Multimodal Interaction in the Car - Combining Speech and Gestures on the Steering Wheel

Bastian Pfleging, Stefan Schneegass, Albrecht Schmidt
Institute for Visualization and Interactive Systems (VIS)
University of Stuttgart
Pfaffenwaldring 5a, 70569 Stuttgart, Germany
{firstname.lastname}@vis.uni-stuttgart.de

ABSTRACT

Implementing controls in the car becomes a major challenge: The use of simple physical buttons does not scale to the increased number of assistive, comfort, and infotainment functions. Current solutions include hierarchical menus and multi-functional control devices, which increase complexity and visual demand. Another option is speech control, which is not widely accepted, as it does not support visibility of actions, fine-grained feedback, and easy undo of actions. Our approach combines speech and gestures. By using speech for identification of functions, we exploit the visibility of objects in the car (e.g., mirror) and simple access to a wide range of functions equaling a very broad menu. Using gestures for manipulation (e.g., left/right), we provide fine-grained control with immediate feedback and easy undo of actions. In a user-centered process, we determined a set of user-defined gestures as well as common voice commands. For a prototype, we linked this to a car interior and driving simulator. In a study with 16 participants, we explored the impact of this form of multimodal interaction on the driving performance against a baseline using physical buttons. The results indicate that the use of speech and gesture is slower than using buttons but results in a similar driving performance. Users comment in a DALI questionnaire that the visual demand is lower when using speech and gestures.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g. HCI)**]: User Interfaces

Keywords

Automotive user interfaces; multimodal interfaces; gesture interaction; speech interaction.

1. INTRODUCTION

Modern cars offer a large number of information and entertainment functions. In addition, controls are required for operating comfort and assistance functions. Functions that have been added over the last decade include navigation functions and access to location-based information (e.g., next gas station), interaction with large music collections (e.g., 1000s of songs on a MP3 player), control of adaptive cruise control, and activation of semi-autonomous parking. Many of these functions do not relate directly to the primary driving task, but are related to secondary and tertiary driving tasks [7]. Nevertheless, users will need (or want) to operate these devices in many cases while driving. In addition to interaction with controls in the car, we see an increase of interaction with mobile devices such as smart phones—either through proxy controls in the car or by directly using these devices. Although the usage of mobile devices (e.g., for messaging, social networking, or Internet access) is predominantly prohibited while driving, we see a desire to use them on the road. This trend will even increase with the rise of semi-autonomous driving modes (e.g., lane keeping, adaptive cruise control) as the cognitive load for driving is reduced. Especially commuters aim at utilizing the time to and from work for (social) interaction and communication.

Figure 1: The study setup with multitouch steering wheel and speech input in the simulator.

With each added function, assistance system, or infotainment system the car cockpit becomes more complex. Buttons and manual controls still play the most important role in the automotive design space [12]. However, given the large number of functions, this traditional approach does not scale as the space within the driver's reach is limited. It is apparent that with the large number of functions a one-to-one mapping between manual controls and functions is not possible any more. Engineers and UI designers have to decide which controls to make directly accessible via a physical button and which functions will require a more complex interaction

AutomotiveUI'12, October 17-19, Portsmouth, NH, USA.

procedure. Current infotainment systems therefore often rely on hierarchical menus to provide access to all functions. Only the essential or favorite ones get a button. Accessing the menu-based interface typically includes navigation using touch screens, turn-and-push controllers, or touch pads. Drawbacks for functions not accessible by buttons are that the controls are hidden and the time to execute a certain function is much longer than with a physical button. Additionally, all of these systems require visual attention (e.g., reading the menu). With our research we aim to offer alternatives for controlling the functions with minimal driver distraction.

Assessing the design space beyond menus, knobs, and buttons, the following potential input modalities present alternatives for in-car interaction: (1) speech interaction, (2) free 3D gestures, (3) touch gestures, and (4) gaze interaction.

Speech interaction and voice control are widely implemented for selected functions (e.g., to input a destination into a navigation system or to initiate a phone call). Nevertheless, so far only a minority of drivers regularly uses speech input due to various reasons, including the effort for learning and remembering commands [18]. With an increasing number of functions this problem becomes more important as the hurdle for taking up speech as a modality is even increased. Using gestures poses a similar problem, as here, too, users have to remember the commands they can execute by gesture and the appropriate gesture. Visual representations on a screen can also be manipulated by touch, not requiring to remember commands but visual attention.

Minimizing visual distraction and reducing drivers' workload are central design goals. In our research we suggest to revisit the idea of multimodal interaction as it provides a great potential benefit over systems operating on a single modality. Since the visionary work of Bolt [3], different research projects have investigated on multimodality and general guidelines have been shaped (e.g., Reeves et al. [19]). However, so far no specific usage pattern or interaction style for an integration of different modalities has been widely adopted in the car.

In this paper we propose a multimodal interaction style that combines speech and gesture in the following way: voice commands are used to select visible objects (mirror, window, etc.) or functions; simple touch gestures are used to control these functions. With this approach recalling voice commands is simpler as the users see what they need to say. Using a simple touch gesture, the interaction style lowers the visual demand and provides at the same time immediate feedback and easy means for undoing actions. To design the system, we first conducted a formative study on user-elicited speech and gesture commands to inform our design. Based on these results, we implemented a functional prototype that allows evaluating the suggested interaction style.

The contributions of this paper are (1) a set of basic gestures and voice commands for in-car interaction, (2) an investigation for which functions multimodal interaction using speech and gestures is appropriate, (3) a description of our prototypical implementation, and (4) an evaluation of the proposed interaction style in a driving simulator.

2. RELATED WORK

Research in human computer interaction in the automotive context has grown in the last years, and finding enabling interaction that is at the same time pleasant and minimally distractive is a common goal. A major challenge is to combine means for interaction for primary, secondary, and tertiary driving tasks [7]. With advances in (semi-)autonomous driving, more primary driving activities (e.g., steering and lane keeping) are assisted by computers. Based on the general layout of a car and of the driver's interaction area Kern et al. proposed a design space for in-car user interfaces (UIs) [12]. This design space can help to identify potential overload areas, reason about driver distraction, and help to assess trends in automotive UIs. Our assessment showed that there is a trend towards adding interaction elements onto the steering wheel. The rational is simple as controls added to the steering wheel are in easy reach for the driver and do not require the users to take their hands off the wheel. However, looking at the resulting design space, it is apparent that this trend is limited by the number of buttons and controls that can be added into this area. For us this motivated the use of the steering wheel as input space, but in contrast to the recent trend of adding buttons, we chose to explore touch as one modality.

Using touchscreens for menu interaction and for entering text are typically application that are found in cars. Touchpads [8] are another option, they are commonly used as remote input device or to write text. Spies et al. investigated in [22] a haptic touchpad as a mean for controlling in-car UIs. In this approach visual attention is required. Döring et al. used gestures of a multi-touch steering wheel [5] for a gesture based interaction style with different applications, such as navigation or a music player. In their work a gesture set was created in a user-centered design process. The comparison of the gesture set with classical means for interacting with an infotainment system showed that using gestures reduces the visual demand for interaction tasks. However, the use of gestures introduces a similar problem as buttons: scalability. By using gestures that do not need visual attention, the gesture rapidly becomes complex and hard to remember. By using touch interaction that relates to the displayed content on the screen, the benefit of reduced visual attention is lost. This motivated us to investigate gestures further, as one modality in a multimodal UI.

In order to lower workload and driver distraction, different input modalities are being evaluated. Gaze and body posture are two examples of implicit modalities that can be used to provide more natural forms of interaction that have the potential to reduce cognitive load. Gaze interaction was explored by Kern et al. [11]. Here, the last fixation of the user before switching attention from the in-car screen to the real world was detected. This fixation was then used to highlight an area on the screen and by these means ease the attention switch from the road to the display. We expect that in a future version of our system a similar concept could be used to detect which items people look at and use this as further modality to disambiguate commands. Similarly, body posture (e.g., head position) could be used to detect the objects towards which the user's commands are aimed at.

Voice input has been investigated for in-car interaction for years, and many efforts focus on the improvement of recognition accuracy of speech input [23]. Nevertheless, voice interaction is still not widely accepted in the automotive domain [18]. Beside some remaining technical difficulties, the lack of conceptual clarity is another problem. This topic is addressed by the use of a natural voice interfaces, as discussed

in [1], however, this approach has its limitation with regard to immediate feedback and visibility of commands. The perceived user experience (UX) is another crucial aspect, in particular for speech UIs. This issue is investigated for in-car speech input in [9]. Based on these findings we designed a multimodal approach combining gesture and speech with a focus on UX.

Multimodal systems are defined by Oviatt [16] as "those that process two or more combined user input modes — such as speech, pen, touch, manual gestures, gaze, and head and body movements — in a coordinated manner with multimedia system output". Multimodality can be seen as offering alternative channels (e.g., an action can be accomplished by using either of the available modalities) or as interaction using two or more modalities at the same time. Müller and Weinberg [15] make a more sophisticated distinction of multimodality in the car and describe three methods for combining different modalities: fused modalities, temporally cascaded modalities, and redundant modalities. An example for fused modalities is given by the "put-that-there" approach by Bolt [3]: pointing gestures were used accompanied with speech input containing deictic references like "that" or "there". This idea has been particularly applied for 2D map interactions (e.g., [21]). Pointing at real objects in the 3D space of a car is, however, more difficult and several functions cannot be associated with a physical location in the car. To avoid these problems, we swapped the modalities of speech and gesture. We follow the approach of temporally cascaded modalities: first, using speech to select a real object and one of its functions and, second, offering touch gestures to specify parameters. A similar approach combining speech and other modalities in cars has been made in industry and in research [6, 15] where a concept of combining freehand gestures and speech input for making phone calls was presented. In contrast, our approach aims at a general interaction style that covers many functions and goes well beyond a single application control.

3. CONCEPT

The major design goal of this user interface is to ensure good usability in the context of usage in the car. Adapting common usability guidelines, traditionally stated for desktop systems, we identified several drawbacks of current unimodal interactive systems and user interface designs. Our idea is to create a multimodal interaction style that addresses these shortcomings.

3.1 Challenges of current solutions

Learnability. Current implementations of command-based speech interaction techniques require the user to learn and remember commands in order to achieve a satisfactory user experience. Natural voice user interfaces (e.g., Dragon Drive[1]) have been created to tackle the problem of remembering by allowing a wide range of natural commands. However, the driver needs to knows the capabilities of the car that can be controlled by voice. In case of ambiguity such systems require an additional clarification of commands [1], which makes the interaction more cumbersome. Touch interaction or gestures as a single modality have a similar disadvantage. They require the user to remember a potentially large number of gestures (e.g., one for each command) that can be differentiated or they increase the complexity of gesture sequences. Regarding the expressiveness of gestures, studies have shown that it is hard to remember a large number of commands in the form of complex gestures. Such a large number needs to be covered in environments with a multitude of functionalities such as infotainment and entertainment systems in cars.

Design Goal 1: *Minimize the effort for learning and remembering.*

Visibility. Current speech interfaces as well as gestural interface do not offer visibility of command options, like menus do. A good interface design is created that it respects the *visibility principle* and offers means for the users to visually perceive choices. Users should see the options they have to do a task, but the design should reduce distraction by hiding *unnecessary alternatives*. For speech interfaces, it is difficult to serve this principle by means of visualizing the available interaction grammar or commands. Additionally, providing meaningful feedback for every operator action is *time-consuming* and maybe even annoying - especially if the feedback should be non-visual in order to keep a low visual distraction. It is similarly difficult for gestural interaction to provide visibility of interaction possibilities as for speech.

Design Goal 2: *Create an interaction style that maximizes the visibility of command options in the car.*

Granularity. Using sliders or gestures, interaction can be very fine-grained. For speech interaction, the granularity of a single interaction is low. As speech commands take a certain while to be spoken (in general longer than a button press or a simple finger movement) the granularity of the provided interaction primitives is usually designed bigger in order to not increase the overall interaction time. Although combining basic commands like "move window up" and "stop window" is possible, it is questionable if a precise window control can be realized due to delays between the user saying a command, parsing the command by the system, and perceiving a system response (i.e., window movement).

Design Goal 3: *Provide fine-grained opportunities for interaction.*

Undo. Modern UIs in the desktop domain have massively benefited from means to revert/undo actions taken. This helps users to explore systems without too much worrying that something goes wrong. Similarly, in the car an easy undo of actions is essential as even small errors may be distracting and time-consuming to be corrected, e.g., by giving a fully formed sentence. A potential command might be: "Close the driver's window by 80 %.". If the driver notices after the execution that the window had been moved too far, another command has to be said to partially or completely undo the last action and to achieve the goal.

Design Goal 4: *Support means for simple partially or completely undo of actions.*

These 4 design goals are hard to realize by using current speech interaction techniques. Simple (single stroke) touch and pointing gestures are well suited to realize design goal 1, 3, and 4. For example, they allow a fine-grained granularity of interaction and provide means to easily undo an action (e.g., by doing the reverse gesture/movement).

3.2 Interaction Style

To address the described issues and to implement the design goals, we propose a new multimodal interaction style

[1]Nuance - `http://www.nuance.com/products/dragon-drive/index.htm`

that combines speech and gesture interaction (Figure 2): In a first step, speech, which reliability is increasingly improved, is used to select and qualify one or multiple objects (e.g., "window") and their function (e.g., "open") to be manipulated. If an object offers only one function that can be manipulated, the selection process can be as short as just saying the name of this object and implicitly choosing its function, e.g., "cruise control". If multiple instances of an object exist (e.g., windows) the desired objects need to be qualified (e.g., "passenger window", "backseat windows"). The interaction can also be started by just saying the object's name ("window"). If the selection is ambiguous, the system will ask for a suitable quantification until the object and function selections are well-defined. Objects that offer more than one function require the user to also clearly select the function. The *disambiguation cycle* (see Figure 2) assures an explicit selection of object(s) and function by providing speech prompts to refine the selection. Depending on the context, the speech prompt could also be combined with a visual presentation of options (e.g., on the head-up display) to qualify the object or function. The disambiguation could also be supported by observing the user's gaze (e.g., voice command "mirror" & gaze to the left to select the left mirror). As the objects are visible in the corresponding environment, it is easy to remember the items of the interaction space and to comply with the visibility principle. Using single words as starting point can also help to increase the users' willingness to explore. With this approach a large amount of items and functions can be addressed without an increased memory load on the users' side.

After the selection of the interaction object(s) and function, the user can perform a gesture to complete the intended action. This form of interaction (e.g. moving a finger up and down on a touchpad) allows for a fine-grained manipulation and provides simple means for undo of an action. As the action is executed at the same time, immediate feedback is given by means of manipulating the objects, e.g., the mirror changes its orientation as the user moves the finger over the touchpad.

Overall, speech allows selecting functions and objects by just naming them (including a range of synonyms) without a need for hierarchical structures or explicit navigation. Touch gestures support fine-grained control of functions and easy undo/redo means. In the context of the car, previous research has shown that gestures are powerful as they require minimal attention and can be performed without taking the eyes off the road [5], whereas interaction with (graphical) menus and lists is visually much more demanding and results in a higher distraction. Finding intuitive gesture commands to manipulate functions can be difficult and hence particular care has been taken to find appropriate gestures. Our developed multimodal interaction style adheres to all goals stated above. By separating selection of object or function from the manipulation of the function, the same touch gesture can be reused for several actions that are distinguishable by their speech invocation (1:n mapping from gestures to functions) and hence gestures remain simple.

4. FORMATIVE STUDY

To explore the combined use of gesture and speech in the specific context of the car, we conducted a first study to investigate user-defined voice commands and gestures. In this study, we wanted to address two research questions: (1) "How do users name or address the objects and functions they need to control without prior training?" and (2) "Which gestures do users perform in order to control a function on a selected object?" We performed the study following methods on user-defined gesture sets [24] to extract and redact user-defined gestures as well as speech commands. We expect that the presented interaction style will reduce the visual demand during interaction. Furthermore, such an approach could potentially be applicable beyond the car for all settings where the functions and objects to control are visible (e.g., smart environments) and where fine-grained control and undo/redo are important.

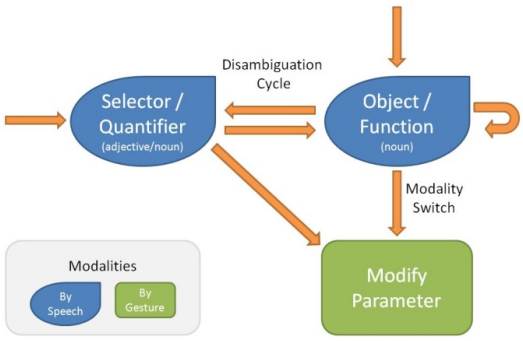

Figure 2: Diagram of the speech-gesture interaction style - Interaction objects are selected by speech and then manipulated by (touch) gestures.

4.1 Study Design and Setup

For the user study, we chose a scenario of controlling 26 secondary and tertiary functions in a car (i.e., safety and comfort functions). All selected functions are simple one-step functions, mapping a single command onto a single interaction object. The selection of tasks also concentrated around well-know functions that should be common for the average driver as well as they should somehow be visible to the driver. To gather the set of functions, we consulted the manuals of several car models in order to take into account the most common features. For each function, the driver was first presented one or two images to identify the object to interact with and was asked to produce a voice command for this object. Next, the participant was asked by a pre-recorded instruction to perform a gesture on the steering wheel to manipulate the object shown before.

We conducted the study in a lab environment. The setup included a speech-enabled multitouch steering wheel that was also used to control a PC-based driving simulation [13] presented on a 24" screen to provide a more realistic driving scenario. Besides the necessary images to identify voice commands, no visual or auditory feedback was given as a response to voice commands or gestures. An Android-based tablet was integrated into a modified Logitech G27 steering wheel to enable multitouch input. A voice and gesture recording app was developed to present the different scenarios on the tablet. While waiting for gesture input, only a white background was shown without displaying any gesture trace.

4.2 Participants and Procedure

We recruited study participants through institution mailing lists. They were required to have a driver's license. In total 12 people participated (2 female) aged between 20 and

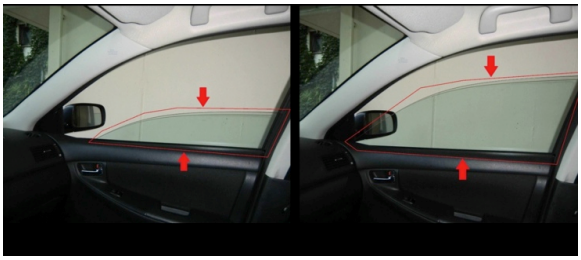

Figure 3: Example instruction image for one task (close the passenger window by 1/3). First, the driver had to identify and name the object that has been modified. Next, a gesture should be performed to achieve the action shown in the images.

39 years (AVG 28.2 years, SD 6.58 years). The participants had an average driving experience of 10.6 years (SD 7.3) and car usage ranged from once a month to every day. Five participants owned a car and 10 were right-handed. Four of the participants already had experiences with speech input/interaction. Nine of them already used a touch-enabled device (mainly phones/tablets) before.

After a short introduction to the research context and a demographics questionnaire, participants were seated in front of the simulation environment. We chose the simulator in order to simulate a realistic environment. The driver had to drive along a 2-lane infinite highway where blocking obstacle indicated necessary lane changes. A fixed speed was pre-programmed and the drivers were instructed to keep at least one hand on the steering wheel. They should only avoid obstacles while performing a gesture. Further, they were told to use either one or multiple fingers to do a gesture.

The main part of the study consisted of 26 tasks (see Table 1), which were presented to each participant in a permuted order to avoid learning effects. Each task consisted of three parts. (1) The participants were presented one or two augmented photographs on the steering wheel screen showing the initial state and the final state of an object in the car (see Figure 3). We asked the participants to verbally address the object/function and spontaneously provide a unique name. In order to not prime the use of a certain wording we gave no textual or verbal instruction. If addressing was not precise enough to specify the object, the experimenter asked to refine the command (e.g. user: "seat"; experimenter: "which seat?"). (2) Once the participant addressed object and function, we asked to suggest a gesture she would use to specify the parameters (e.g., moving the window on the driver side half way down). This instruction was given as a voice instruction by the tablet. (3) The participants were asked to rate the difficulty of conducting the command and gesture. No auditive or visual feedback was given for both voice commands and touch gestures. The three steps were repeated for each of the tasks. They were controlled by the experimenter in order to prevent advancing before a suitable voice command/gesture was presented. There was no time limit given for any of the tasks. At the very beginning, one additional task was presented to explain the procedure. During all steps, the participants were asked to think aloud in order to be able to recognize their mental processes afterwards.

At the end of the study, users filled out a questionnaire on the acceptance of this type of interaction. Participants were encouraged to provide feedback and to suggest improvements in an interview.

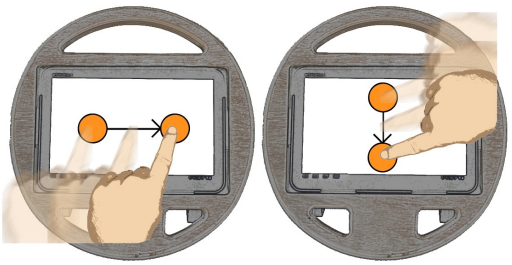

Figure 4: Examples of directional gestures used in our study.

4.3 Results

Users' Evaluation of the Concept. Participants rated the overall acceptance of a system that uses the presented multimodal approach in average 3 on a 5-point Likert scale ranging from 1 (poor acceptance) to 5 (high acceptance). For each of the task categories, we asked the participants to rate on a 5-point Likert scale from 1 ("don't agree at all") to 5 ("fully agree") whether they assume that the proposed interaction is suitable for this category. In average, the participant rated the external mirrors (AVG 4.08, SD 1.11) and the seat heating (AVG 4.00, SD 1.08) highest and the wipers worst - slightly below undecided (AVG 2.92, SD 1.26).

Difficulty of Voice Commands & Gestures. Of all tasks executed, 30.1 % of the given commands and gestures were rated "very easy", 46.5 % were graded "easy" and 15.7 % "medium". In 6.1 % of the presented cases, users rated gestures as "difficult", leaving 1.3 % of commands and gestures that were "very difficult" (one task was not rated, 0,3%).

Voice Commands. In 305 of the 312 tasks (97.8 %), people were able to find appropriate terms/voice commands for the objects and/or their functions presented during the first part of each task. The most difficult object was the head-up display where seven participants did not succeed in finding an appropriate term. As the images already suggested the intended action by showing the situation before and after executing the task, the participants could have chosen to directly name the function of the object (e.g., "move driver's seat"). Nevertheless, only a minority chose this option (16.1 %). Most users named the objects themselves (e.g., "driver's seat", 82.1 %). The evaluation of the voice command showed further that the participants used a variety of terms for the same object (e.g., right mirror, right exterior mirror, mirror on the passenger side, exterior mirror on the passenger side, adjust exterior mirror right . . .). A similar variation of commands could be noted throughout all tasks. Even though this variation occurred, the object (and partially its function) could be identified accordingly. As a conclusion, we think that the denotations of visible objects have potential for intuitive voice commands.

Touch gestures. Looking at the recorded touch gestures, the study reveals a high agreement on gesture commands among participants. Overall, the participants did not have problems to think of touch gestures and chose very similar and simple gestures to control most of the functions. For 309 of the 312 tasks, the participants were able to produce a meaningful touch gesture. For 78.1 % of these 309 gestures the participant used only one finger, 12.9 % resp. 6.8 % of gestures were done with two or three fingers. In 1.6 % of the cases, 4 fingers were used and five fingers were used only once. One third of the gestures was performed using the left hand, both hands were involved only twice.

Table 1: Collection of the different tasks and the frequency of directional gestures for each scenario. Other gestures are summarized as '*'.

#	Object	Function	↑	↓	←	→	*
1	driver seat	move forward	8	-	4	-	-
2		move back	-	7	-	5	-
3		move up	12	-	-	-	-
4		move down	-	11	-	-	1
5		headrest up	11	-	-	-	1
6		headrest down	-	12	-	-	-
7		backrest fwd.	8	-	4	-	-
8		backrest back	-	8	-	4	-
9		inc. heating	12	-	-	-	-
10		dec. heating	-	12	-	-	-
11	right ext. mirror	move left	-	-	12	-	-
12		move up	11	1	-	-	-
13	front wiper	off	-	4	4	1	3
14		permanent	2	-	-	4	6
15		intermittant	1	-	1	4	5
16	rear wiper	clean+wipe	1	-	-	2	7
17	passenger window	open partially	-	8		-	4
18		close partially	8	-	-	-	4
19	backseat window (l)	auto open	-	11	-	-	1
20		auto close	12	-	-	-	-
21	cruise control	accelerate	-	11	-	-	1
22		decelerate	11	-	-	-	1
23	left air vent	more air	11	-	-	1	-
24		less air	-	11	1	-	-
25	head-up display	inc. brightness	10	-	-	1	1
26		dec. brightness	-	11	-	-	1

If we consider the style of the 309 recorded meaningful gestures, 86.7 % of them were easy directional drawing gestures. As shown in Figure 4, the main direction of such a gesture was either up/down (37.9 % resp. 34.3 %) or left/right (8.4 % resp. 6.1 %). These gestures were conducted with one or multiple fingers and the participants drew their gestures either as a straight or slightly curved line. For a real-world implementation, these directional gestures also allow an easy undo feature: To undo an action, just the direction of the drawing gestures has to be inverted. Moving the finger(s) to more than one direction (e.g., "zoom gestures", moving left and right as one gesture) occurred for 7.8 % of the performed gestures. Another 5.5 % of the gestures were conducted without a certain direction (e.g., circular gesture, single tap).

For six of the presented tasks (23.0 %, see Table 1), all 12 participants performed the same gestures. Similarly, for 8 tasks (30.8 %) all but one participant made the same gesture as did 10 of them for two other tasks (7.7 %). The performed gestures were either drawing one or multiple fingers up or down. These gestures were used to move an item upwards, increase a value, or do the opposite action. The only exception was task 11 where a drawing to the left was noted to move the exterior mirror to the left. For the task of moving seat (#1, # 2) or backrest (#7, #8) to the front or back, still 8 respectively 7 participants performed the same up-/down-pointing gestures. The rest of the participants used drawing gestures pointing to the left or to the right. This might have been caused by the visual representation of the seat as the front was at the left. Additionally, almost all participants stated that they tried to create gestures consistently.

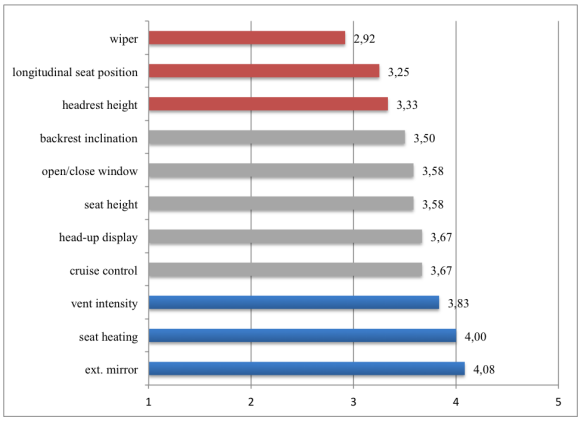

Figure 5: Participants' ratings of the suitability of our interaction style for different tasks.

Overall, the drawing gestures are based on embodied conceptual metaphors and seem well suited to control the parameters of most objects' functions.

5. PROTOTYPE

We designed and implemented a prototype that is used to evaluate the proposed multimodal interaction approach. The prototype is integrated into our driving simulator and car function simulation setup. One part of the system is an android application[2], running on a tablet PC attached to the center of the steering wheel. It allows the user to enter multitouch gestures on its entire interactive surface, which are then sent to the main interaction framework. We implement the voice recognition using a consumer microphone and the Microsoft Speech SDK on a Windows 7 computer. It uses a simple grammar to understand users' speech commands and implements the *disambiguation cycle* (see Figure 2). Our architecture is built to allow addressing further input modalities such as gaze or posture input. The whole communication is based on the EIToolkit developed by Holleis and Schmidt [10]. This toolkit allows a loosely coupled architecture as shown in Figure 6 by utilizing UDP broadcast.

The formative study provided the basis for the speech commands and gestures used in our implementation. For speech commands, we included variations for each object. The input data is processed in the interaction framework. The generated output includes the information about the manipulated object and the action that takes place. This output (e.g. a moving mirror or opening window) can be processed in the car or be interpreted and visualized by the car environment simulator.

For our second study, the prototype is connected to a driving simulator and to a car interior simulator. This simulator itself consists of five displays (one for the driving scene, four to simulate the cockpit) arranged around the driver's seat to simulate the driving task (see Figure 1): A 55" LCD TV in front of the driver shows the road based on the output of a driving simulation. Two screens are arranged on the left and right of the driver showing a recorded view of the driver/passenger window visualizing the mirror and the window itself. Mirrors and windows are interactive and change their position depending on the driver's input. Another display visualizes the rear window showing an interactive

[2]Google Code: TUIOdroid - `http://code.google.com/p/tuiodroid`

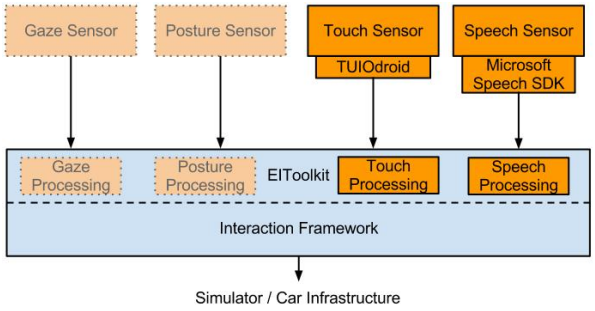

Figure 6: Architecture for the proposed multimodal automotive interaction interaction style.

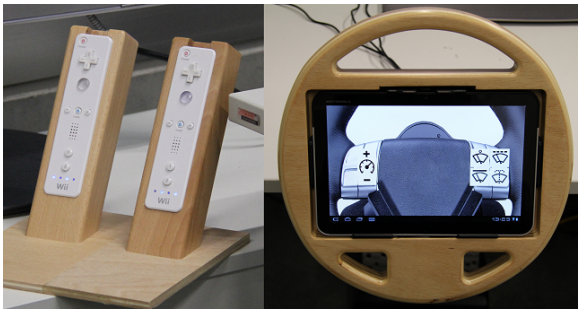

Figure 7: Controls used to approximate a traditional automotive interface.

wiper. At last, a Netbook screen shows speed and heating information in front of the driver, similar to a common dashboard.

6. EVALUATION

In a second study we investigate the possible distraction of the proposed system compared to a traditional setup which is known from current cars. We focus on primitive, one-action tasks neither requiring a long interaction time nor the use of more than a single input element in the traditional setting (e.g., opening the window). Basically, this interaction does not need long practicing from participants and is easy to understand [20].

Another reason for choosing these simple actions is that it seems obvious that functions hidden somewhere in the menu need more time to execute and result in a long interaction time leading to high visual distraction due to the attention shift as described in [20]. During the study, participants only have a short period of time for practicing each interaction technique. Confronting them with an actual in-car menu would take significantly more time to master and some function would maybe not be found at all.

6.1 Study Setup

We used two different conditions used during the user study: the traditional and the multimodal setup, both integrated into our driving simulator environment. The traditional setup consists of digital and tangible input elements (see Figure 7) arranged in the style of current cars. We used two WiiMote controls to simulate the window and mirror buttons whereas the speed limiter and wiper controls were simulated as buttons on the tablet PC attached to the steering wheel. These are similar to buttons found on current steering wheels. This setup simulates a current car environment. As driving simulation, we used the Lane Changed Task (LCT) [14], which is a standardized method to compare driving performance for different conditions. The LCT measures the average distance to baseline of each ride.

6.2 Participants and Procedure

In total, 16 participants took part in the user study (all male; 21 to 29 years old; AVG 23.8 years; SD 2.35 years) which all possessed a valid driver's license. Each participant was an experienced driver with a driving experience ranging from 3 to 10 years (7 years on average).

The study used a within subject design and, therefore, all participants drove four laps, namely: baseline 1, multimodal interaction, current interaction, and baseline 2. Both interaction approaches were alternately changed (randomized over the participants). At first, we asked each participant to set up the seat to reach all controls as good as possible. Afterwards, we introduced the study purpose to them. The first lap is a baseline lap where each participant drives along the track without any secondary task. On the second and third lap, each participant performed as many tasks as possible while driving the whole LCT track either with the multimodal or the baseline interaction approach. Before each participant performs the actual lap, we introduced them to the interaction approach and encouraged them to practice at least once with each object. After each condition they filled in a System Usability Scale (SUS) [4] and Driving Activity Load Index (DALI) [17] questionnaire. In the end, they performed a second baseline lap.

6.3 Results

Analyzing the mean distance to baseline from both interaction approaches, no significant difference is found. The mean distance to baseline is an indicator for how much the secondary task is influencing the driving performance. The multimodal interaction approach caused a slightly higher average distance (1.80 m to 1.74 m), however, significance cannot be reasoned, $t(15) = -1.03$; $p = .32$, $r = .26$.

In addition to the quantitative data measured, we asked the participants to give feedback after using an interaction technique. We measured the feedback with SUS and DALI questionnaires to extract perceived usability and perceived task load. The SUS results show that the participants rate the multimodal approach lower (SUS score 69.06) than the baseline approach (SUS score 79.38). However, it is still rated good [2] and the lower score is likely to be related to the frustration of users with the speech recognition. The task load analysis shows similar results: the multimodal approach receives a slightly higher value than the baseline (17 to 15), indicating a higher workload. Investigating the DALI more detailed, it is important to mention that the visual demand is higher for the traditional approach than for the multimodal approach, which is an important finding because the lesser the visual demand the more time drivers' have their eyes on the road.

Our study indicates that multimodal interaction does perform similar to nowadays widely used interaction objects for one-action interaction. The fact that all participants are used to the baseline approach enhance our findings.

7. CONCLUSION

Inspired by the design principles for interactive systems: learnability, visibility, granularity, and easy undo, we implemented a new multimodal interaction style for the automotive

domain. In this interaction style we combined simple speech commands and minimal touch input on the steering wheel. A further benefit of this approach is that the visual demand is lower than in menu based systems. We show that the overall distraction of this multimodal interaction is comparable to current interaction approach but offers greater flexibility. A further opportunity that arises from this interaction style is that users are offered a simple starting point for voice interaction in the car.

We plan further studies where we investigate how well this approach scales to abstract objects, such as radio stations, in the automotive context, which do not have a visible physical manifestation. Naming these objects could be difficult, as the users may require learning and remembering them. Additionally, we are working on extending the number of interaction objects to more complex parts of the infotainment system such as the navigation system. It should benefit from the new input methodology in comparison to other interaction methods used today. For instance, searching for a point of interest on a map can be cumbersome. We think that a combination of speech and gesture may improve the interaction. In future work we plan to extend the approach with further modalities, in particular gaze and body posture. We would like to investigate how they can be used to disambiguate interaction objects in the car and expect that this will ease the overall interaction process further. Integrating an eye-tracker into the setup also allows to analyze the driver's gaze behavior in detail providing means to further assess the visual demand of the proposed system.

8. REFERENCES

[1] I. Alvarez, A. Martin, J. Dunbar, J. Taiber, D.-M. Wilson, and J. E. Gilbert. Designing driver-centric natural voice user interfaces. In *Adj. Proc. AutomotiveUI '11*, pages 42–49. ACM, 2011.

[2] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, May 2009.

[3] R. A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *Proc. SIGGRAPH '80*, pages 262–270, 1980.

[4] J. Brooke. Sus: A quick and dirty usability scale. In *Usability evaluation in industry*. Taylor and Francis, London, 1996.

[5] T. Döring, D. Kern, P. Marshall, M. Pfeiffer, J. Schöning, V. Gruhn, and A. Schmidt. Gestural interaction on the steering wheel: reducing the visual demand. In *Proc. CHI '11*, pages 483–492. ACM, 2011.

[6] C. Endres, T. Schwartz, and C. A. Müller. Geremin": 2D microgestures for drivers based on electric field sensing. In *Proc. IUI '11*, pages 327–330. ACM, 2011.

[7] G. Geiser. Man Machine Interaction in Vehicles. *ATZ*, 87:74–77, 1985.

[8] I. E. González, J. O. Wobbrock, D. H. Chau, A. Faulring, and B. A. Myers. Eyes on the road, hands on the wheel: thumb-based interaction techniques for input on steering wheels. In *Proc. GI '07*, pages 95–102. ACM, 2007.

[9] A. Goulati and D. Szostak. User experience in speech recognition of navigation devices: an assessment. In *Proc. MobileHCI '11*, pages 517–520. ACM, 2011.

[10] P. Holleis and A. Schmidt. Makeit: Integrate user interaction times in the design process of mobile applications. In *Pervasive Computing*, volume 5013, pages 56–74. Springer Berlin / Heidelberg, 2008.

[11] D. Kern, A. Mahr, S. Castronovo, A. Schmidt, and C. Müller. Making use of drivers' glances onto the screen for explicit gaze-based interaction. In *Proc. AutomotiveUI '10*, pages 110–116. ACM, 2010.

[12] D. Kern and A. Schmidt. Design space for driver-based automotive user interfaces. In *Proc. AutomotiveUI '09*, pages 3–10. ACM, 2009.

[13] D. Kern and S. Schneegass. CARS - configurable automotive research simulator. *i-com*, 8(2):30–33, 2009.

[14] S. Mattes. The lane-change-task as a tool for driver distraction evaluation. *Most*, pages 1–5, 2003.

[15] C. Müller and G. Weinberg. Multimodal input in the car, today and tomorrow. *Multimedia, IEEE*, 18(1):98 –103, Jan. 2011.

[16] S. Oviatt. The human-computer interaction handbook. chapter Multimodal interfaces, pages 286–304. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003.

[17] A. Pauzie. A method to assess the driver mental workload : The driving activity load index (dali). *Humanist*, 2(April):315–322, 2008.

[18] C. Pickering, K. Burnham, and M. Richardson. A review of automotive human machine interface technologies and techniques to reduce driver distraction. In *2nd IET Conf. on System Safety*, pages 223 –228, Oct. 2007.

[19] L. M. Reeves, J. Lai, J. A. Larson, S. Oviatt, T. S. Balaji, S. Buisine, P. Collings, P. Cohen, B. Kraal, J.-C. Martin, M. McTear, T. Raman, K. M. Stanney, H. Su, and Q. Y. Wang. Guidelines for multimodal user interface design. *Commun. ACM*, 47(1):57–59, Jan. 2004.

[20] S. Schneegass, B. Pfleging, D. Kern, and A. Schmidt. Support for modeling interaction with in-vehicle interfaces. In *Proc. Automotive UI '11*, pages 3–10. ACM, 2011.

[21] R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, I. Brewer, A. MacEachren, and K. Sengupta. Speech-gesture driven multimodal interfaces for crisis management. *Proc. IEEE*, 91(9):1327 – 1354, Sept. 2003.

[22] R. Spies, A. Blattner, C. Lange, M. Wohlfarter, K. Bengler, and W. Hamberger. Measurement of driver's distraction for an early prove of concepts in automotive industry at the example of the development of a haptic touchpad. In *Proc. HCII '11*, pages 125–132. Springer-Verlag, 2011.

[23] U. Winter, T. J. Grost, and O. Tsimhoni. Language pattern analysis for automotive natural language speech applications. In *Proc. AutomotiveUI '10*, pages 34–41. ACM, 2010.

[24] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proc. CHI '09*, pages 1083–1092. ACM, 2009.