Hands-free but not Eyes-free: A Usability Evaluation of Siri while Driving

Helene Høgh Larsen
Science, Policy and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen, Denmark

Alexander Nuka Scheel
Science, Policy and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen, Denmark

Toine Bogers
Science, Policy and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen, Denmark
toine@hum.aau.dk

Birger Larsen
Science, Policy and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen, Denmark
birger@hum.aau.dk

ABSTRACT

Distractions while driving are a major cause of traffic accidents and chief among these is the use of mobile phones. Driver distractions typically fall into four categories—visual, cognitive, bio-mechanical, and auditory—and different technological solutions have been proposed to address these. Intelligent Personal Assistants (IPAs), such as Siri, is a recent example of such a technological solution that offers the potential for hands-free phone interaction through a voice-controlled interface. IPAs could potentially reduce visual and bio-mechanical distractions if they are usable enough to not increase a driver's cognitive load. We present the results of a controlled experiment with the aim of understanding how the use of Siri while driving compares to manual interaction in terms of usability and distractions. We also tested these two interaction types in the lab in order to understand how the main driving task influences Siri's (perceived) usability. Our study shows that Siri is not ready for every-day use in the car: interacting with Siri while driving is likely to be unsafe for most participants, especially less experienced drivers. Participants were distracted by Siri due to its over-reliance on visual feedback as well as frequent time-outs by Siri when waiting for a response from a driver occupied with the road environment. Speech recognition quality in a noisy car as well as problematic multi-lingual speech recognition in general are other issues that resulted in low usability and more cognitive distractions. While interacting with Siri may be hands-free, it does not provide an eyes-free and distraction-free experience yet,

KEYWORDS

Siri, usability testing, driving, IPA, intelligent personal assistants

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377962

ACM Reference Format:
Helene Høgh Larsen, Alexander Nuka Scheel, Toine Bogers, and Birger Larsen. 2020. Hands-free but not Eyes-free: A Usability Evaluation of Siri while Driving. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377962

1 INTRODUCTION

In the period between 2014 and 2016, the Danish Council for Traffic Safety estimated that distractions while driving were the cause of 33% to 50% of all traffic accidents in Denmark [5]. One of the major causes of distraction was mobile phone use, even though drivers are generally aware of the dangers of mobile phone use while driving [6]. For instance, texting and calling while driving increases the risk of an accident by a factor of 3.8 to a factor of 8.3 respectively [20]. A likely explanation for the continued use of mobile phones despite this awareness is the (perceived) essential role that mobile phones play in people's everyday lives in terms of completing navigation, communication, and general search tasks [3] and the habitual nature of reaching for one's mobile phone when faced with such tasks, regardless of the context of use.

According to Ranney et al. [28], driver distractions typically fall into one of four categories: (1) *visual*, when a driver is looking at objects or events other than the road environment; (2) *cognitive*, when a driver is thinking about something not related to driving the vehicle; (3) *bio-mechanical*, when a driver is doing something physical that is not related to driving (e.g., reaching for their phone or eating food); and (4) *auditory*, when a driver is distracted by sounds not related to driving that prevent them from making the best use of their hearing. While many forms of manual mobile phone use have been outlawed altogether, some technological solutions such as hands-free calling are allowed in some countries as they are meant to reduce bio-mechanical distractions, even though they still impair driving coordination and control [25, 35].

Another, more recent type of technology that is advertised as having the potential to make mobile phone use safer while driving are so-called *intelligent personal assistants*. Intelligent Personal Assistants (IPA) are voice-controlled software agents that support task-oriented exchanges with the user, such as making restaurant reservations, getting directions, dictating text messages, and searching the Web. The voice-controlled, conversational style of interaction that IPAs provide, could be beneficial in many hands-free scenarios, such as aiding the visually impaired or driving a car. The latter scenario is addressed by, among others, Apple's CarPlay[1] offering, which allows a driver to control their mobile phone through a combination of in-car touch-screen controls and voice control using Apple's own IPA called Siri. Potentially, the use of such voice-controlled IPAs while driving could reduce both biomechanical and visual distractions, while possibly increasing auditory distractions. Ideally, cognitive distraction would stay at the same levels, provided the IPA is usable enough for use in true hands-free settings.

While the usability of IPAs has been the subject of a handful of studies already [2, 18, 19, 23], all of them were performed in a controlled environment where the participants' sole task was interacting with the IPA and where voice control could be supplemented with manual interaction, such as scrolling through results lists or selecting options. Only Strayer et al. [34] compared different IPAs in a driving setting and found that interacting with an IPA while driving added significantly to the driver's cognitive workload compared to driving without any distractions. However, there has been no work on examining how interacting with an IPA compares to manual smartphone interaction while driving a vehicle. It is also unclear how drivers perceive the usability of an IPA when IPA interaction becomes a secondary activity with driving being their main priority. In addition, all of these usability tests have used English as the interaction language, with the recent work by Bogers et al. [2] on the Danish version of Siri being the sole exception.

In this paper, we take several steps towards addressing some of these open issues. We present the results of a controlled experiment with the goal of better understanding how usable IPAs are in natural, hands-free situations where the user is occupied with another primary task, such as driving a car, and to what degree the IPA is successful at reducing driver distractions. Similar to Bogers et al. [2], we use Danish as our interaction language and we focus exclusively on Siri, as it was shown to be the most popular IPA in Denmark by far around the time we conducted our experiment [2] and because interacting with an IPA in one's native language as opposed to non-native ones is likely to result in a lower cognitive load. Through our controlled experiment, we answer the following pair of research questions:

> **RQ1** How does interacting with Siri while driving compare to manually interacting with an iPhone in a car (under strict security precautions) in terms of usability and distraction?
> **RQ2** How is interaction with Siri impacted by driving compared to interacting with Siri in a controlled setting?

Our first research question enables us to draw conclusions about how serious an alternative Siri could be for in-car phone use, compared to manually interacting with a smartphone while driving. Because the latter is illegal in many countries, including Denmark, we take strict security precautions to prevent illegal and reduce dangerous behavior. Our second research question gives us insight into the question of how generalizable the results of previous usability tests of IPAs are, given that they were conducted under controlled conditions and as the primary task. Such a controlled setting could be seen as the best possible result for Siri, which gives us insights into whether Siri has been adequately adapted to a driving setting. A smaller difference in perceived usability between these conditions could be seen as good news for Apple (and other IPA producers). We conclude our paper by providing several design implications for IPAs to streamline their use when driving motorized vehicles.

[1] https://www.apple.com/ios/carplay/

2 BACKGROUND

2.1 Mobile Phones & Driving Safety

The link between mobile phone use and driving safety has been studied extensively over the past decades as well as the impacts of different interaction styles—hand-held versus hands-free. There seems to be little doubt in the literature about the dangers of texting and calling while driving: Klauer et al. [20] reported that these activities increase the risk of an accident by a factor of 3.8 to a factor of 8.3 respectively, despite the fact that drivers are generally aware of these dangers [6]. Studies of drivers' eye fixations while performing demanding cognitive tasks have shown that their visual field narrows both horizontally and vertically—meaning that rather than scanning the road environment for hazards, drivers spend more time staring directly ahead, leading to tunnel vision [10, 35]. In general, the more complex or time-consuming the second activity in addition to driving is, the more demands are placed on the driver, causing them to become less observant or to make poor decisions about how to control the vehicle safely.

Perhaps surprisingly, there does not seem to be a clear difference in distraction levels between hand-held versus hands-free interaction [16], despite the latter's potential to reduce bio-mechanical distractions. Nevertheless, several studies have found that hands-free calling and texting still impairs driving coordination and control [25, 35]. The conversational style of IPA interaction could lead to the interesting comparison between mobile phone use while driving versus conversations with other passengers. While research does show that drivers are more distracted when speaking themselves compared to listening, in general people do drive better when conversing with passengers compared to mobile phone conversation, because the traffic and driving tasks become a part of the passenger conversation, thereby increasing situational awareness [7]. This suggests that, while potentially distracting, interacting with an IPA could beneficial compared to direct mobile phone use.

2.2 Usage & Usability of IPAs

Even though the current generation of IPAs, such as Alexa, Cortana, Google Voice and Siri, have only been around for less than a decade and needed time to mature in terms of capabilities, they have seen a rapid rise in popularity. In April 2018, 41.4% of surveyed US adults reported using the IPA on their smartphone, while 19.7% were found to use a smart speaker IPA [17]. This popularity is mirrored in terms of increasing research attention for IPAs and other conversational agents. Usage of IPAs was studied by, among others, Garcia et al. [8], who conducted a questionnaire about IPA usage in Argentina, Brazil, Chile, Germany, Spain, the UK and the US. They found that IPA usage in other countries lags behind the US, but that around 50-60% of IPA users use them several times a week. Bogers et al. [2] performed a similar study in Denmark, and found that Siri was by far the most popular IPA due to support for the Danish language, but that only 20% of respondents used Siri more than once a month.

Other researchers have attempted to evaluate the usability of IPAs, such as Kiseleva et al. [18], who investigated a range of different scenarios, such as device control, Web search, and structured search. They discovered that user satisfaction is highly task-dependent, which is similar to what other studies evaluating IPA usability have found [2, 23]. Bogers et al. [2] tested the usability of Siri in Danish, as all of the other work on usability testing IPAs had been done in English. They found that multi-lingual speech recognition can be problematic: mixing English named entities into Danish requests resulted in increased time, effort and the number of speech recognition errors by Siri. They also found that prior experience with Siri as well as the way results are presented to the user had a strong positive influence on user satisfaction. We expect similar challenges and patterns to emerge from our study.

What all of these studies have in common, however, is that they were all performed in a controlled environment where IPA interaction was the participants' sole focus, and where voice control could be supplemented with manual interaction, such as scrolling through results lists or selecting options. This is in contrast to the use of IPAs in hands-free settings. To the best of our knowledge, only Strayer et al. [34] have compared the usability of different IPAs in a driving setting. They found that interacting with an IPA while driving added significantly to the driver's cognitive workload when compared to driving without any distractions. However, Strayer et al. [34] did not compare IPA interaction while driving to manual smartphone interaction—the most common (and in most situations illegal) activity—nor to interaction with IPAs as a primary activity in a controlled environment, which means a baseline for cognitive load is missing from their work—something we address in our study.

2.3 Multi-tasking & Information Behavior

In the majority of the user-based studies of IPAs, the interaction with the IPA is the sole task the participants have to focus on [2, 18, 19, 23], making it difficult to know how the introduction of a second task would affect usability, task performance and distraction levels. Spink et al. [32] reviewed a broad range of research on multitasking behavior. They find that the cognitive sciences tend to focus on interruption behavior and decreased effectiveness, whereas information behavior research is beginning to regard multitasking as a natural and essential coping mechanism that is becoming increasingly obvious with the proliferation of communication devices.

In the fields of information behavior and mobile interaction, there has been work exploring how the fragmented attention caused by a concurrent physical activity could affect search behavior. Harvey and Pointon [12, 13] performed controlled experiments where participants had to complete different search tasks while either sitting down, navigating an obstacle course, and walking on a treadmill. They found that those sitting down were able to generate more accurate and precise search queries than participants in other conditions better search queries and that participants on the treadmill felt the most rushed. Participants navigating the obstacle course were least likely to forget their immediate surroundings, because of the increased need to be aware of their surroundings while navigating the obstacle course. They also tended to be less absorbed in the search tasks. Related work by Oulasvirta et al. [26] shows that smartphone interaction also tends to be reduced to shorter bursts of interaction when the user is occupied with another physical activity. The interaction between smartphone use and physical activity goes in both directions: several studies have shown that performing common smartphone tasks while walking, such as calling or texting, have a negative impact on a person's speed and direction and make people more cautious in their physical activity [21, 22].

In addition to physical activity, Harvey and Pointon [14] also examined the effect of ambient noise on search behavior and satisfaction. They found that noisy environments induce stress on users and make them feel additional time pressure, negatively affecting their search behavior. A similar study was performed by Sarsenbayeva et al. [30], who investigated the effects of different types of ambient noise on typical smartphone usage tasks. They found that ambient noise has a significant effect on interaction and that different ambient noises affect users differently, with some slowing down text entry while others affect task completion time. All of this suggests that the physical act of driving a car and the ambient noise involved are likely to have a negative impact on IPA interaction, its perceived usability, and distraction levels.

3 METHODOLOGY

3.1 Design

We designed our study with two goals in mind. The first one was to compare interaction with Siri while driving to the traditional alternative of manually interacting with an iPhone in terms of usability and distraction (RQ1). This independent variable, *interaction type*, has two conditions—Siri vs. manual interaction—which correspond to the lower-level branches in Figure 1. To ensure that no illegal activity took place, manual iPhone interaction was only allowed after participants had pulled over and parked the car.

In order to determine how the driving process itself influences how people interact with their phones using either Siri or manual control (RQ2), we also manipulated a second independent variable, the *interaction context*, by repeating the same experiments in a controlled laboratory environment. This should give us insights into the question of how generalizable the results of previous usability tests of IPAs are: is there an impact on usability and cognitive load IPA interaction is the driver's secondary focus? This comparison between car and lab is represented by the top branches in Figure 1, with the right-hand side duplicating the conditions on the left in a controlled lab environment.

As shown in Figure 1, a total of twenty-four participants took part in our study. Eight of them used Siri while driving and eight of them controlled their phone manually while in the car. Because of difficulties with recruiting enough participants for each of the four final conditions, we used a within-group design for the laboratory experiments. Here, eight participants were asked to complete tasks using both Siri and manual interaction, with the order of these two conditions randomized.

While the authors' home university only has an Institutional Review Board (IRB) available for medical studies, we had two senior faculty members review the study design and propose changes where necessary to improve the safety of our participants and eliminate any potentially illegal activity.

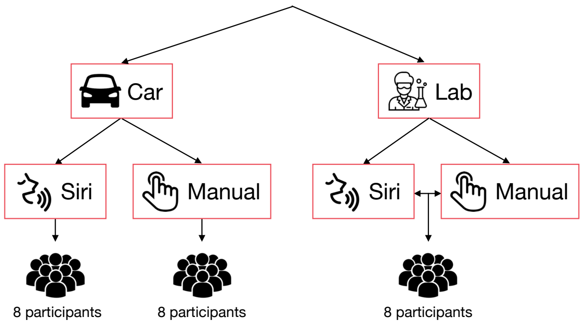

Figure 1: Illustration of the split-plot design with comparisons between interaction type (Siri vs. manual) and interaction context (car vs. lab) and the number of participants in each group at the bottom.

3.2 Participants

We recruited study participants using a combination of convenience and purposive sampling by word-of-mouth and requests on social media. Participants were compensated for their participation with food and beverages. In total, twenty-four participants were selectively recruited to balance gender across the age range of 20 to 38 years old—which represents the segment of the population that most actively uses IPAs according to Bogers et al. [2]. In addition, our participants had to match the following criteria:

- They should have owned a valid driver's licence for at least the past two years
- They should not rate themselves as an inexperienced or insecure driver on a self-assessment scale of driving experience
- They should speak and understand Danish fluently
- They should have experience with driving in the Greater Copenhagen area
- They should own an iPhone (to ensure that usability judgments reflected the quality of the IPA instead of lack of familiarity with the phone's OS)

Our participants consisted of 12 men and 12 women with a mean age of 26 with a range of 23-35 years. They were relatively inexperienced in the use of Siri: only two participants used Siri daily, one used it less than monthly, while 71% (n = 17) had tried out Siri once, but never used it. Only three participants had never interacted with Siri before. Participants had had a driver's license for 8.6 years on average and 75% self-assessed their driving experience as above average. Interestingly, 29% of participants admitted to having used their phone manually while driving before. We anonymized our data and refer to our participants as TP1-TP24 in this paper.

3.3 Materials and Equipment

Access to the Siri IPA was provided using an Apple iPhone 6S with iOS 11.2.6 with language set to Danish. Siri was reset after each experiment. Voice activation of Siri using the phrase "Hey Siri" was enabled for each participant. The iPhone was placed in the car's center console using a car phone mount (see Figures 3a and 2). The car was a 2016 Suzuki Swift, 1200cc manual gear, five-door right-hand drive model. Video was recorded with a GoPro Hero 5 from the inside-top right-hand corner of the windscreen and as point-of-view video from the Pupil Core eye-tracking glasses. We used 200hz binocular head-mounted Pupil Core glasses from Pupil Labs. Data was recorded on a laptop with the Pupil Capture software and analysed with Pupil Player. Mini-QR code markers around the phone were used to aid automatic Area of Interest (AoI) detection (cf. Figure 2). The iPhone screen was recorded using Apple's native screen recording app. Audio was recorded from the GoPro and Pupil Core as well as a Tascam DR-05 dictaphone and the Voice Recorder app on a Samsung S8 smartphone as backup. The experiment manuscript and post-test interview guide were read from an iPad Air. Experimenter 1 was seated next to the driver and Experimenter 2 behind the driver. The experimental setup in the lab is shown in Figure 3b. Perceived workload was assessed using the NASA Task Load Index (NASA-TLX) [11]. A downside of the NASA-TLX is that it is a post-event instrument. Other methods, like the tapping test [1] or the Sternberg Memory Test [33] provide a means of measuring cognitive load during a usability test. However, for safety reasons we did not want to further increase the cognitive load of our drivers, so we used the NASA-TLX instrument.

3.4 Procedure

3.4.1 Pre-test questionnaire. After being welcomed and introduced to the purpose of the study, participants were asked to sign a consent form. Participants who were recruited for the two driving conditions had to show us their driver's license before continuing. Participants then started the study by completing a pre-test questionnaire. This questionnaire consisted of four groups of questions: attitudes towards technology, prior experience with IPAs, driving experience, and demographics.

3.4.2 Controlled experiment. The main element of our study is a controlled experiment where users are subjected to one (or more) of the four conditions: Siri while driving (Car-Siri), manual control while driving (Car-Manual), Siri in the lab (Lab-Siri), and manual control in the lab (Lab-Manual). As explained in Section 3.1, our lab participants were exposed to both Lab-Siri and Lab-Manual conditions in random order. In the two Manual conditions, participants were allowed to use any combination of browser and

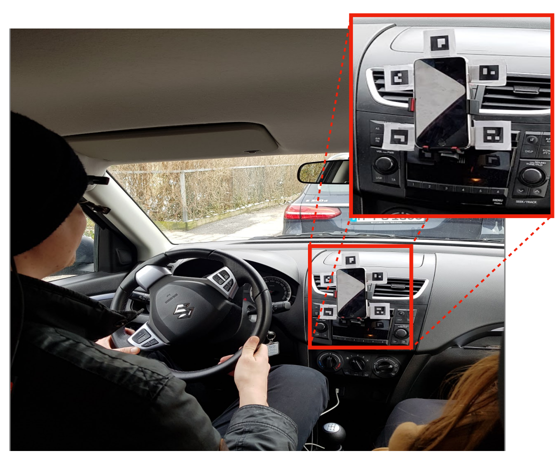

Figure 2: Placement of eye-tracking markers around phone.

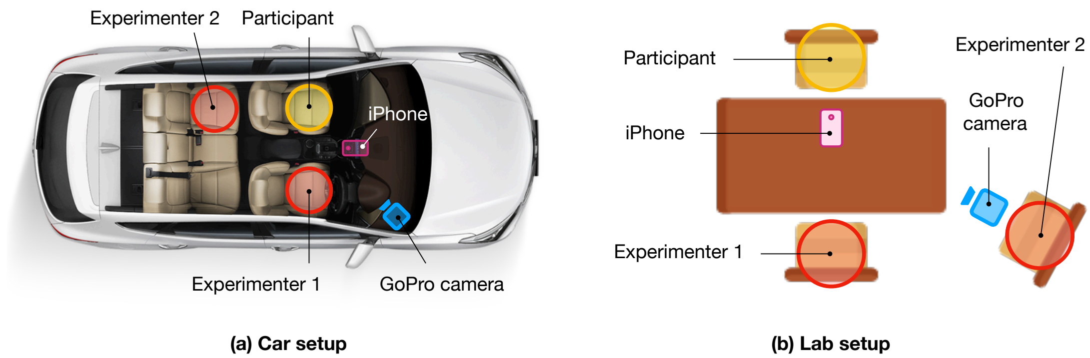

Figure 3: Experimental setup in the car (a) and the lab (b).

apps to solve their tasks, although typically they used Google Maps along with the native Notes, Messages, Weather and Music apps.

At the start of the experiment, all participants had to put on the eye-tracking glasses before starting their tasks so they could be adjusted and calibrated to the eyes of the specific participant. All participants assigned to the Siri conditions were first asked to go through the Siri configuration sequence to adjust it to their voice. They were then given two training tasks: multiplying two numbers and asking for the time in Sydney, Australia. All participants assigned to the driving conditions were reminded to take their time to adjust their seat and mirrors for optimal comfort and safety. We then asked them to drive around for a least five minutes in order to familiarize themselves with the car.All participants were instructed to end up at the same location from which they drove the same 4.3 km route, which included six right-hand turns, six left-hand turns and two 180-degree turns. During the experiment, participants were given turn-by-turn driving directions by Experimenter 1.

After completing their training, each participant was assigned five different tasks in random order, which were read aloud to them by Experimenter 1, one at a time with a break of approx. 1-2 minutes in between each task. Table 1 shows a list of the five tasks. These tasks were partly based on earlier work on IPA usability [24] and voice search log analysis [9], who found that voice queries with narrow information needs were more common for IPA use. We only kept tasks that we considered to be realistic in a driving setting. In the lab conditions, where users were exposed to both the (Lab-Siri) and (Lab-Manual) conditions, we assigned our participants small variations of these tasks where relevant so as to avoid a learning effect. For instance, when they were on their second condition, task 1 featured a different fast-food restaurant and task 5 featured a different song they had to play. All lab-based experiments took place in a meeting room at the authors' university during the same time period as the driving experiments. The lab setup was kept as similar to the car setup as possible. Small-talk was kept to a bare minimum in all four conditions to avoid distractions and to make comparisons between the conditions more reliable.

3.4.3 Post-test interview. After completing the experiment, participants were subjected to a post-test interview, which focused on their satisfaction with the interaction during the experiment and, in the Car-Siri and Lab-Siri conditions, how they themselves perceived Siri's ability to successfully complete the tasks. Finally, we had them retrospectively assess their cognitive workload using the NASA-TLX. All interviews were transcribed and thematically analyzed.

3.4.4 Dependent measures. The goal of our study was to assess interaction with Siri in terms of usability and distraction, because we expected the usability of an IPA to have an impact on the cognitive distractions it induces. We follow the Quality in Use Integrated Measurement (QUIM) model by Seffah et al. [31] in determining how to measure usability and select the following factors: (1) *effectiveness*, or how many errors do users make, how severe are they, and how easily can they recover; (2) *efficiency*, or how quickly can they perform tasks once users have learned the system; (3) *satisfaction*, or how pleasant is it to use the system; and (4), *learnability*, or how easy is it for users to accomplish basic tasks the first time they use the system? The latter is especially relevant, given the relative novelty of IPAs in general and the low percentage of active Siri users in our sample. The fifth factor, safety, corresponds to what we refer to as *distraction*, i.e., to what degree does the system take the attention away from driving? Table 2 has an overview of the metrics and methods used to measure these five factors.

3.4.5 Safety precautions. To ensure the safety of all persons involved in the study as well as the other road users, we took several precautions to eliminate illegal behavior and reduce risks as much as possible. Precautions taken before driving included not recruiting people who self-identified as "inexperienced and/or not comfortable with driving a car" and having our participants perform their

Table 1: List of tasks used in the usability test

ID	Type	Description
1	Directions	You are hungry and want to go to the nearest McDonalds. Use Siri/the phone to help you find it.
2	Note	You suddenly remember that you have to buy cucumbers on your way home. You want to write this in a note so you don't forget this. Use Siri/the phone to do this.
3	Text message	You have received a text message that you want to know the content of and reply that you will be there in 10 minutes. Use Siri/the phone to do this.
4	Weather	You are in doubt whether you should bring a raincoat when you go out tonight. Use Siri/the phone to check the weather.
5	Music	You would like to listen to the song "Billie Jean". Use Siri/the phone to play this song.

Table 2: Overview of the top-level dependent measures, their component measures and which methods they were measured with (QUE = pre-test questionnaire, ET = eye-tracking, VR = screen and video recording, INT = post-test interview).

	Data collection methods			
	QUE	ET	VR	INT
Efficiency				
Task completion time			✓	
No. of attempts at task completion			✓	
No. of steps until task completion			✓	
Effectiveness				
% of tasks completed			✓	
No. of abandoned tasks			✓	
Perceived ability of Siri to complete tasks				✓
Satisfaction				
Facial expressions			✓	
User satisfaction				✓
Learnability				
Prior experience with Siri	✓			
Progression in task completion			✓	
Distraction				
Cognitive workload				✓
% of fixations on phone vs. all fixations		✓		
Facial expressions			✓	

training tasks with Siri *before* they started driving to ensure their full attention on the task, hopefully reducing the cognitive load later on. During their training drive at the start, participants were asked to drive around the neighborhood without any distractions, such as the IPA or the eye-tracking equipment, to help them get used to the car and gain confidence. During the main experiment, participants all drove the same quiet route in Sydhavn during daylight hours outside rush hour on roads with low traffic density and a maximum speed limit of 50 km/h. While driving, the experimenters served an extra pair of eyes and alerted the participants to possible situations they might have missed. Participants were not allowed to touch their mobile phone at any point while driving. In the manual interaction condition, participants were forced to pull over first before being allowed to operate the iPhone, while in the Siri condition any button presses were done by the experimenters.

4 RESULTS & ANALYSIS

4.1 Efficiency

The first dependent measure we focus on is efficiency, i.e., how quickly can the participants perform their tasks in terms of time, steps and attempts. A task completion attempt often consists of multiple steps; whenever a participant had to start a task over by saying "Hey Siri", this was counted as a new attempt. We analyse only completion times and step counts for completed tasks, i.e., tasks that were not abandoned or assessed by us as incomplete. Task completion times are measured from the moment Experimenter 1 finished reading the task aloud until the participant completed the task—either self-assessed or assessed by us when reviewing the video recording. To simulate a realistic and above all legal experience, we included the time needed to find a parking spot in the Car-Manual condition in task completion times.

On average, participants in the Car-Siri condition spent 95.7s completing their tasks (*SD* = 88.1), compared to 36.2s (*SD* = 19.5) in the Car-Manual condition. In the Lab-Siri condition the mean completion time was 30s (*SD* = 13.5) and for Lab-Manual it was only 16.0s (*SD* = 6.8). Unsurprisingly, completing tasks in the car took more time than in the lab, which is due to the participant's attention being elsewhere: on the road environment in the Car-Siri condition and on finding a parking spot in the Car-Manual condition. Completing tasks with Siri while driving was by far the most time-consuming, which took 2.6 times as long as in the Car-Manual condition, which includes the time needed to find a parking spot. In the lab, using Siri only took 1.9 times as much time. Unlike the findings by Luger and Sellen [24, p. 5291], our participants did experience speaking as a faster interaction method.

Figure 4 shows a break-down of task completion times across our five tasks. We observe the same trends at the level of individual tasks, with the tasks of getting directions (#1) and dictating a text message (#3) showing the slowest performance by Car-Siri. This is probably due to the complexity of these two tasks, as they also require more steps in order to be completed, which can be seen in Figure 5a. Nevertheless, Siri's poor hands-free performance on these two tasks is not good news, as these are arguably the most useful hands-free capabilities an IPA should offer while driving.

To get a more complete picture of the effort required to complete a task, we examine not only the number of steps, but also the number of attempts needed before successful task completion. One way of thinking of the number of required attempts is that the more attempts are needed, the less intuitive interaction with Siri was apparently, since participants needed to reformulate their request for Siri to understand it. Common causes of having to attempt a task multiple times included Siri timing out due to the driver being distracted by a traffic situation or the participant interacting with Siri when Siri was not ready. In the majority of cases, participants

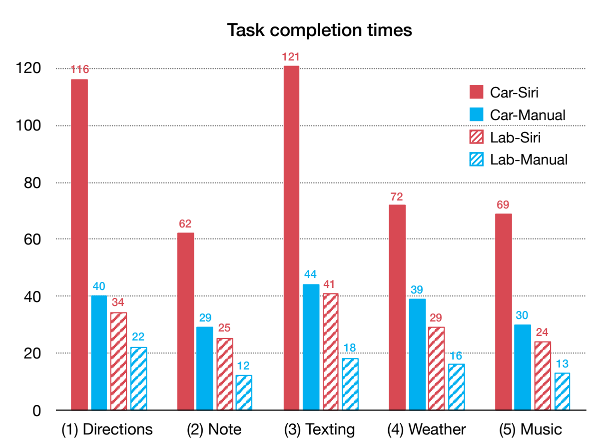

Figure 4: Task completion times (in seconds) for the four different conditions.

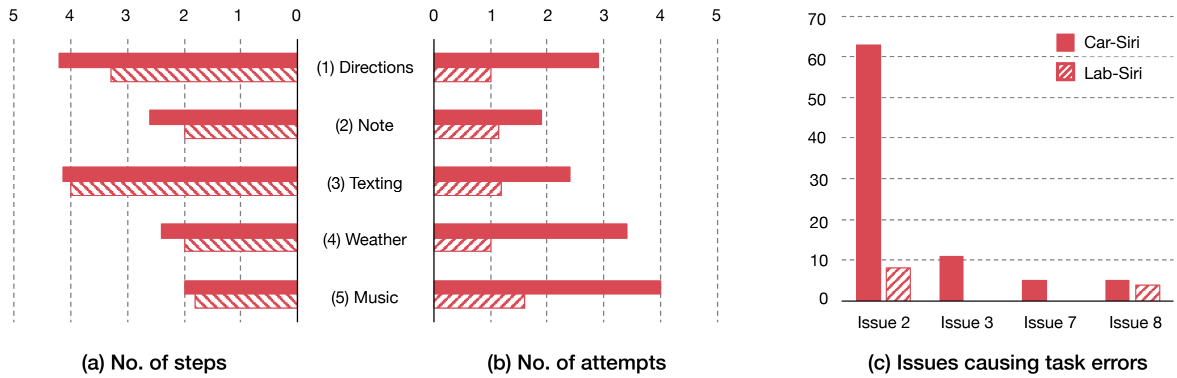

Figure 5: Comparison of Car-Siri and Lab-Siri conditions on (a) number of steps taken to complete the task, (b) number of attempts at completing a task, and (c) distribution of the issues causing errors according to the taxonomy by Cowan et al. [4].

simply repeated their original query, but louder or better articulated. In one other case, a participant had to use two attempts, because they tried to complete a multi-step interaction in a single step, cause Siri to ignore the second step.What is evident from Figures 5a and 5b is that while it appears to be easy to find the right request formulation for some tasks, they can still take many steps to complete. In contrast, other tasks, such as the music task (#5), required many attempts for successful completion even though final completion only required few steps. The note task (#2) appears to have been the easiest to complete with the lowest average step and attempt counts and average completion time. An alternative to saying "Hey Siri" is to press the Siri re-activation button, which only appears after Siri has already been activated. Pressing this button can be used to make Siri listen without having to start an entire session over, for example after Siri's response window has timed out. During the Car-Siri condition, these button presses were always performed by Experimenter 1. We did not include these button presses as explicit attempts, but there were considerably more of them in the Car-Siri at 32 compared to 3 button pressed in the Lab-Siri condition. This suggests they are a meaningful alternative to saying "Hey Siri", even though they are not a legal option for solo drivers.

4.2 Effectiveness

In order to assess the effectiveness of Siri, we analyzed our video and screen recordings for task completion and abandonment, and performed a qualitative error analysis based on transcriptions of the usability tests and post-test interviews. In the Car-Siri condition, with five tasks attempted by each of the eight participants, we recorded a total of 39 task completion attempts (one was lost from the recordings). Of these 39 tasks, 82.0% ($n = 32$) were completed successfully and 17.9% ($n = 7$) were abandoned. The weather (#4) and music (#5) tasks were the ones that were most often abandoned, although the inter-task differences are small because only 7 tasks were abandoned in total.

When we compare these results to the other conditions, we can see clear differences. In the Lab-Siri condition, only one participant had to give up on a task, corresponding to 2.5% of all attempted tasks, which is considerably lower than in the Car-Siri condition. No participants ever had to abandon tasks in the Car-Manual and Lab-Manual conditions, suggesting that using Siri while driving is the least effective solution of all of them, and that effectiveness is also impaired by having to drive a car and navigate the road environment at the same time.

To better understand the reasons for having to attempt a task multiple times or even abandoning a task, we performed an a priori content analysis of their transcribed comments during the controlled experiment and their relevant post-test interview. Our analysis is representative of the experience of novice IPA users, who made up the majority of our participants. As our a priori categories, we relied on the taxonomy of six key usability issues identified by Cowan et al. [4]: (1) issues with supporting hands-free interaction; (2) general problems with speech and accent recognition; (3) problems with controlling third-party apps or services; (4) social embarassment due to public use (not relevant for us due to our setup); (5) the human-like nature of IPAs; and (6) issues related to trust, data privacy, transparency, and ownership. Two additional categories emerged from our analysis: (7) missed window of opportunity, for when responses to Siri timed out; and (8) lack of experience with Siri's capabilities.

Figure 5c shows that four of the eight usability issues plagued interaction with Siri and that it was the use of Siri while driving that most most problematic. Participants in the Car-Siri condition were frustrated by timed-out response windows (#7) and problems controlling Spotify (#3), but the overwhelming majority of problems were due to problems with speech recognition (#2). A particularly large gap was observed between female and male participants due to the severe problems Siri had with recognizing female voices in the car. No gender gap existed in the Lab-Siri condition, but in Car-Siri it affected female participants' performance on a variety of measures: error rate, perceived effort, and completion times. As a result, usability was measured and perceived to be much lower for female participants. We speculate that this may be due to interference between the higher frequency of their vocal register and the background noise(s) the car produces.

4.3 Learnability

We assigned our five tasks in random order to reduce the influence of fatigue or learning effects, so we could better determine whether participant learned how to best interact with Siri during the experiment. However, we could not find any evidence of users getting better or faster at completing their tasks with Siri as the controlled experiment progressed, which suggests that the short-term learnability of Siri is not very high. We also examined whether self-assessed technology adoption behavior and prior experience with IPAs (and Siri in particular) had an influence on participants' performance, but we found no evidence of this.

4.4 Satisfaction

Rubin and Chisnell [29, p. 4] define satisfaction as the absence of frustrations. We have evidence of this from the facial expressions in videos and from the post-test interviews. One participant described the use of Siri as "an emotional roller-coaster", that is, even though it took a lot of effort, there was a certain satisfaction when successful. There were also examples of participants thanking Siri or giving it a thumbs up in the videos. We did not systematically analyse facial expressions due to resource constraints, but the general impression from looking through the videos is that the majority of statements and expressions were negative or neutral. Statements from the post-task interviews indicate that tasks were completed with ease and with no frustrations in the Lab-Manual condition, but that some participants experienced frustrations in Lab-Siri and that most participants experienced frustrations in the Car-Siri condition.

4.5 Distraction

Arguably the most desirable aspect of Siri while driving would be safety: if Siri offers a less distracting experience than manual interaction, that would make it both a legal and preferable experience. Our main reason for using the eye-tracker for data collection was to get a better understanding the level of *visual distraction*: how do participants divide their visual attention between the road environment and the iPhone? One measure of visual attention are the gaze points, which capture what the eye is looking with frequency of 200 Hz. A series of gaze points that are close in time and/or space constitute a fixation, which represent the eyes locking towards an object. By looking at the percentage of gaze points that were directed at the iPhone, we can assess how much time participants spent looking at their phone and how much time on the road environment. In the Car-Siri and Car-Manual conditions our participants looked at the iPhone 10.3% and 9.5% of the time, while in the lab, participants spent 16.9% of their time looking at the phone. These numbers suggest two things: (1) one might expect manual interaction to require more time looking at the phone than voice interaction, but this is not borne out by the evidence; and (2) that the driving task requires more of their visual attention to be directed to the road environment. In addition, in the Car-Siri condition, participants showed more saccades, which indicates that their attention went back and forth more often between the road environment and the iPhone. This was not the case in the Car-Manual condition.

To assess the level of *cognitive* distraction, we used the NASA-TLX instrument, which showed that using Siri imposed a much higher cognitive load. Participants rated their experience in these two conditions as much more cognitively demanding on a scale of 0-100 (where higher is more cognitively demanding) with average scores of $M_{\text{Car-Siri}} = 62.0$ ($SD = 18.0$) and $M_{\text{Lab-Siri}} = 66.0$ ($SD = 20.0$) respectively. Surprisingly, participants in the Lab-Siri conditions found Siri to be slightly more cognitively demanding, although we did not check for statistical significant differences because of the small sample size ($n = 8$). Cognitive load of manual phone interaction was much lower at $M_{\text{Car-Manual}} = 34.0$ ($SD = 8.0$) and $M_{\text{Lab-Manual}} = 40.0$ ($SD = 20.0$) respectively, which is likely due to the large gap in experience with the two interaction types.

The impact of *bio-mechanical* distraction is difficult to assess, because participants were not allowed to touch their phones during any of the driving experiments. However, Siri's response windows timed out frequently in the car, and an experimenter had to press the home button to re-activate Siri. In a real-world scenario this would constitute a bio-mechanical distraction. As for *auditory* distraction, we did not test for this explicitly, but we expect the sound produced by Siri to be at the same level as a regular passenger conversation and to therefore not have a meaningful effect on driving.

5 DISCUSSION & CONCLUSIONS

In this paper we presented the results of a study designed to evaluate how usable and how safe the use of Siri when driving a car, the kind of natural, hands-free settings that IPAs are designed to support. Here, we discuss our findings with regard to influences of interaction type and context on the perceived usability and distraction levels. We also suggest potential design solutions as well as avenues for future work.

5.1 Interaction type

Our first research question focused on the comparison between the use of Siri while driving compared to manual interaction with the smartphone. To avoid unsafe and illegal situations, we adjusted the latter scenario by requiring participants to pull over and park before being allowed to manually operate the phone. Earlier studies have shown that our setup for manual interaction with the iPhone while driving does not represent a 100% realistic scenario: unfortunately, many drivers operate their mobile phone directly while driving instead of pulling over first—something that several of our participants also admitted to doing occasionally. We are aware that this does affect the ecological validity of our comparison to a degree, but not taking this precaution would have been unethical.

Through a mixed methods approach, we found that, as expected perhaps, Siri's usability was considerably lower than manual interaction, partly due to a large experience gap between the two interaction types, but also due to specific usability issues with Siri. Participants in the Car-Siri condition were slower and made more errors and attempts at task completion than in the Car-Manual condition, and they abandoned nearly one in five of their tasks completely. Participants also found that using Siri was more demanding mentally and required more effort to use, which is bad news for an interaction style that is supposed to reduce cognitive distraction. We also found this was worse for less experienced drivers, who are precisely the group that should not be overburdened when driving. In contrast, our male participants with higher than average driving experience and knowledge of the neighborhood in which the driving tests took place, found Siri easier to use and more usable. This suggests that having a mental surplus due to experience could have offset Siri's usability problems. It would be interesting to compare different levels of cognitive load in more detail to determine where Siri becomes an asset as opposed to a distraction.

Apart from the obvious added auditory distraction that comes with conversational interaction, the level of visual distraction did not decrease when using Siri, as too much time was spent on confirming input and reading output messages. In addition, it appears that the human habit of looking at your conversational partner also translates to looking at Siri when interacting with her, at least for novice users—something that applies to the majority of us at this stage in IPAs' technology adoption lifecycle [2]. Based on our findings, we can conclude that using Siri while driving is not a safer and less distracting alternative to manual phone use.

5.2 Interaction context

Through our second research question, we aimed to investigate how interaction with Siri is impacted by the driving process compared to interacting with Siri in a controlled setting where this interaction is the primary and only task for the user. Having to interact with Siri in addition to driving a car certainly seemed have hurt its perceived usability among our participants, which is reflected in considerably longer task completion times as well as the number of abandoned tasks, which was a factor seven higher in the Car-Siri condition. This suggests that Siri has not been adequately adapted to and designed for use in a driving context yet. When comparing manual interaction in different contexts we do not see the same differences: manual interaction was equally effective in both conditions and lower task completion times were mostly due to the need to find parking in the Car-Manual condition, although we cannot rule out that part of the difference was due to a residual cognitive load brought on by driving [34], although we did not test for this.

Another driving-specific issue with Siri is that user utterances are normally displayed on-screen so the user can check them; any results lists produced by Siri in response to a question are typically presented in the same way. However, relying on such visual feedback could lead to serious visual distractions when driving. A possible solution to this could be to read back the user's input and results to them instead of showing it on-screen.

Siri offers a limited temporal window for user responses after being activated using "Hey Siri', which was a source of frustration in the Car-Siri condition. If a participant temporarily had to focus more on the road environment, Siri would often time out, forcing the user to start over. This problem did not exist in the Lab-Siri condition. Allowing for more time to formulate questions and replies so timeouts are less frequent would be a good idea when interacting with Siri while driving. This matches the findings of Hoedemaeker and Neerincx [15], who argue that it is important to properly attune in-car interfaces to the momentary cognitive load spikes that naturally occur while driving. In general, even when not used in the Apple's official CarPlay mode, Siri should be more aware of the context and automatically adapt its interaction and feedback mechanisms when used in a hands-free context.

One of our more peculiar findings was related to a gender gap in the Car-Siri condition: on all dimensions—error rate, perceived effort, completion times—female drivers experienced considerably lower usability of Siri due to the speech recognition problems. While there were no differences between the two genders in the Lab-Siri condition in terms of speech recognition quality, female drivers had the hardest time getting Siri to understand them correctly. We speculate that this may be due to interference between the higher frequency of their vocal register and the background noise(s) the car produces. Despite our best efforts, it is possible this was a specific fluke of our experimental setup, but this potential gender gap in performance and satisfaction merits further study.

Regardless of the interaction context, Siri has several other serious usability problems in Danish, that were also encountered by Bogers et al. [2]. For instance, we also found issues with multilingual speech recognition. Tasks that required the user to include English named entities in their Danish request were prone to being misunderstood by Siri's Danish speech recognition model. An example of this was task 5, where participants had to ask Siri to play the English song "Billie Jean", which the Danish speech recognition model could only recognise when pronounced with a Danish accent. This resulted in increased levels of frustration among participants as it took them more attempts and more time to correct Siri and complete their tasks. Such frustration can easily lead to cognitive distractions and can be dangerous when driving a car. One possible way of improving (multi-lingual) speech recognition for non-English languages could be to train those models in mixed-language settings. This could, for instance, include modeling the pronunciation of the most popular English expressions and named entities (e.g., movies, TV shows, celebrities, and sports teams) to ensure their accurate recognition. Future research should focus on a more controlled evaluation of how tolerant different IPAs are of mixing multiple languages. Evaluating IPAs in languages other than English could also add to our understanding of how usable IPAs are non-Anglo-Saxon parts of the world.

Another serious usability problem of Siri and other IPAs—and one also reported by Bogers et al. [2]—is the limited memory and understanding of context that Siri possesses. Multi-stage tasks are poorly supported by Siri, because it has limited capabilities to remember information from previous interactions or resolve anaphoric expressions referring back to an earlier interaction. This severely impacted user satisfaction and would be worth focusing on in future IPA development.

Finally, some other interesting ideas for future work would be to perform large-scale tests in a driving simulator to achieve a perhaps better trade-off between safety and realism. A test of Apple's CarPlay integration would also be interesting as well as usability testing of IPAs in other hands-free settings.

5.3 Conclusions

Due to all of the aforementioned challenges and limitations, we believe that, while legal, Siri is not yet a safer alternative to manual mobile phone use while driving. It is likely that this also extends to other IPAs, given the small differences in terms of usability between IPAs uncovered in the related work [18, 23]. While Siri may provide a hands-free experience to users, the added distraction and reduced situational awareness means that it is certainly not an eyes-free experience. Given that prior work has shown that there is little difference in terms of safety between using mobile phones handheld or hands-free [16], perhaps it would be most prudent from a safety perspective to outlaw all mobile phone use while driving, regardless of the interaction method. Unfortunately, related work also shows that over time people tend to revert to their old illegal habits [27]. This makes it even more important to make sure that alternatives such as Siri are usable and safe for use.

REFERENCES

[1] M. J. Albers. 2011. Tapping as a Measure of Cognitive Load and Website Usability. In *Proceedings of the 29th ACM International Conference on Design of Communication.* ACM, 25–32.

[2] Toine Bogers, Ammar Ali Abdelrahim Al-Basri, Claes Ostermann Rytlig, Mads Emil Bak Møller, Mette Juhl Rasmussen, Nikita Katrine Bates Michelsen, and Sara Gerling Jørgensen. 2019. A Study of Usage and Usability of Intelligent Personal Assistants in Denmark. In *Proceedings of the 2019 iConference.* Springer, 79–90.

[3] Matthias Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. 2011. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. In *MobileHCI âĂŸ11: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services.* ACM, 47–56.

[4] B. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira. 2017. "What can I help you with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *MobileHCI '17: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services.* ACM, 1–12.

[5] Danish Road Traffic Accident Investigation Board. 2014. *Why do road traffic accidents happen?* Technical Report. Retrieved from http://www.hvu.dk/SiteCollectionDocuments/HVUdec14_UK_HvorforSkerUlykkerne.pdf.

[6] Marco Dozza, Carol A. C. Flannagan, and James R. Sayer. 2015. Real-World Effects of Using a Phone while Driving on Lateral and Longitudinal Control of Vehicles. *Journal of Safety Research* 55 (2015), 81–87.

[7] Frank A. Drews, Monisha Pasupathi, and David L. Strayer. 2008. Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied* 14, 4 (2008), 392.

[8] Marta Perez Garcia, Sarita Saffon Lopez, and Hector Donis. 2018. Everybody is Talking about Voice Activated Virtual Assistants, but how are People really Adopting and Using Them? Lessons from a Multi-Country Study. In *Proceedings of the 32nd Human Computer Interaction Conference.* BCS.

[9] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 35–44.

[10] Joanne L. Harbluk, Y. Ian Noy, Patricia L. Trbovich, and Moshe Eizenman. 2007. An On-Road Assessment of Cognitive Distraction: Impacts on Drivers' Visual Behaviour and Braking Performance. *Accident Analysis and Prevention* 39 (2007), 372–379.

[11] S. G. Hart and L. E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183.

[12] Morgan Harvey and Matthew Pointon. 2017. Perceptions of the Effect of Fragmented Attention on Mobile Web Search Tasks. In *CHIIR '17: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* ACM, New York, NY, USA, 293–296.

[13] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The Effects of Fragmented Attention on Mobile Web Search Tasks. In *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, USA, 155–164.

[14] Morgan Harvey and Matthew Pointon. 2018. Noisy Signals: Understanding the Impact of Auditory Distraction on Web Search Tasks. In *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval.* ACM, New York, NY, USA, 241–244.

[15] Marika Hoedemaeker and Mark Neerincx. 2007. Attuning In-Car User Interfaces to the Momentary Cognitive Load. In *International Conference on Foundations of Augmented Cognition.* Springer, 286–293.

[16] Yoko Ishigami and Raymond M. Klein. 2009. Is a Hands-Free Phone Safer than a Handheld Phone? *Journal of Safety Research* 40, 2 (2009), 157–164.

[17] Brett Kinsella. 2018. Over Half of Smartphone Owners Use Voice Assistants, Siri Leads the Pack. https://voicebot.ai/2018/04/03/over-half-of-smartphone-owners-use-voice-assistants-siri-leads-the-pack/. (2018). Last visited September 10, 2018.

[18] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, USA, 45–54.

[19] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *CHIIR '16: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* ACM, New York, NY, USA, 121–130.

[20] Sheila G. Klauer, Feng Guo, Bruce G. Simons-Morton, Marie Claude Ouimet, Suzanne E. Lee, and Thomas A. Dingus. 2014. Distracted Driving and Risk of Road Crashes among Novice and Experienced Drivers. *New England Journal of Medicine* 370, 1 (2014), 54–59.

[21] Eric M. Lamberg and Lisa M. Muratori. 2012. Cell Phones Change the Way we Walk. *Gait & Posture* 35, 4 (2012), 688–690.

[22] Sammy Licence, Robynne Smith, Miranda P. McGuigan, and Conrad P. Earnest. 2015. Gait Pattern Alterations during Walking, Texting and Walking and Texting during Cognitively Distractive Tasks while Negotiating Common Pedestrian Obstacles. *PLoS one* 10, 7 (2015), e0133281.

[23] Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-based Natural User Interfaces. In *International Conference on Applied Human Factors and Ergonomics.* Springer, 241–250.

[24] E. Luger and A. Sellen. 2016. "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *CHI âĂŸ16. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 5286–5297.

[25] Luis Nunes and Miguel Angel Recarte. 2002. Cognitive Demands of Hands-Free-Phone Conversation while Driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 5, 2 (2002), 133–144.

[26] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA, 919–928.

[27] Oscar Oviedo-Trespalacios. 2018. Getting Away with Texting: Behavioural Adaptation of Drivers Engaging in Visual-Manual Tasks while Driving. *Transportation Research Part A: Policy and Practice* 116 (2018), 112–121.

[28] T. A. Ranney, W. R. Garrott, and M. J. Goodman. 2001. *NHTSA Driver Distraction Research: Past, Present, and Future.* Technical Report 2001-06-0177. Society of Automotive Engineers. SAE Technical Paper.

[29] Jeff Rubin and Dana Chisnell. 2008. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (2nd ed.). Wiley.

[30] Zhanna Sarsenbayeva, Niels van Berkel, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2018. Effect of Distinct Ambient Noise Types on Mobile Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2, Article 82 (2018), 23 pages.

[31] Ahmed Seffah, Mohammad Donyaee, Rex B. Kline, and Harkirat K. Padda. 2006. Usability Measurement and Metrics: A Consolidated Model. *Software Quality Journal* 14, 2 (2006), 159–178.

[32] Amanda Spink, C. Cole, and M. Waller. 2008. Multitasking Behavior. *Annual Review of Information Science and Technology* 41, 1 (2008), 93–118.

[33] S. Sternberg. 1966. High-Speed Scanning in Human Memory. *Science* 153, 3736 (1966), 652–654.

[34] David L. Strayer, Joel M. Cooper, Jonna Turrill, James R. Coleman, and Rachel J. Hopman. 2017. The Smartphone and the Driver's Cognitive Workload: A Comparison of Apple, Google, and Microsoft's Intelligent Personal Assistants. *Canadian Journal of Experimental Psychology* 71, 2 (2017), 93.

[35] Paul J. Treffner and Rod Barrett. 2004. Hands-Free Mobile Phone Speech while Driving Degrades Coordination and Control. *Transportation Research Part F: Traffic Psychology and Behaviour* 7, 4-5 (2004), 229–246.