered. When operating in-car devices and services distraction effects have to be minimized in order to guarantee that driving behaviour is not influenced. The encountered speed errors resulting in low speed are typically countermeasures when the mental workload gets too high.

Special care has to be taken in the design of the first prototype to keep the mental load of the driver as low as possible while simultaneously enabling him to operate complex functionalities of in-car systems.

7.4. Conclusions

In summary, the performed Wizard-of-Oz experiments can be considered to have been completed successfully. When designing the experiments it had not been clear if a limited set of pre-defined answers would be sufficient to enable a natural language dialogue. The variety of system prompts needed to manage the situation throughout the experiment was very low. This was due to the restricted nature of the used domain, and a certain dialogue guidance triggering the user to react in a specific way. The fact that even with this relatively simple set of responses the wizard was able to keep up a natural conversation with the test subjects proves the feasibility of a true technical solution concerning dialogue strategies even in the complex context of a spoken language dialogue system.

The subjective satisfaction of the test participants can also be seen as a positive outcome of the experiments. A broad functionality, its ease of use and little distraction effects were considered as the main benefit of the system. The level of surprise when being told that in fact they had been talking to a human being instead of a virtual co-driver indicates that the Wizard-of-Oz methodology is an effective means to be applied early in the system design process.

Also, before conducting the presented experiments it was not clear to which extent potential users were willing to accept natural language dialogues. Our results have shown that natural language is clearly preferred over command-and-control input. As the sequence and structure of typical natural language dialogues for in-car applications is unknown, the performed tests allowed to gain a first insight on how a typical dialogue taking place in a spoken dialogue system might look like.
Subsequent research in simulated and real environments will be conducted throughout the VICO project to ensure a user-oriented system design and development.

8. Summary

Knowledge about typical natural language dialogues in the context of driver information systems is of great benefit for the development of a spoken language dialogue system for in-car applications. Also, information about the expectations of potential users of a VICO-like system as well as investigations on user acceptance are needed to guarantee a user-oriented design and development process. The performed Wizard-of-Oz experiments have provided this information crucial for a successful and efficient system design process.

Without imposing any restriction on the test subjects concerning vocabulary, sentence or even dialogue structure to be used during the conversations with VICO, typical dialogues could be identified. An insight on the expectations of potential users towards a vocal-driven natural language human-machine interface and its inherent dialogue structure could be gained. Feasibility of a VICO system has been demonstrated, very positively rated by the test subjects. Based on the knowledge acquired in the performed Wizard-of-Oz experiments a first prototype VICO system will be realized. Following the multi-cycle design process the resulting initial prototype will then be further evaluated both by experts and test subjects.

9. Acknowledgements

The authors gratefully acknowledge all members of the VICO team that contributed to performing these experiments. Special thanks to the Human Factors Team at Robert Bosch GmbH, who facilitated the linkage with the driving simulator, supervised and analyzed the experiments. This work was partly funded by the European Commission in the framework of the Information Society Technologies (IST) Programme project reference IST-2000-25426 (VICO). The views and conclusions contained in this document are those of the authors.

10. References

N. Dahlbäck, A. Jönsson and L. Ahrenberg (1993). Wizard of Oz Studies – Why and How. *Knowledge-Based Systems*, Vol. 6, No. 4, pp. 258-266, December 1993.

P. Geutner, L. Arévalo and J. Breuninger (2000). VODIS – Voice-Operated Driver Information Systems: A Usability Study on Advanced Speech Technologies for Car Environments. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 378-381, Beijing, China, October 2000.

D. Manstetten, W. Krautter, B. Grothkopp, F. Steffens and P. Geutner (2001). Using a Driving Simulator to Perform a Wizard-of-Oz Experiment on Speech-Controlled Driver Information Systems. In *Proceedings of the 1st Human-Centered Transportation Simulation Conference (HCTSC 2001)*, Iowa City, Iowa, November 2001.