“Stop over There”: Natural Gesture and Speech Interaction for Non-critical Spontaneous Intervention in Autonomous Driving

Robert Tscharn, Marc Erich Latoschik, Diana Löffler, and Jörn Hurtienne
Julius-Maximilians-Universität Würzburg
Würzburg, Germany
{robert.tscharn,marc.latoschik,diana.loeffler,joern.hurtienne}@uni-wuerzburg.de

ABSTRACT

We propose a new multimodal input technique for Non-critical Spontaneous Situations (NCSSs) in autonomous driving scenarios such as selecting a parking lot or picking up a hitchhiker. Speech and deictic (pointing) gestures were combined to instruct the car about desired interventions which include spatial references to the current environment (e.g., ”stop over [pointing] there” or ”take [pointing] this parking lot”). In this way, advantages from both modalities were exploited: Speech allows for selecting from many maneuvres and functions in the car (e.g., stop, park), whereas deictic gestures provide a natural and intuitive way of indicating spatial discourse referents used in these interventions (e.g., near this tree, that parking lot). The speech and pointing gesture input was compared to speech and touch-based input in a user study with 38 participants. The touch-based input was selected as a baseline due to its widespread use in in-car touch screens. The evaluation showed that speech and pointing gestures are perceived more natural, intuitive and less cognitively demanding compared to speech and touch and are thus recommended as NCSSs intervention technique for autonomous driving.

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**;

KEYWORDS

Multimodal interaction; autonomous driving; speech; pointing gestures; natural user interfaces; intuitive use; workload

ACM Reference Format:
Robert Tscharn, Marc Erich Latoschik, Diana Löffler, and Jörn Hurtienne. 2017. “Stop over There”: Natural Gesture and Speech Interaction for Non-critical Spontaneous Intervention in Autonomous Driving. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI’17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3136755.3136787

1 INTRODUCTION

Autonomous vehicles are currently considered a game changer for the future of individual transportation. Self-driving cars promise to significantly relief drivers from the overall burden of piloting in today’s challenging traffic environments. This relief will come along with a potential increase in the overall passenger comfort and safety. Although technology will introduce its own risks, today 90% of all accidents stem from human errors and autonomous driving promises to foster accident-free traffic [39,41].

However, autonomous driving does not completely take the user out of the loop. User interception to regain direct control over the piloting is considered a necessity in critical spontaneous situations (CSSs), e.g., for legal reasons as soon as the system does not perform reliably any more. But autonomous driving also includes a variety of scenarios where a spontaneous intervention might be intended by the driver without being a critical or dangerous situation (Non-critical Spontaneous Situations, NCSSs). Both CSSs and NCSSs are constrained by available time frames for input, but only CSSs imply critical and threatening consequences if the driver does not intervene with the system. For NCSSs, a missing or wrong intervention is followed only by non-critical consequences. In this work, we focus on the human-machine interface for these NCSSs such as spontaneously picking-up additional passengers (friends or hitchhikers), spontaneously stopping to provide help and assistance for car accidents occurring on the way, or the profane selection of a specific parking spot.

In contrast to the manual interception in CSS, NCSSs have relaxed time constraints and call for alternative and non-disruptive user intervention methods. They require prompt adjustment of the car’s current planning, control, and decision states but without risking additional errors caused by a manual overtake of control. These alternative interventions should seamlessly be integrated in the current driving situation. They should be easy to learn, intuitive to use (defined as allowing to subconsciously apply prior knowledge, e.g., from human-to-human interaction, using a minimum of cognitive resources [28]) and not increase the stress of the passengers or workload for the driver. Intervention speed is important but not first priority and overall safety should not be decreased.

A major aspect of NCSSs is a close correlation to the current environment at the intervention time. It can safely be assumed that deictic information (pointing to or selecting by touch a specific location) will play a major role in most NCSS interventions to differentiate between and identify objects, places, or directions. Today’s graphical user interfaces provide deictic information input by means of touch displays. Hence, touch-based selection provides a well-known input metaphor for NCSS interventions. Additionally, touch displays already found their way into many of today’s in-car systems.

Unfortunately, today’s in-car touch displays often do not align their graphical representations to the environment of the car and

ICMI’17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136787

even Head-Up Displays (HUDs) are projected to the driver's field-of-view regardless of the head position. Hence, they require the users to refocus their visual attention and to remap their egocentric perspective of the surrounding to a potential allocentric perspective on a screen and vice versa affecting cognitive load and situational awareness. We propose an alternative to NCSS interventions which does not break the situational awareness. Inspired by Bolt [2] and Ishii [16], the seamless interface between the user and the surrounding objects is based on natural multimodal interaction ("based on a clear and enjoyable path to unreflective expertise" [43:ix]) utilizing the deictic expressive power of combined gestures and speech to refer to objects, places, and directions in NCSS interventions. In reminiscence to Bolt's ground-breaking work we named our system "Stop-over-There". Our system increases situational awareness and blends seamlessly into the user context. It allows users to intuitively and directly refer to the current environment of the car with gestures and speech which makes graphical displays unnecessary. Thus, minimized attention shifts should reduce cognitive load.

Future autonomous vehicles will provide elaborated sensors and object recognition facilities necessary for the overall autonomous functions and environment representation. Because object recognition is beyond the scope of the work proposed here, we evaluated the usefulness of our system and the proposed intervention method using a driving simulation based on a semi-immersive (wall-projected) Virtual Reality (VR) [33:322,36]. The VR system implicitly provides spatial grounding due to its internal representations. Additionally, the virtual scenarios are augmented with semantic information and objects relevant for interaction are labeled with attributes (color, size, object-class) that the system can access as Semantic Entities [22,23]. Using the system, we investigated the research question whether a combination of speech and pointing gesture will make the input in NCSSs more intuitive for the driver. We compared our system against a straight-forward touch-based input alternative. Our hypothesis that 'speech and pointing gestures' is more intuitive and less cognitive demanding than 'speech and touch' could be confirmed. These results support the claimed benefits of our approach and contribute by exploring a promising NCSS interventions for future autonomous driving.

2 RELATED WORK

2.1 Automation and Intervention

The transaction from manual to autonomous driving can be divided in several stages. Endsley et al. [8] proposed different Levels Of Automation (LOA) for describing the role of a human operator in socio-technical systems. She argued that the function of the operator shifts from actively controlling (manual control) to observing the system (full automation). This classification has also influenced international automation classifications of automotive systems [11,29]. Neither completely manual or completely autonomous levels are regarded as the most challenging ones [4,35] and unexpectedly taking over full manual control in a time critical scenario is a challenging to impossible task for the driver [34].

Interestingly, drivers using almost fully automated systems like the Tesla Autopilot tend to ignore potential risks of these systems. In 2016, Dikmen and Burns [7] reported insights from an online survey with frequent users of this system. Even though drivers perceive automation failures (speed limit not recognized correctly), they do not rate them as risky when the system usually does not need human intervention. This finding is in line with Casner et al. [4] stating that the safest and most trustworthy automation is when no human being is intervening at all or when roles are clearly defined and separated [10]. Hence, changing the vehicle's behavior only by high-level input (within the fully automated mode; similar to an intervening co-pilot and a driver) provides safety and comfort during autonomous driving in NCSSs [17,42].

2.2 Modalities for NCSS Interventions

A variety of modalities is available for high-level NCSS interventions. Yet to apply them successfully, different prerequisites should be met. First, to enable a seamless intervention in the current NCSS, the modality must allow a high *spatial accuracy* for distinguishing intended discourse referents (a specific parking lot or a person). Second, NCSS interventions can be expected to occur not on a regular basis and the intervention style should be *intuitive to use* and require only minimal cognitive workload [28]. Third, it must cover a wide *range of possible maneuvers* and spatial references (e.g., take an exit, pick up a friend along a road, select one of many parking spots). Fourth, it must be *feasible* in the automotive context. Using these four dimensions as a framework, the intervention modalities tablet, speech, and gesture are discussed in terms of their appropriateness for conveying NCSS interventions in the following. Even though other and more exotic modalities have been suggested like, for example, brain computer interfaces [12] or joysticks [18], we focus on these three modalities on account of their pervasiveness in next-generation vehicles.

Tablet. For NCSS interventions in autonomous vehicles, Kauer et al. propose and evaluate a touch-based interface [17]. Here, the driver can overtake a preceding vehicle by selecting a context-independent 'overtake'-button. In a similar approach, Walch et al. suggest a cooperative decision making process which involves both the vehicle and the driver [42]. Their results indicate that this interaction is perceived as comfortable and accepted by drivers. Still, it is only applicable if a few potential maneuvers are available from which can be chosen from. The spatial accuracy of this touch-based approaches is constrained, since the spatial reference points in the world are pre-defined. An alternative could be a tablet-based top-down-view of the environment to select reference points and maneuvers directly in the environment, thereby increasing both the spatial accuracy and number of possible maneuvers. However, this would increase also the complexity of the interface raising issues of intuitive use and cognitive workload. The feasibility of either approach is high since touchscreens are already implemented in many vehicles.

Speech. Speech has been successfully introduced to the automotive market and is already pervasive in the automotive context [1]. Also, its role for in-car human-to-human interaction has been investigated and Cohen et al. provide an extensive interaction corpus based on analyzing in-depth speech and gestures between driver and co-pilot during different driving scenarios [5]. Speech interaction between a navigating co-pilot and driver uses mostly buildings or public spaces, persons or vehicles, and roads or driveways as discourse referents ("take a left turn at the green building"). Walch et al. [42] also successfully evaluated the potential of speech as a NCSS intervention. Still, the spatial accuracy of speech is low since it is difficult to define discourse referents by speech alone (even though sometimes possible for defined scenarios with low complexity like, for example, "the next traffic light"). The intuitive use of speech-based input for NCSS interventions should be high since it can borrow from human-to-human interaction. A broad range of different maneuvers can be communicated by speech alone. Finally, speech recognition already can be found in today's cars and the feasibility of this modality can be regarded as high.

Gestures. Like speech, gestures have already been introduced to the car-related customer market. Freehand gestures as well as micro-gestures on the wheel have been used to control mainly entertainment functions [15]. Mostly, symbolic gestures (e.g., 'fist' for 'stop' or 'play') have been suggested to control in-car systems like the music player or light systems [15,25]. Still, especially deictic (pointing) gestures in human-to-human interaction in the automotive context often support verbal navigation commands [5]. NCSS interventions might build upon this natural way of interaction and include context dependent (deictic) gestures for the spatial identification of discourse referents used in speech commands. Rümelin et al. [38] already explored the potential of unimodal deictic gestures to the car's environment for human-machine interaction. Still, even though pointing gestures with a high spatial accuracy are feasible in the car-context, the variety of NCSS interventions that can be conveyed via unimodal gestures is limited.

In sum, different intervention modalities could be or have been proposed for NCSS interventions but all suffer from drawbacks that justify their multimodal combination instead of unimodal approaches. The following section therefore summarizes literature on multimodal interaction that calls for its application in NCSS interventions.

2.3 A Multimodal NCSS Intervention

Multimodal interaction with digital systems has a long research tradition. Most prominently, Bolt [2] introduced his "Put-that-There" application, that used speech and pointing gestures to naturally interact with a digital system. Insights from this research provide valuable guidance and inspiration for the automotive context [9,19,37,40]. From a technical point of view, multimodal interaction can be expected to be more robust in discourse referencing, because the driver's intention is redundantly conveyed over more than one information channel [37,40]. Also, one modality can enhance and support another modality, a pattern that has been reported by Cohen et al. [5] also in human-to-human interaction (pointing gesture to a house to facilitate speech recognition of "The house at the right corner"). Thus, even though multimodal interaction is not always the best and preferable approach [30], the potential of multimodal interaction should be considered in the automotive domain [26]. For example, Pfleging et al. [32] combine speech and tablet-based gestures (e.g., up/down) to control the vehicle's interior (windows, mirrors). They underline the importance of speech for the selection of functions and tablet-based gestures to provide fine-grained control over these functions. However, despite the prior introduction of multimodal interaction in the automotive context, to the knowledge of the authors, no multimodal system directly addresses the challenge of NCSS interventions in autonomous vehicles.

Thus, we propose a multimodal approach to NCSS interventions that utilizes speech in combination with pointing gestures. Speech allows for selecting from a wide range of interventions and pointing gestures can add the spatial accuracy needed for specifying discourse referents in more complex scenarios (e.g., selection of one specific parking lot). To explore the potential of this input technique, we also evaluate this combination in comparison to a baseline not using pointing gestures but a touch-based top-down view of the environment. Since touchscreens are a core part of modern vehicles and top-down-views can be visualized on them in autonomous vehicles (e.g., [44]), combining them with speech represents a straight-forward solution for communicating high level commands with referencing to the current environment. By this, we contribute by applying the visions of Bolt [2] in the automotive context and providing experimental data for the superiority of this interaction type in terms of its intuitive use and low cognitive demand.

3 SYSTEM DESCRIPTION

We have iteratively developed an interactive system capable of processing multimodal input as a research platform for the evaluation of the proposed multimodal interventions using speech and pointing gestures. Following Latoschik [20,22], the system combines a (projected) semi-immersive VR with a multimodal interface. The system provides a simulation of autonomous driving in different NCSSs and supports processing and fusion of various multimodal input streams (speech, tablet, pointing gestures). The system's simulation capabilities ensure replicability of different NCSSs and interventions.

3.1 Concept

The system supports four different NCSSs (without traffic). Per NCSS, different interventions are possible with only minor differences between both (e.g., select one from many parking lots). Each NCSS intervention can be carried out using either 'speech and pointing gestures' or 'speech and touch'. If the autonomous vehicle is moving, the current traveling path is visualized via a blue semi-transparent strip that is updated as soon as an intervention leads to a changed traveling path. In **'Parking Lot'** (Fig. 1, top left), the participant must select one of six parking lots while the vehicle is standing still on a concrete area. A parking lot in the shadow (only one of six is shaded by a tree) or between two cars (only one is between two cars) can be chosen. Since the car is standing still until the user has entered a complete command in the system, initially no path is visualized. In **'Waving Friend'** (Fig. 1, bottom left), the participant is autonomously driving on an urban main road. Here, the task is to spot a waving friend and take the road she is standing most closely to. Two different variants are available with the waving friend standing at a road on the right or left side. **'Highway Exit'** (Fig. 1, top right) is located on a highway the participant is driving on autonomously. The highway is one-way traffic and no contraflow can occur. The only task is to take the next exit, either on the right or left side. In **'Lateral Swift Maneuver'**

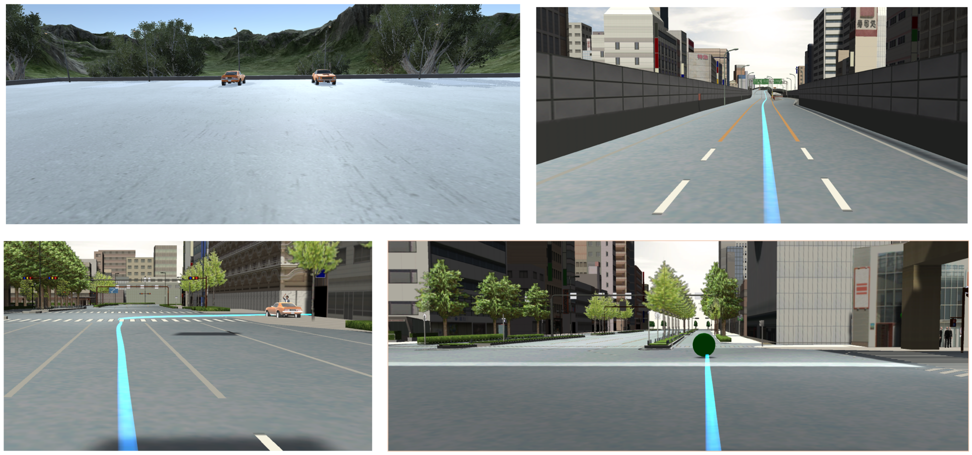

Figure 1: Scenarios investigated in the user study. Top left: Parking Lot. Top right: Highway Exit. Bottom left: Waving Friend. Bottom right: Lateral Swift Maneuver.

Table 1: Examples for NCSS interventions used in this work.

NCSS	Exemplary intervention	Time constraints
Parking Lot	”Park [pointing] there”, ”Use [pointing] this parking lot”	low
Waving Friend	”Take [pointing] this street”, ”Drive right [pointing] there”	low to medium
Highway Exit	”Take [pointing] this exit”, ”Exit the highway [pointing] there”	medium
Lateral Swift Manoeuvre	”Swift in [pointing] this direction”, ”Drive over [pointing] there”	medium to high

(Fig. 1, bottom right), the participant is driving in an urban area. As soon as the vehicle crosses a specific way point along the path, a green ball pops up in the current trajectory and the participant has to indicate the system to avoid the obstacle either to the right or to the left (the car performs a lateral swift to the right or left side to avoid the obstacle). Compared to the first three NCSSs, the last scenario is more standardized. Even though 'Parking Lot', 'Waiving Friend', and 'Highway Exit' are clearer examples for NCSSs, 'Lateral Swift Maneuver' inheres time-pressure but allows for a more standardized measurement of reaction times (Table 1).

To allow for a system supporting these NCSSs and the two interventions, the following requirements were stated for the system architecture:

R1 Internal representation of the environment (e.g., based on sensor data).
R2 Recognition and processing of speech input.
R3 Recognition and processing of gesture input.
R4 Multimodal fusion of both input channels.
R5 Identification of intention and translation to maneuverer.
R6 Dereferencing of discourse referents.

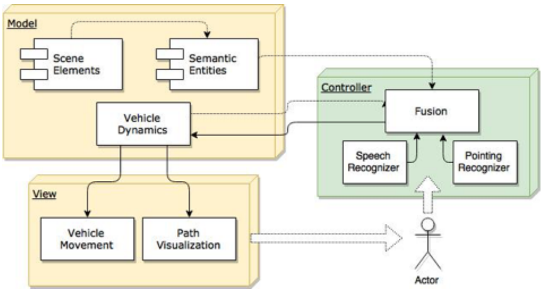

Figure 2: System architecture. Most relevant scene elements (streets, trees, parking lots) are stored as Semantic Entities that are accessed by the fusion module and linked to user input to extract intervention commands and update visualization.

3.2 Implementation

We decided to initially support a semi-immersive VR based on Unity 3D that incorporates all parts of the system including data fusion and internal representations. Based on a MVC-architecture [3], the model-component provides internal information of the driving simulation and driving dynamics, a view-component the visualization of this simulation, and a controller-component collection and fusing driver input to trigger system behavior (Fig. 2). For internal representation (R1) of possible discourse referents (e.g., street, parking lot, tree), relevant objects and areas were represented as Semantic Entities [20] to access object properties (e.g., object class, sizes, color). This allowed for identifying discourse referents (”the street on the right”) using deictic information (pointing gestures to the screen; touch).

The system uses Cortana (mode: current hypothesis for each time stamp) for natural speech input and Kinect v2 for body tracking and recognition of pointing gestures to the projected environment (R2, R3). We implemented a temporal Augmented Transition Network (tATN) as proposed by Latoschik [21] to fuse both modalities

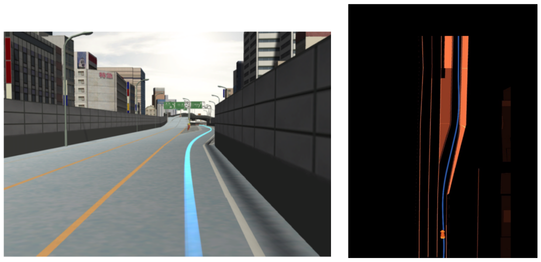

Figure 3: Projection (left) and the top-down view (right).

(R4). The tATN listens for new user input (speech token, pointing gesture, touch-event) and checks if constraints for the transition between two states are met. Following this approach allows the final derivation of exactly one NCSS intervention (R5). By pointing, for example, at a tree along the road, its Semantic Entity is selected by the system (Fig. 2) and used as the discourse referent for a speech command. In the tablet-based version, the top-down view was enhanced by invisible buttons that allowed for selecting the same discourse referents and therefore Semantic Entities (Fig. 3). Since error-free dereferencing of possible discourse referents is essential for the proposed intervention style (R6), the pointing vector is based on the joints ’head’ and ’right fingertip’. We applied a hierarchical best guess approach [24,31] based on the minimal angle between the pointing vector and the head-to-target (target internally represented by its Semantic Entity) vector.

4 EVALUATION

The system is able to internally represent and process scene context based on Semantic Entities for various NCSSs and extracts and executes NCSS interventions. We developed the system iteratively in a user-centered design process (iteration 1: 6 participants; iteration 2: 13 participants). NCSSs and interventions were improved according to user feedback to provide a proper setting and a realistic baseline for the system. For example, the top-down view of the tablet-based intervention was initially too cluttered with irrelevant objects.

In our user study, we compared the two multimodal NCSS interventions: ’speech and pointing gesture’ and ’speech and touch’. For the latter, we implemented a tablet-based top-down view of the current environment that was synchronized in real-time with the above described system (Fig. 3). The difference between both approaches primarily lies in the spatial grounding and anchoring of deictic references in user space vs. device space. For ’speech and pointing gesture’, the driver directly indicates the discourse referent relative to her own perspective (egocentric). In ’speech and touch’, this indication is indirect and relative to other objects in the world (allocentric).

For the evaluation, we used an Intel Core i7-3770 processor (3.40GHz), 16GB 1600MHz DDR3 working memory, and a NVIDIA 2000D graphic chip with 1GB DDR5 working memory. The semi-immersive VR was projected via a Optomo Short Throw Beamer (ML750ST LED) with 700ANSI Lumen and a maximum contrast of 15000:1. The distance between projection and participant was 2.6m and the projected simulation had a width of 2.4m and height of 1.4m.

4.1 Sample

38 German native speaking participants were recruited via a local student panel for course credit (age: M = 22.3, SD = 3.05, 16 male). All but one of them held a driving license and their manual driving experience was on average 6622 km per year (SD=12083 km, min: 0 km, max: 70000 km). Participants were familiar with digital in-car interaction technologies: 92.1% had used a navigation system and 44.7% a hands-free phone system in a car. Also, 76.3% had previously interacted with a digital system using speech recognition and 28.9% were used to full-body tracking technologies (e.g., Kinect).

4.2 Technical Evaluation

First, we analyzed the system’s performance for successfully identifying (disambiguate) important discourse referents as indicated by the deictic pointing gestures. Every participant completed a standardized pointing test consisting of 12 pointing gestures to 6 randomized pointing targets to test the implemented direction-to-semantic-entities resolution. The results were analyzed with two different spatial resolution thresholds: 5° an 10°. The resolution accuracy of the prototype was identified as 99.1% (10° resolution) and 81.0% (5° resolution). These results are in line with known geometric resolution properties of pointing gestures indicating that additional information from speech and discourse will be necessary in case of spatially near-by objects. Overall, our technical implementation provides the necessary gesture input in real-time with the required spatial resolution for later fusion and analysis.

The evaluation of the system’s performance in terms of its capabilities to successfully analyze the complete multimodal utterance was initially performed manually for two of the four NCSSs (’Parking Lot’ and ’Waving Friend’). Here, our system achieved a successful interpretation of the desired intervention in 68.4% and 56.8% of all trials in the two NCSSs over all 38 participants. These low success rates were further analyzed and could be completely routed back to failed recognitions of the utilized speech recognition system. Our gesture detection, analysis and fusion worked as expected. Notably, during these trials we did not individually train the speech recognition to the participants to keep a reasonable low trial duration for the participants. Hence, improvements of the speech recognition process will directly improve the current system performance.

4.3 Experimental Design

To reliably evaluate the proposed intervention method, the experimenter triggered all system-actions using the Wizard of Oz method [6] as soon as a full command (complete speech command plus pointing gesture or touch) had been performed by the participant. By this, biases due to the system’s performance were minimized and system behavior standardized. To check the comparability of delays in both conditions, the subjectively perceived delay between command and system feedback (the current path was visibly updated via the blue strip in front of the car) was rated after each trial. The study was realized using a within-subjects design with two independent variables: two levels of interventions (’speech and pointing gestures’ vs. ’speech and touch’) and four levels of NCSS (described above). Participants completed all four NCSSs with both interventions. The presentation order of the factor ’intervention’

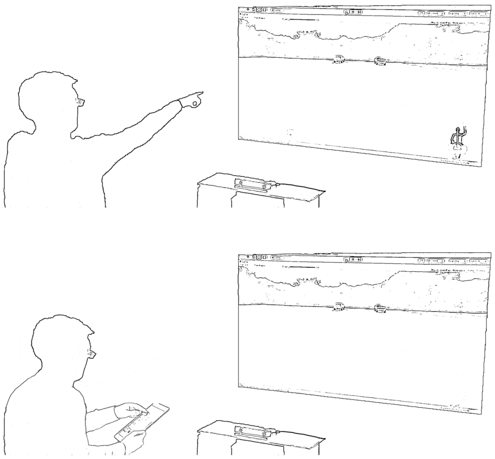

Figure 4: Setup of the study for 'speech and pointing gesture' (top) and 'speech and touch' (bottom). Example: "Drive to [pointing gesture or tap on the tablet] this parking lot". A Kinect and a small microphone were used in both conditions for data collection and a projection (no head-mounted display) displayed the semi-immersive VR.

was blocked (half of all participants started with the four NCSSs using 'speech and pointing gestures', the other half 'speech and touch'). The order of NCSSs within each intervention block was completely randomized for all participants.

4.4 Material

To measure the participants' *cognitive workload* during the scenarios, two standardized measures were used: The NASA-TLX, a widely used and time-proven instrument for estimating the overall human workload during performing a task [13,14] and the SMEQ which focuses on the perceived mental workload. To compare both interventions in terms of their *intuitive use*, the QUESI-questionnaire was applied [28]. Five 7-point Likert-scales were used to also collect data on scenario- and intervention-specific aspects. In questions one and two, participants rated how '*natural*' and '*spatially precise*' the intervention had been. In question three, participants rated how '*fast*' the system executed a command after it was given by them (Wizard of Oz check for latency differences between conditions). In questions four and five, participants rated how '*aware*' they had been about their environment during the interaction phase and how '*frequent*' this scenario would be in real traffic.

4.5 Procedure

After giving informed consent and completing a pre-questionnaire, participants were introduced to the two interventions: 'speech and pointing gestures' and 'speech and touch'. First, the pointing gesture recognition was tested using the Live Preview of Visual Gesture Builder [27]. Indicated by a cursor controlled by the experimenter, participants pointed with their right hand at different parts of the projection they sat in front of. No NCSSs were presented but participants were informed that the system would recognize which objects or areas they would be pointing at later in the following simulation. Afterwards, the speech recognition was explained. For speech recognition, a small microphone was pinned to participants during the entire experiment. They were informed that they could use natural language and the system would understand appropriate commands ("drive over there", "take this exit"). Finally, the tablet-based intervention was instructed and they were told that the system could recognize touch in the environment and combine them with speech commands. It was emphasized that for a complete command, speech had to be combined with, depending on the intervention, a pointing gesture to or a tap on an object or area.

In the main part, the four NCSSs were tested with the first intervention. Each NCSS was presented using the full screen mode of the Unity 3D game-view and started with a black screen, during which the experimenter read out a short and standardized use-case instruction. After the experimenter removed the black screen, the camera perspective (from the driving seat of the simulated vehicle) was projected to the screen. Except for 'Parking Lot', where the vehicle was not moving until a complete command had been recognized (see section 'Concept'), participants started the autonomous vehicle by the German phrase "Start". The acceleration of the vehicle was fast enough to allow for an appropriate traveling time until any intervention was required to complete the given use-case. A specific end state was defined after which the experimenter closed the NCSS (e.g., after reaching the intended parking lot). After each NCSS, participants completed the NASA-TLX, the SMEQ-scale, the five Likert-scales (see section 'Materials'), and wrote down qualitative feedback. After a block of all four NCSSs with one intervention, participants completed the QUESI questionnaire. This procedure was repeated for the second intervention (four NCSS; newly randomized order). One of the two NCSS-variants (e.g., take left highway exit) was used for the first intervention and the second (take right highway exit) for the second intervention. The order of variants was blocked. Finally, participants completed the standardized pointing task used for the technical evaluation. Depending on the participant, the entire experiment lasted between 40 and 60 minutes.

4.6 Hypotheses

Contrary to the combination of 'speech and touch' (from now on called ST), the combination 'speech and pointing gesture' (from now on called SP) builds on prior knowledge from natural human-to-human interaction. Thus, it was expected that

H1: SP is rated as significantly more intuitive to use and more natural than ST.

Also, a direct interaction with the environment should require less cognitive demand compared to an indirect one. Thus, it was expected that

H2: SP is rated as significantly less cognitive demanding than ST.

Finally, in the condition SP, attention can remain in the environment, whereas in ST, attention has to be split to two "screens". Based on the measures up-down head rotations and the question 'How aware were you about the environment', we expected that

H3: SP leads to significantly more situational awareness from the environment than ST.

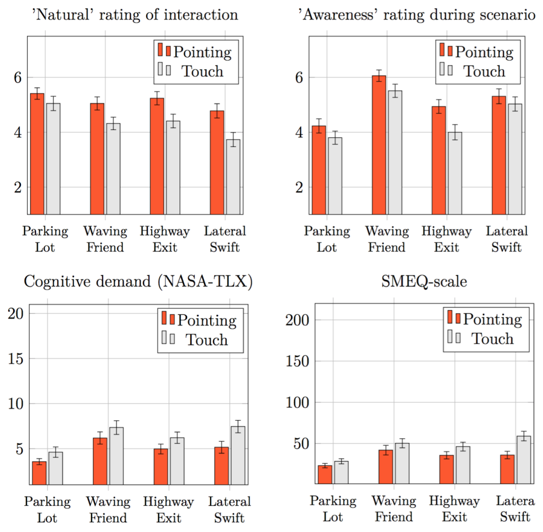

Figure 5: Results of the user-study (Mean and SE).

5 RESULTS

Differences in degrees of freedom result from single missing data in the dataset (technical reasons of the tracking device, participants forgetting to rate one item) or adjustment (e.g., violation of sphericity assumption). To compare delays of the Wizard of Oz method over conditions, we analyzed the ratings of the ’fast’ question via an ANOVA with ’intervention’ and ’NCSS’ as repeated factors. No significant differences were found for the factors ’intervention’, *F*(1,35)=2.6, *p*=.12, or ’NCSS’, *F*(1,35)=1.78, *p*=.16. Qualitatively, SP was described as ”intuitive” and ”easy”. ST was perceived as ”somewhat unnatural” and ”confusing, because I have to switch between the tablet and the projection”. Participants also reported that ST ”distracts from environment” and was ”more difficult to use”. Though, some participants remarked that multimodality is unnecessary in some scenarios like ’Highway Exit’ and could be replaced by a unimodal speech command like ”take the next exit”. Other participants emphasized the importance of an overview (”I like the overview of the tablet”) and one suggested a combination of SP and a top-down view displayed on a side tablet.

Intuitive Use and Natural Interaction. The intervention SP was rated significantly higher on the QUESI (*M*=3.97, *SD*=0.57) than ST (*M*=3.66, *SD*=0.72), *t*(37)=2.76, *p*=.009, *Cohen’s d*=0.45. Comparing the five subscales (alpha-correction to .01) showed that SP was rated significant lower regarding on the dimensions ’cognitive demand’ and ’needed learning effort’, *t*(37)=4.32, *p*<.001, *d*=0.72 and *t*(37)=2.99, *p*=.005, *d*=.5, respectively. An ANOVA was conducted with ’intervention’ and ’NCSS’ as repeated factors and the rating of the ’natural’ question as dependent variable. The two main effects ’intervention’, *F*(1,36)=18.37, *p*<.001, η_p^2=.34, and ’NCSS’, *F*(3,36)=9.23, *p*<.001, η_p^2=.20, reached level of significance, but not the interaction between both.

Workload and Situational Awareness. In an ANOVA (same repeated factors as above) and the subscale ’cognitive demand’ of the NASA-TLX as the dependent variable, the main effect for ’intervention’ was significant, *F*(1,36)=11.86, *p*=.001, η_p^2=.25, with lower values for SP compared to ST. Also, the main effect of the factor ’NCSS’ was significant, *F*(2.54,36)=10.49, *p*<.001, η_p^2=.23. For the SMEQ-scale, also both main effects of ’intervention’ and ’NCSS’ were significant, *F*(1,34)=14.09, *p*=.001, η_p^2=.29, and *F*(2.48,34)=9.95, *p*<.001, η_p^2=.23 (with adjusted degrees of freedom by Greenhouse-Geisser correction due to violation of the sphericity assumption), but not the interaction between both. The ratings on the ’awareness’ question revealed a significant factor ’intervention’, *F*(1,34)=9.77, *p*=.004, η_p^2=.22, and a significant factor ’NCSS’, *F*(3,34)=20.66, *p*<.001, η_p^2=.38, but again no interaction between both. Over all NCSS, participants reported a significantly higher perceived spatial precision of the intervention SP (*M*=5.27, *SD*=0.94) compared to ST (*M*=4.82, *SD*=1.04), *t*(35)=2.10, *p*=.04, *d*=.35. As a possible measurement of the participant’s distraction in the most standardized NCSS ’Lateral Swift Maneuverthe variance of head rotation (up-down) was analyzed as recorded by the Kinect sensor (based on manual inspection, head tracking worked reliably in both conditions for 30 participants). The variance of head rotation in degrees was significantly lower for SP (*M*=0.61, *SD*=0.59) than for ST (*M*=1.67, *SD*=1.76), *t*(29)=3.17, *p*=.004, *d*=.65

Time pressure and reaction time. The subscale ’time pressure’ of the NASA-TLX was also analyzed. An ANOVA (same repeated factors as above) revealed a significant factor ’NCSS’, *F*(3,34)=7.98, *p*<.001, η_p^2=.19, but not ’intervention’. Post-hoc test using Helmert contrasts showed that ’Parking Lot’ was rated significantly less time critical compared to the other three NCSS, *F*(1,34)=17.45, *p*<.001, η_p^2=.34.

As an objective measurement of the participants’ behavior, the reaction time was analyzed. Only the NCSS ’Lateral Swift Maneuvre’ allowed for a strictly standardized calculation of a reaction time (starting from when the green ball spawned on the road). Here, the first speech input was recognized earlier by the system in SP (*M*=1.47s, *SD*=0.31s) than in ST (*M*=1.91s, *SD*=0.61s), *t*(23)=3.56, *p*=.002, *d*=.81. Also, the pointing gesture or tablet-touch occurred earlier for SP (*M*=1.54s, *SD*=0.47s) than for ST (*M*=2.04s, *SD*=0.72s), *t*(23)=3.25, *p*=.004, *d*=.68.

6 DISCUSSION

Autonomous driving will require novel intervention approaches especially in NCSSs. To explore the potential of multimodal interventions, we implemented an interactive system that allows for the comparison of two input techniques: a) speech and pointing gestures and b) speech and touch. To address this research question, we utilized a semi-immersive VR and the Wizard of Oz method. Notably, our system already fulfills all initially defined requirements (R1-R6) and is capable of processing multimodal input to interpret speech and pointing gestures in a given environment. Here we introduced the overall concept, system architecture, and a first evaluation: (1) The tablet-based baseline provides a straight-forward multimodal solution using available in-car displays (e.g., [44]). (2) The simulation provides replicable NCSSs that are otherwise hard to achieve under real-world conditions. (3) The Wizard of Oz aspect provides standardization and avoids influences of system performance on the results of the input technique evaluation.

38 participants completed four NCSSs in a semi-immersive VR, each with both interventions. The hypothesis was that ’speech and

pointing gesture' would be rated as more intuitive to use and more natural (H1), require less cognitive resources (H2), and lead to a higher situational awareness (H3) than in the condition 'speech and touch'. Results confirm the hypothesizes regarding the overall effects between both intervention styles. 'Speech and pointing gesture' is rated significantly more intuitive and natural over all NCSSs. Also, participants report a higher subjectively perceived awareness for the intervention 'speech and pointing'. In the most standardized NCSS ('Lateral Swift Maneuver' participants reacted objectively faster with the intervention 'speech and pointing gesture'. One explanation for this could be that participants have to split their attention to two 'areas of interest' when using the intervention 'speech and touch' (the simulated environment and the tablet with the top-down environment). The split of attention might also be represented in the fact that participants rotated their head significantly more in the up-down direction when choosing targets via the tablet. Also, participants reported that switching between the two "screens" is more challenging and stressful compared to the intervention 'speech and pointing gesture', generally described as natural and direct. In sum, the maneuver-based approach is perceived as useful. This supports the claim of [17] and [42] that high-level interventions are a promising and approach for autonomous driving.

The indirectness of the intervention 'speech and touch' is also reflected in its significantly high cognitive load. Low switching costs from a driving-unrelated task such as reading to a NCSS intervention must be regarded as critical for its acceptance in autonomous driving. Generally, both 'speech and pointing gesture' and 'speech and touch' provide the same functionality. Speech allows for selecting from a broad variety of maneuvers while the pointing gesture or tap on the tablet allow for a spatial referencing of specific objects or areas. But the intervention style 'speech and pointing gesture' requires lower cognitive resources over all investigated NCSSs.

One could expect larger differences between both interventions when a NCSS is perceived as highly time critical. This is not the case. Indeed, the NCSSs differ significantly in almost all collected measures, including the time criticality as measured in the NASA-TLX subscale 'time pressure'. 'Parking Lot' is the least demanding NCSS, probably because of its rather static setting until the final command has been given. But the interaction between the two factors 'NCSS' and 'intervention' is not significant in any of the measures and 'speech and pointing gesture' does not benefit more in NCSS with a higher time pressure. One explanation could be that none of the NCSS was perceived (as also not intended) as extremely time critical. Even though one could argue that time criticality may play a role in NCSSs with very short time frames, these were not subject of this intervention style. Also in level 5 of automotive transportation [29], driver initiated intervention under time pressure is not of interest [4].

One participant also suggested a combination of speech, pointing gestures and the tablet-based top-down view of the environment to supplement the intervention of speech and pointing gestures with an overview of the current situation similar to a navigation system. This was interesting since the environment was not too complex in this study and no traffic was added to the scenarios. Still, users might benefit even more from this three-fold approach. For example, displaying surrounding cars could provide an advantage compared to the pointing gestures without the top-down view.

In sum, the combination of speech and pointing gestures is perceived as intuitive to use in the context of autonomous driving. It provides a more direct and natural NCSS intervention compared to a touch-based approach. Our multimodal NCSS intervention is perceived as differently appropriate over all NCSSs. Future studies could therefore build on these findings and address a variety of questions. First, an open question is which NCSSs benefit most from this type of multimodal intervention and under which circumstances a unimodal maneuver-based interface [42] might be sufficient even though it provides only a limited set of maneuvers and a low spatial accuracy. Second, relevant discourse referents were not made visible or highlighted in the NCSSs of this work. Using augmented reality elements in the windshield of a real car might enhance the intuitive use of pointing gestures to the environment even further. And third, the impact of this approach to a more diverse user group must be considered. For example, older adults with less prior knowledge about navigation systems and tablets could benefit more from our NCSS intervention that borrows from natural human-to-human interaction.

7 CONCLUSION

We propose and evaluate a natural and intuitive NCSS intervention for autonomous driving by combining speech and pointing gestures. This multimodal approach exploits advantages of these two input techniques: using speech allows for selecting from a very broad variety of different commands while pointing gestures can enhance the speech command by spatially specifying discourse references. The chosen system architecture effectively decouples the multimodal processing from the user tracking and scene recognition task. Hence it could be conveniently combined with current advances of in-car sensors and object recognition for autonomous driving and could be used for physical prototypes.

The proposed intervention scores high along the discussed four dimensions spatial accuracy, intuitive use, range of maneuvers, and feasibility. But besides these dimensions, what are the most important benefits of pointing and speech? A fast intervention is important but not crucial for autonomous driving since the autonomous vehicle will most probably be constrained by safety margins and a sudden and abrupt change of the traveling path is not likely. More important is that the proposed intervention is rated and perceived as natural and intuitive. Thus, humans are able and willing to transfer prior knowledge obtained from human-to-human interactions to the human-to-vehicle interaction. Also, the cognitive demand for this transfer is low and enables for direct interaction with the environment. NCSS interventions using speech and pointing gestures allow for a seamless interaction and do not put any additional cognitive burden on the driver by the interaction itself. Even though this intervention builds on Bolt's "Put-That-There" metaphor [2], not only multimodality itself is important for a seamless interaction with the environment, but also *which* modalities are combined. Our results indicate that combining direct input techniques (pointing gestures) with speech is a promising candidate for future ways of human-to-vehicle NCSS interventions.

REFERENCES

[1] Leonardo Angelini, Andreas Sonderegger, Jürgen Baumgartner, Francesco Carrino, Stefano Carrino, Maurizio Caon, Omar Abou Khaled, Jürgen Sauer, Denis Lalanne, and Elena Mugellini. 2016. Comparing Gesture , Speech and Touch Interaction Modalities for In-Vehicle Infotainment Systems. *Actes de la 28ieme conference francophone sur l’Interaction Homme-Machine on - IHM ’16*: 188-196.

[2] Richard Bolt. 1980. ”Put-that-there.” *Proceedings of the 7th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’80*: 262-270.

[3] James Bucanek. 2009. Model-View-Controller Pattern. In *Learn Objective-C for Java Developers.* Apress, Berkeley, CA, 353-402.

[4] Stephen Casner, Edwin Hutchins, and Don Norman. 2016. The Challenges of Partially Automated Driving. *Communications of the ACM 59*, 5: 70-77.

[5] David Cohen, Akshay Chandrashekaran, Ian Lane, and Antoine Raux. 2014. The HRI-CMU Corpus of Situated In-Car Interactions. *International Workshop Series on Spoken Dialogue Systems Technology*: 201-212.

[6] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies - why and how. *Knowledge-Based Systems 6*, 4: 258-266.

[7] Murat Dikmen and Catharine Burns. 2016. Autonomous Driving in the Real World: Experiences with Tesla Autopilot and Summon. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI ’16)*, Ann Arbor, MI, USA., 225-228.

[8] Mica Endsley and David Kaber. 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics 42*, 3: 462-492.

[9] Martin Fischbach. 2015. Software Techniques for Multimodal Input Processing in Realtime Interactive Systems. In *Proceedings of the International Conference on Multimodal Interaction - ICMI’15*: 623-627.

[10] Yannick Forster, Frederik Naujoks, and Alexandra Neukum. 2016. Your Turn or My Turn? Design of a Human- Machine Interface for Conditional Automation Your Turn or My Turn? Design of a Human-Machine Interface for Conditional Automation. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI’16*, 253-260.

[11] Tom Michael Gasser. 2012. *Ergebnisse der Projektgruppe Automatisierung: Rechtsfolgen zunehmender Fahrzeugautomatisierung.* Bergisch Gladbach.

[12] Daniel Göhring, David Latotzky, Miao Wang, and Raul Rojas. 2013. Semi-autonomous Car Control Using Brain Computer Interfaces. Springer, Berlin, Heidelberg, 393-408.

[13] Sandra Hart. 2006. Nasa-Task Load Index (NASA-TLX) 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*, 9: 904-908.

[14] Sandra Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology 52,* C: 139-183.

[15] Renate Hauslschmid, Benjamin Menrad, and Andreas Butz. 2015. Freehand vs. micro gestures in the car: Driving performance and user experience. *2015 IEEE Symposium on 3D User Interfaces (3DUI) 336*: 159-160.

[16] Hiroshi Ishii and Brygg Ullmer. 1997. Tangible bits: towards seamless interfaces between people, bits, and atoms. In *Proceedings of the 8th international conference on Intelligent user interfaces*, 234-241.

[17] Michaela Kauer, Benjamin Franz, Michael Schreiber, Ralph Bruder, and Sebastian Geyer. 2012. User acceptance of cooperative maneuverbased driving - A summary of three studies. *Work 41*, SUPPL.1: 4258-4264.

[18] Martin Kienle, Daniel Damböck, Heiner Bubb, and Klaus Bengler. 2013. The ergonomic value of a bidirectional haptic interface when driving a highly automated vehicle. *Cognition, Technology and Work 15, 4*: 475-482.

[19] Denis Lalanne, Laurence Nigay, Philippe Palanque, Peter Robinson, Jean Vanderdonckt, and Jean-Francois Ladry. 2009. Fusion engines for multimodal input: a survey. *International Conference on Multimodal Interfaces*: 153-160.

[20] Marc Erich Latoschik. 2001. A General Framework for Multi-Modal Interaction in Virtual Reality Systems: PrOSA. In *The Future of VR and AR Interfaces-Multimodal*, 21-25.

[21] Marc Erich Latoschik. 2002. Designing transition networks for multimodal VR-interactions using a markup language. In *Proceedings of the International Conference on Multimodal Interfaces - ICMI’02*: 411-416.

[22] Marc Erich Latoschik. 2005. A user interface framework for multimodal VR interactions. In *Proceedings of the International Conference on Multimodal Interfaces - ICMI’05*, 76-83.

[23] Marc Erich Latoschik and Christian Fröhlich. 2007. Semantic reflection for intelligent virtual environments. In *Proceedings - IEEE Virtual Reality*, 305-306.

[24] Marc Erich Latoschik and Ipke Wachsmuth. 1998. Exploiting distant pointing gestures for object selection in a virtual environment. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 185-196.

[25] Sebastian Loehmann, Martin Knobel, Melanie Lamara, and Andreas Butz. 2013. Culturally independent gestures for incar interactions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 538-545.

[26] Jannette Maciej and Mark Vollrath. 2009. Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention 41*, 5: 924-930.

[27] Microsoft Corporation. 2014. Visual Gesture Builder (VGB). Retrieved February 24, 2017 from https://msdn.microsoft.com/de-de/library/dn785304.aspx

[28] Anja Naumann, Jörn Hurtienne, Johann Habakuk Israel, Carsten Mohs, Martin Christof Kindsmüller, Herbert A. Meyer, Steffi Husslein, and IUUI Research Group. 2007. Intuitive use of user interfaces: Defining a vague concept. *Engineering Psychology and Cognitive Ergonomics*: 128-136.

[29] NHTSA. 2013. Preliminary statement of policy concerning automated vehicles. National Highway Traffic Safety Administration. National Highway Traffic Safety Administration: 1-14.

[30] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Communications of the ACM 42*, 11: 74-81.

[31] Thies Pfeiffer and Marc Erich Latoschik. 2004. Resolving object references in multimodal dialogues for immersive virtual environments. In *Proceedings - Virtual Reality Annual International Symposium*, 35-42.

[32] Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. 2012. Multimodal interaction in the car - combining speech and gestures on the steering wheel. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 155-162.

[33] Bernhard Preim and Raimund Dachselt. 2015. *Interaktive Systeme: Band 2: User Interface Engineering, 3D-Interaktion*. Springer Vieweg: Wiesbaden.

[34] Jonas Radlmayr and Klaus Bengler. 2015. Literaturanalyse und Methodenauswahl zur Gestaltung von Systemen zum hochautomatisierten Fahren. *FAT-Schriftenreihe 276*: 1-57.

[35] Jonas Radlmayr, Christian Gold, Lutz Lorenz, Mehdi Farid, and Klaus Bengler. 2014. How Traffic Situations and Nondriving-Related Tasks Affect the Takeover Quality in Highly Automated Driving. In *Human Factors and Ergonomics Annual Meeting*, 2063-2067.

[36] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The Office of the FutureâĂŕ: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. *SIGGRAPH '98 Proceedings of the 25th annual conference on Computer graphics and interactive techniques*: 1-10.

[37] Leah M. Reeves, Jean-Claude Martin, Michael McTear, TV Raman, Kay M. Stanney, Hui Su, Qian Ying Wang, Jennifer Lai, James A. Larson, Sharon Oviatt, T. S. Balaji, Stephanie Buisine, Penny Collings, Phil Cohen, and Ben Kraal. 2004. Guidelines for multimodal user interface design. *Communications of the ACM 47*, 1: 57-59.

[38] Sonja Rümelin, Chadly Marouane, and Andreas Butz. 2013. Free-hand pointing for identification and interaction with distant objects. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*: 40-47.

[39] JR Treat. 1977. Tri-Level Study of the Causes of Traffic Accidents: An overview of final results. In *Proceedings of the American Association for Automotive Medicine Annual Conference*, 391-403.

[40] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters 36*, 189-195.

[41] Volvo. 2013. Volvo Trucks European Accident Research and Safety Report 2013. Retrieved February 20, 2017 from https://www.kenallenlaw.com/2013/02/new-volvo-truck-study-9-out-of-10-truck-accidents-in-europe-caused-by-human-factor-including-distracted-driving-while-ntsa-proposes-new-trucking-regs-to-fmsca-after-2011-nevada-crash/

[42] Marcel Walch, Tobias Sieber, Philipp Hock, Martin Baumann, and Michael Weber. 2016. Towards Cooperative Driving: Involving the Driver in an Autonomous Vehicle's Decision Making. In *Proceedings of AutomotiveUI'16*: 261-268 .

[43] Daniel Wigdor and Dennis Wixon. 2011. Brave NUI world: designing natural user interfaces for touch and gesture. Elsevier.

[44] 2017. driveAI. Retrieved August 20, 2017 from www.drive.ai