Self-Driving Cars Should Use an Assertive Voice to Grab a Distracted Driver's Attention

Priscilla N. Y. Wong[1], Duncan P. Brumby[1], Harsha Vardhan Ramesh Babu[2], Kota Kobayashi[2]
[1]University College London, [2] ustwo London

ABSTRACT

Automated driving will mean that people can engage in other activities and an important concern will be how to alert the driver to critical events that require their intervention. This study evaluates how various levels of assertiveness of voice command in a semi-AV and different degrees of immersion of a non-driving task may affect people's attention on the road. In a simulated set-up, 20 participants were required to execute actions on the steering wheel when a voice command was given while playing a mobile game. Regardless of how immersed the driver was in the game, a more assertive voice resulted in faster reaction time to the instructions and was perceived as more urgent than a less assertive voice. Automotive systems should use an assertive voice to effectively grab people's attention. This is effective even when they are engaged in an immersive secondary task.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **User interface design**; *Sound-based input / output*; • **Computer systems organization** → *Robotics.*

KEYWORDS

Autonomous vehicles, voice assistant, Immersion, Assertiveness

ACM Reference Format:
Anonymous Author(s). 2019. Self-Driving Cars Should Use an Assertive Voice to Grab a Distracted Driver's Attention. In *Automotive User Interfaces and Interactive Vehicular Applications, September 22–25, 2019, Utrecht, Netherlands.* ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

With autonomous vehicles (AVs) becoming more and more advanced, on-road tests with AVs have increasingly been carried out. At this stage of development, the public is still not confident that AVs are as reliable as human drivers [11]. This belief is even further accentuated by recent fatal accidents. For example, an Uber AV killed a pedestrian in Tempe, Arizona in 2018 [29]. The footage of that accident showed that at the moment of the accident the human driver was not paying attention to the road and missed important cues that the autonomous system had failed because they were immersed in using their smartphone instead [29]. This highlights that current AV systems lack sufficient feedback to let drivers know about its state and the appropriate actions that they should engage in (i.e., stay attended to the road). Therefore, we are interested in how to alert drivers to events that require their input and the means to effectively grab their attention.

Semi-AVs, vehicles that are autonomous in some parts of the road and manual in other parts e.g., the Tesla's Enhanced Autopilot [9], are suggested in the industry and in the literature that they should have pre-alerts installed in them. One type of pre-alert is handover requests which takes place when the vehicle is transitioning from autonomous to manual driving or vice versa for safety-critical situations. Most studies about handover requests therefore focused on when and how these requests should be given for drivers to smoothly disengage with secondary non-driving tasks and engage with primary driving task [28, 32]. However, these handover requests do not play a role in informing drivers about a problem that the system cannot pick up e.g., the disabled emergency braking system in the Tempe AZ Uber accident [29]. So in a case of a 'malfunction', we should not simply rely on these requests. Therefore, this study suggests that it might be useful if automated cars can also give more frequent updates about lower-level hazards so that drivers may stay alert to their general surroundings [16].

The aim of this study is to prevent people from being complacent about automated systems by using the concept of voice commands. The idea of drivers being informed by verbal messages is not a novel one. Navigation systems have been around for decades to direct drivers on roads. More recent research explored different variations of voices that deviate from the conventional monotone voices as it was suggested that people are sensitive to the slightest changes in acoustic elements in speech [10, 28]. The current study therefore asks what kind of voice a vehicle should have to effectively grab drivers' attention. This leads to the research questions: Do drivers react differently when they perceive a voice command differently? Does a more immersive secondary task influence people's reaction to the voice commands? Using a simple simulated set-up, we investigate how

AutoUI '19, September 22–25, 2019, Utrecht, Netherlands

the nature of a voice command and a secondary task may impact on people's reaction times and perceptions such as sense of urgency. This is done by presenting voice commands which vary in their level of assertiveness while drivers are immersed in a secondary task to different extents.

The following sections of the paper first reviews the literature related to voice commands and why varying assertiveness in them may impact on people's attention. After describing in detail how a driving simulator study is set up, we present the results that address the research questions. This involves the analyses of people's reaction times and accuracy in reaction to the voice commands which are varied in their level of assertiveness. Their various perceptions of the voice commands are being studied as well. The reactions of participants who engage in different immersive tasks are also compared. The findings are then discussed in relation to the relevant literature and theories, the study's limitation and the implications on design and future work.

2 RELATED WORK

In-car Voice Alerts

In recent years, researchers and developers have been exploring how conversational agents can be incorporated into the in-car system e.g. Android Auto and Apple Carplay [24]. These systems are verbally activated systems that 'listen' and respond to people's instructions to carry out, for example, telematics and infotainment related tasks. They are essentially a built-in virtual assistant such as Siri in a car. A core concern for the development has been on improving the communication between the in-car voice assistants and drivers.

Multitasking in driving is difficult because it stresses people's cognitive workload which has limited resources [5, 7, 34]. As people try to interleave between tasks such as driving and texting, they would encounter dual task interference. This means that as people are trying to maintain the performance of one task, it would affect their performance of another on-going task. Hence, research about in-car voice assistants is important as they can act as mediators to aid a smoother transition between driving and non-driving tasks.

Researchers believe that better interactions between drivers and voice assistants may provide a safer driving environment. For example, Iqbal et al.'s study [13] showed that an alert that warns drivers of critical road situations was effective in reducing people's driving errors in such as turning and chances of collision while drivers were talking on the phone at the same time. However, despite the success of reducing errors, Iqbal et al. acknowledged that there is a tradeoff with the quality of non-driving tasks as conversing on a cell phone became more difficult. Moreover, loading people with a distraction task in a simulated semi-AV environment, Politis et al. [28] investigated how audio warning alone or in combination with visual and tactile cues affected drivers' handover time. They showed that voice commands in combination with other cues led to better driving performance (i.e., less lateral deviation) after handover than voice commands alone. The present study decide to utilise voice commands as alerting tools as evidences showed that voice commands in various forms are considerably effective in drawing people's attention to their driving.

There is an increasing number of research that focus on voice assistants in semi-AVs. Unfortunately, most research did not explore systems that help people prioritise their attention on the road in preparation for emergency situations. In other words, drivers are often put in a passive position in waiting for the vehicle to warn them of emergency situations. For example, in Politis et al.'s study [28], participants were told that they could engage with a 'secondary task' freely unless a warning was given. However, the unexpected always happens very suddenly and cannot be prepared for in advance e.g. a pedestrian rushing out from the side. Thus, despite being occupied by a secondary task, drivers have the responsibility to understand the road situation and intervene the vehicle at any time [8]. Therefore, the current study explores an alternative approach which provides frequent alerts informing participants of low level hazards which may potentially help them stay attended constantly.

Getting constant updates from conversational agent was previously explored by Koo et al. [16]. Unlike many previous studies in the literature that explored the conventional semi-AV that switches between automated and manual driving, Koo et al. [16] studied conversational agent in a semi-AV that had an automatic braking system. It is a system where the car interrupts participants' driving activity by braking automatically. They found that by informing a combination of simple messages such as "The car is braking" (information about the action of the car) and "Obstacle ahead" (information about the reason of the action) increased driving performance e.g. less collisions, speeding, road sign and red-light misses etc. Despite the fact that drivers might have been overloaded with information which resulted in anxiety, providing reasons for the action was nonetheless beneficial for drivers. Alerts that provide reasons for the vehicles' actions have not been explored in the context of semi-AV systems i.e., systems where the vehicle drives itself unless human interferes. This study incorporates this concept into these systems to help drivers better understand and be more aware of their road surroundings. In this case, as the purpose of the alerts was to raise people's attention, anxiety which was seen as unwanted in the previous study [16] may help people stay focused on the road and prevent them from being complacent about the autonomous system.

Anthropomorphism and Assertiveness in Voices

There is an increase in tendency in research to apply anthropomorphism to recent technological developments. Conversational agents are no exception. In fact, speech plays a crucial role in human lives - it is a distinctive identification [4] and fundamental and unique way of communication [22] by using languages [25] for humans. People tend to automatically make attributions related to human-to-human interactions, e.g. genders and personalities, to voices even those that are from machines [23]. It shows that it is human's natural instinct to make use of cues in speeches to make sense of the world and to formulate their reactions and behaviours accordingly [1, 27, 30]. However, traditionally, in-car voices such as navigation and verbal alerts are straightforward, rigid and non-anthropomorphic, e.g., those in Koo et al.'s study [16]. We believe that by eliciting personality in conversational agents in cars may therefore help drivers attend to the road better.

It was shown that the concept of assertiveness is an effective way in delivering verbal messages in the literature. Large and Burnett [17] studied people's ratings on various navigation voices that were differed in gender and identities and were readily available on the market, including the traditional TomTom British female and male voices, Snoop Dogg and Yoda. It was found that people's likelihood to choose a navigation voice for everyday use was correlated with the assertiveness of the voice. The more assertive participants found the voice, the more likely they were to choose it as an everyday navigation. The positive association between assertiveness and trustworthiness suggested that people might have preferred the more assertive voice because they find it more trustworthy. However, the wording of the messages was not varied in a way that the assertiveness of the voices was controlled for whereas Shechtman [31] believed that words for example "needed" and "must" make messages sound more assertive. Also, Large and Burnett did not take direct measures such as react time and accuracy but only self-reported perception of the voices. Therefore, their findings were unable to show how much the voices were able to capture participants' attention. Nonetheless, Large and Burnett's finding [17] that assertiveness in a voice may affect people's choices for a voice assistant is a useful piece of information in this study. It is because assertive voice commands may potentially be an effective tool in drawing people's attention. Therefore, we are interested in exploring the concept of assertiveness further by carefully manipulating assertiveness in voices and taking direct measures of people's driving behaviour.

It is possible, however, that the effectiveness of assertiveness might be context-dependent. When Large et al. [18] explored a more diverse variety of commands and conversational exchanges, they found that participants took a polite turn-taking approach and expected the in-car conversational agent to do the same. It seems like depends on the context and the type of information, the drivers have different preferences in the agent's conversational style. It might be that participants prefer a more assertive voice over a polite non-assertive one when they are simply following instructions e.g., navigation directions, but not when the agent takes on more responsibilities and engage in conversations, for instance, giving reminders (e.g. time of a meeting), suggestions (e.g. music) and asking questions about the driver e.g. his/her interest or first name. Taking into account that the preference for assertiveness might be context-specific and difficult to control, this study therefore focuses on exploring voice commands instead of conversations.

Further exploration into the concept of assertiveness demonstrated greater insights into how and why assertiveness may affect one's perception and in turn their behaviours. van der Heiden et al. [32] explored 'assertiveness' in handover requests through increasing the intensity of audio pulses. The type of pulses explored were no audio pulses, three consecutive beeps evenly spread over time and the increasing number of beeps over time. It was found that the beeps that gradually increase in frequency was able to capture driver's attention to the road the most and resulted in the highest sense of urgency. This finding showed that it is possible that people reacted quicker due to the underlying concept of urgency in the pulses. We believe that this effect of urgency is also present in language-based voice commands with the complex elements in languages.

People's perceived sense of urgency was previously shown effective in influencing people's attention on the road. It was suggested that certain words (e.g., "Danger") convey a stronger sense of urgency [3] and lead to faster reaction time in simulated driving [2] than others (e.g., "Warning", "Caution" and "Notice"). Politis et al. [28] adopted the wordings from Baldwin and Moore [3] and investigated multimodal voice commands including audio, visual and tactile cues in handover situations in semi-autonomous contexts. They found that multimodal warnings were more effective i.e., leading to faster handover time, and were perceived as more urgent but more annoying than unimodal ones with visual alone being the least effective. Consistent with Edworthy et al. [10], urgently spoken voice commands were rated more urgent and led to faster transition than non-urgent warnings. Therefore, the manipulation in the wordings and tones was shown effective in influencing people's behaviours when handling driving related matter. However, only sense of urgency alone has been extensively explored in the literature. A direct relationship between assertiveness and urgency in voice commands has not been established before. Therefore,

through exploring assertiveness in the present study, we believe that it might help us understand how the different nature of voices may impact on people's underlying perception of the voices specifically sense of urgency and in turn provide explanation on their behaviours.

Immersive Secondary Task and Assertiveness

It was long known that secondary tasks affect driving performances e.g. lateral deviations [6, 7] and that people interleave between tasks at 'chunk boundaries' which are natural breakpoints of the secondary tasks [7]. But as the development of automation advances, the boundaries between a primary and secondary task has started to become blurred i.e., driving might be seen as the secondary task and non-driving task as primary task now before handover. How people interleave between task in a semi-AV or fully-AV has become more complicated. Note that by convention we still refer driving as the primary task and non-driving tasks the secondary tasks.

This was suggested that interleaving behaviour in semi-AV was particularly influenced by the nature of the secondary task. In this case, it is the amount of time needed for drivers to deactivate the autopilot mode when a hand-over request is given. Petermann-Stock et al., [26] showed that engaging in a cognitively, visually and motorically demanding task resulted in the longest handover time, consistent with several other studies that examined people on similar mentally demanding tasks e.g. a mobile quiz game [12, 19, 21, 36]. Moreover, Vogelpohl et al. [33] found that distracted drivers' attention i.e., gazes towards side mirrors and dashboard, was regained significantly slower from secondary task compared to non-distracted drivers. Therefore, non-driving tasks which significantly shift people's mental engagement from driving to the task experience seem to affect people's resumption of the primary driving task. One explanation is that because people are so immersed and intrinsically motivated to engage in the non-driving task, more effort and time are needed to unwillingly terminate the activity [33].

Jennett et al. [14] quantified this immersive experience, the state of high engagement and the feeling of being "in the media environment" and suggested that it can exist to different degrees. For example, Wong et al. [35] showed that film media was less immersive than gameplay footages followed by actively interactive games. Therefore, the effort in shifting in and out of the media environment may vary depend on how immersive the task is and how motivated people are to continue to interact with it. To our best knowledge, immersion has not been directly manipulated in the literature in the context of automated driving. Additionally, it was suggested that older drivers of the age between 55 and 73 resulted in better driving performance i.e., less accidents when they listened to a voice assistant that people found more authoritative of than a less authoritative one [15]. It is possible that a more assertive voice has a stronger ability to help people maintain focus on the road. Therefore, this study is not only interested in observing the effect of assertiveness of the voice commands on people's behaviours i.e., reaction time and accuracy in response to the instructions in voice commands, but also how the level of immersion in the secondary task may impact on their interleaving behaviours.

Goals and Hypotheses

This study aims to explore the effects of assertiveness in voice commands and the level of immersion in secondary task on driver's interleaving behaviour and their perception of the voices. In a simulated automated driving set up, participants are asked to follow the instructions given in the voice commands to execute actions on the 'vehicle' while playing a mobile game. The voice commands, varying in their level of assertiveness, instruct participants to perform actions on the brakes and the indicators upon the encounters of low-level hazards. The games, either immersive or non-immersive, acted as the secondary task. Reaction time to voice commands, accuracy in following the instruction given by the voice command and perceptions and feelings elicited by voices e.g., preference, urgency and annoyance, are measured.

We propose three main predictions in this study in regard to the effect of assertiveness of the voice commands and the level of immersion of the mobile games. First, we predict that higher assertive voice commands will result in faster reaction time, higher accuracy, urgency and preference than lower assertive voice commands. This hypothesis is formulated based on the prior work that demonstrated how assertiveness may potentially convey high level of urgency and therefore affect people to react quicker to the voices [2, 28]. Second, a more immersive secondary task will result in slower reaction time and lower accuracy in response to voice commands than a less immersive task. This is expected as the previous studies suggested that the more cognitively loaded one is in a secondary task, the longer it will take for people to process other information at the same time [33]. Third, however, with a higher assertive voice, there will be no difference in reaction time and accuracy in following verbal instructions between a more immersive task and a less immersive task. This is because voices that sounded similar in nature as assertiveness were shown to be able to draw people's attention more effectively.

3 METHOD

Participants

Twenty drivers were recruited through opportunistic sampling (12 males and 8 females). The age range was from 21 to

48 years old ($M = 26.30, SD = 7.34$). Six people usually drove in the UK and others mostly drove in their home countries e.g. Poland, America, Canada and China.

Design

A 2 × 2 (Game × Assertiveness) mixed factorial design was carried out. The between subject variable is how immersive the mobile games were. Two games were selected from an initial manipulation check where one game was of significantly more immersive than the other. The within-subject variable is the level of assertiveness of the voice commands which was determined by their wordings and tones that were also previously explored in the manipulation check. Phrases in the two conditions are significantly different in their level of assertiveness. For example, "Please", "suggest" and "if possible" were used in the lower assertive voice commands which were said with a pleasant tone, and "need", "Watch out!" and "immediately" were used in the higher assertive voice commands which were said with a serious tone.

The dependent variables were participant's response time and accuracy of their response to the voice commands, their preferences for, perceptions on and feelings about the voices. Perceptions and feelings include participants' perceived sense of urgency, distraction from the game, trustworthiness, annoyance, clarity and anthropomorphism.

Materials

Primary Task. A set of different voice recordings was previously tested in a manipulation check for their level of assertiveness in different scenarios. It was recorded with a British male voice. Each voice command consists of a combination of a *scenario command* (i.e., information about the road situation) and an *execution command* (i.e., instruction for required action). An example for a scenario command is "*Beware of T-junction ahead.*" Table 1 illustrated all the execution commands used. Scenario commands varied in tone while execution commands varied in tone as well as wording. The tone was varied so they have different level of seriousness and wordings were varied according to Shechtman's manipulation of assertiveness [31].

A driving simulator was set up using the Logitech G25 racing wheel which include pedals and a shifter unit and a 31" Dell 3007 wfp monitor (Refer to Figure 1 for driving simulator set-up).

Four unique driving videos which were between two and a half minute and five and a half minutes were used. Each of them had six scenarios which were alerted with an appropriate voice command. Using different nature of commands, two different versions were created from each video e.g. "*Exiting roundabout ahead. Indicate left if possible.*" (non-assertive with a pleasant tone) and "Exiting roundabout ahead. Look up! Action to indicate left is needed." (assertive with a serious tone). Therefore, there were eight videos in total.

	Non-assertive	Assertive	t
Indicate Left(L)/ Right(R)	*Indicate L/R if possible.*	*Look up! Action to indicate L/R is needed.*	6.18*
Braking	*Please apply the brakes.*	*Watch out! Brake immediately.*	3.08*
Slow Down	*I suggest you slow down gradually.*	*You need to slow down immediately.*	5.04*

Table 1: Execution Commands - Significantly Different in Their Level of Assertiveness
Note: * indicates $p < 0.05$, $df = 14$.

Figure 1: Driving Simulator Set-Up

A voice rating sheet that was developed by Large and Burnett [17] was used with an addition of the rating of urgency (See Table 3 for the complete questionnaire). First set of questions asked participants their perceptions on and feelings about the voice commands i.e., "Do you think that this voice is ... ?" (Q1) following with "Clear", "Distracting from the game", "Trustworthy", "Assertive", "Friendly", "Annoying", "Entertaining" and "Urgent". Moreover, "Does this voice make it feel like there is somebody with you?" (Q2) measures the anthropomorphism of the voices. Participants' preference for the voices were measured with two more specific questions "How likely would you be to use this as your everyday car assistant voice?" (Q3) and "How likely would you be to use this on a one-off occasion such as a day-out?" (Q4) and finally "What is your overall rating of this voice?" (Q5). This study decided to use the rating on Q5 as the measure of the preference for the voices. The higher the participants

rated on the question, the more they preferred the voice. All questions in the voice rating questionnaire (VRQ) are measured on a 7-point Likert Scale with 1 being not at all and 7 being completely.

(1) Do you think that this voice is...?
  - Clear
  - Distracting from the game
  - Trustworthy
  - Assertive
  - Friendly
  - Annoying
  - Entertaining
  - Urgent
(2) Does this voice make it feel like there is somebody with you?
(3) How likely would you be to use this as your everyday car assistant voice?
(4) How likely would you be to use this on a one-off occasion such as a day-out?
(5) What is your overall rating of this voice?

Table 2: Voice Rating Questionnaire (7-point Likert Scale)

Secondary Task. Two mobile games, *Fruit Ninja* and *Smart Shapes*, were selected due to their significant difference in people's level of immersion in the previous manipulation check. Immersion was measured using the Immersive Experience Questionnaire (IEQ) developed by Jennett et al. [14]. *Fruit Ninja* resulted in a significantly higher immersion than *Smart Shapes*. As a secondary task in the experiment, the mobile games were played on an iPhone 7 plus. *Fruit Ninja* is a mobile game that involves players to slice up fruits that randomly appear on the screen by swiping with their fingers. Players have to avoid slicing up bombs which are traps. The game ends when three misses or mistakes have taken place. *Smart Shapes* is a kid's game that help them learn the organisation of shapes, colours and sizes. Players have to move floating blocks to holes that match with the blocks' property.

Procedure

Participants were seated in a lab room and were instructed to give their consent in participating in the study followed by their basic demographic information. They were then told that the set-up they were sitting in was a simulation of a automated driving environment and they were only required to operate on one of the brake pedal and the indicators on the steering wheel.

Participants were told that they were the drivers of this automated vehicle and that even though the car was on autopilot mode, they still had to manually execute actions with the indicators and the brake. They were told that voice reminders would be given prior to the need of the actions to assist the executions. They were then proceeded to the practice trial where participants were allowed to familiarized with the set up with a one and a half minute video which consists of commands for all the actions i.e., indicate left and right, brake and slow down. Participants were instructed to carry out the action consistent to the commands. Note that the commands used in the practice trial were different from the actual study. Participants were also introduced with the secondary task at the same time. Half of the participants received the higher immersive game (*Fruit Ninja*) and the other half received the lower immersive game (*Smart Shapes*). They were given time to play with the game until they understood its rules. Participants were then asked if they understood the tasks and had any questions before they proceed to the main task.

In the main experiment, participants were required to perform four trials with each trial presenting a unique driving scenario. There were six voice commands in each video. The voice commands in half of the videos were assertive and those in the other half were non-assertive. The order of the videos and the assertiveness conditions were counterbalanced across participants.

During the video, participants were to act accordingly to the instruction given by the voice commands while playing their assigned mobile game. For example, if they hear "*Indicate left if possible.*", they would have to respond by pushing onto the left indicator. Reaction time and accuracy in response to the voice commands on the indicator and the brake were recorded. At the end of each trial, participants were asked to fill in a VRQ which consists of measures such as sense of urgency and annoyance to voice commands.

4 RESULT

Data Filtering and Analysis

Reaction times to voice commands were recorded as every first gamepad response after the onset of a voice command. Care was taken to set the start time to the beginning of the utterance of the instruction in the voice command e.g. "*left*" in "*Exiting roundabout head. Indicate left if possible.*" This ensures that the reaction times across different videos were standardized. Accuracy was a measure of whether the keys on the gamepad pressed matches with the action described in the voice command. Accurate responses were coded with 1 and inaccurate response with 0. Missing responses were treated as inaccurate.

A 2 × 2 (Immersion × Assertiveness) mixed factorial ANOVA was conducted on both participants' reaction time and accuracy. A repeated measures ANOVA was also used to evaluate people's survey ratings based on the assertiveness of the voice commands to determine if participants perceive them differently. Effects with a p value < .05 were deemed as significant.

Assertiveness

A significant main effect of Assertiveness on reaction time was found, $F(1, 18) = 13.95, p = .002, \eta_p^2 = .437$. It can be seen in Figure 2 that assertive voice commands resulted in faster reaction time than non-assertive voice commands. However, no significant main effect of assertiveness on accuracy was found, $F(1, 18) = 3.06, p = .098, \eta_p^2 = .145$.

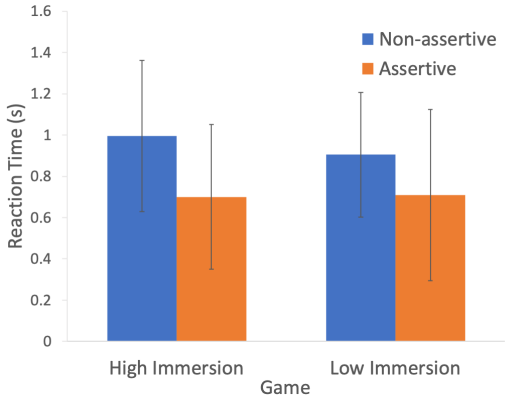

Figure 2: Reaction Time for Different Levels of Assertiveness and Immersion. The error bars represent the standard deviation of the means.

Immersion in Mobile Games

There was no significant difference in reaction time between more immersive condition (Fruit Ninja) and less immersive condition (Smart Shape), $F(1, 18) = 0.075, p = .787, \eta_{p2} = .004$. Also, no main effect of immersion was found in accuracy, $F(1, 18) = 0.689, p = .417, \eta_{p2} = .037$. Further analysis found no immersion × assertive interaction in reaction time, $F(1, 18) = 0.567, p = .461, \eta_{p2} = .031$, nor in accuracy, $F(1, 18) = 0.387, p = .387, \eta_{p2} = .042$.

Voice Rating Questionnaire

The observations of the means of relevant survey ratings in Figure 3 and the result from statistical analyses shown in Table 3 suggested that except for the ratings of urgency and distraction from secondary task, there was little difference between assertive and non-assertive conditions in the subjective ratings.

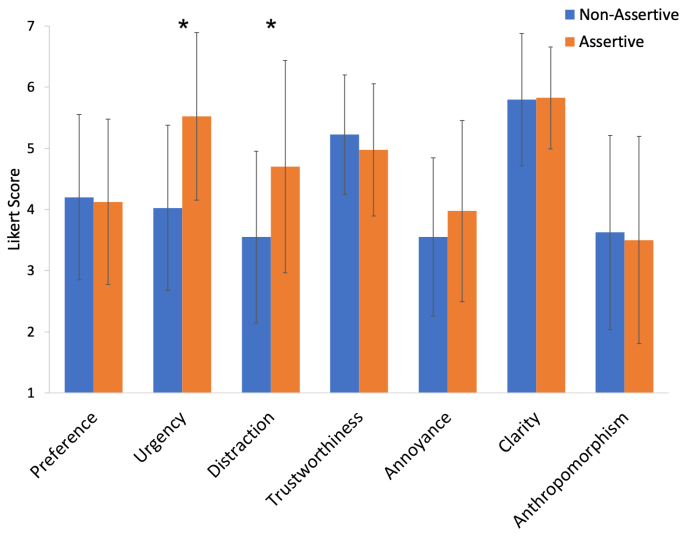

Figure 3: Likert Scale Ratings of Voice Command Related Questions. The error bars represent the standard deviations of the means.
Note: * indicates $p < 0.05$

Urgency and Distraction from Secondary Task

Significant main effects were found for the sense of urgency and people's distraction from the games. Assertive voice commands were perceived to be more urgent and more distracting from their game than that the non-assertive ones (See Figure 3).

Preference, Trustworthiness, Annoyance and Anthropomorphism

No significant difference between the assertive conditions in preference, trustworthiness, annoyance and whether or not the voice felt like a companion.

	$F(1, 19)$	p	η_{p2}
Preference	0.048	.829	.003
Urgency	11.18	.003*	.370
Distraction from Game	10.35	.005*	.353
Trustworthiness	0.446	.512	.023
Annoyance	1.95	.178	.093
Clarity	.008	.928	.000
Anthropomorphism	0.180	.676	.009

Table 3: Repeated Measures ANOVA Results for the Ratings in VRQ
Note: * indicates $p < 0.05$, $df = 14$.

Duration of Voice Commands

The durations of the voice commands were compared between the two assertiveness conditions using a one-way ANOVA in attempt to understand how this acoustic element is different between the assertive and non-assertive voice commands. A significant difference in their length was found, $F(1, 46) = 12.15, p = .001$, $\eta_p^2 = .209$. The more assertive voice commands (M = 4.40 seconds, SD = 0.711) were significantly longer than the less assertive ones (M = 3.68 seconds, SD = 0.733).

5 DISCUSSION

This study aimed to understand the effects of assertiveness in voice commands and immersion of a non-driving task on people's attention in a automated driving environment. A higher assertive voice resulted in a faster reaction time and a higher sense of urgency than lower assertive voice as hypothesized. However, it did not result in a higher accuracy in following the instructions given by the command nor was it more preferred. Our prediction that a more immersive secondary task would delay reaction time and result in lower accuracy was also not supported by our findings. The interaction that we predicted was also not found in our results. Not only was there no difference in reaction times and accuracies between the higher immersion task and the lower immersion task in the higher assertiveness condition, but also in the lower assertiveness condition. It appears that regardless of the level of immersion in the non-driving task, people responded to the respective natures of the commands equally as quickly. We believe that the results can be interpreted in two different directions: assertiveness in voice commands can effectively draw driver's attention from any non-driving task or any non-driving task regardless of how engaging it is may be equally as detrimental to people's attention and response to road environments.

Our results demonstrated that how the different nature in the voice commands regarding their assertiveness had an effect on people's reaction time and perception of the urgency in the voices. The ability for assertive voice to attract people's attention was reinforced by the fact that the voice was able to distract participants more from the games than non-assertive ones. We proposed that assertiveness in voice commands might be more effective in drawing people's attention due to the sense of urgency that it conveys. Though van der Heiden et al. [32] and Politis et al. [28] did not directly investigated assertive voice commands, the present results were consistent with their findings of reaction time where the higher the sense of urgency in the alert, the quicker the people responded to the requests. While Large et al. [17] studied assertiveness in navigation voices, reaction time was not measured in their study. This study therefore provided a novel finding where not only did assertiveness affect people's psychological perception of the situation i.e., urgency, but it also influenced people's actual physical reaction i.e., reaction time.

Similar to Politis et al. [28] and Edworthy et al.'s studies [10], we manipulated the tone of the voices. In line with Politis et al.'s findings [28], we successfully showed that both tones and wordings are important in determining people's perception and in turn their behaviours. Though Politis et al. [28] and the present study explored tones that were based on different foundations i.e., urgency versus serious tones, both findings obtained a difference in urgency. It is possible that there are commonalities present between the natures of the tones which subsequently led to a similar outcome. Future studies may examine how the different kinds of tones may have overlapping for example acoustic properties such as length and volume of the spoken words. The present study found that the duration of the voice commands in the assertive condition was generally longer than that of the non-assertive voice commands. While speed of a signal word may influence people's perceived urgency [10], it seems as though the lengthier the full command, the more assertive and more urgent they were being perceived. This showed how the slightest changes in the dynamic acoustic elements in speech may influence people's behaviour significantly.

However, people did not responded quicker to the voices because they were more trusting to the assertive voices. Unlike Large et al. [17], assertive voices in the present study were not more trustworthy than nonassertive voices. Nonetheless, they scored high in trustworthiness overall. The differences between Large et al. [17] study's semi-autonomous experience and the present automated system was that the present experience was not a conventional one where the vehicle switched in and out of autopilot mode. There was no proper transition time such as handover or takeover time given but required participants to react to situations as soon as the voice commands were given. Both assertive and non-assertive voice commands might have significantly acted as a safety net for participants. Moreover, Koo et al.'s [16] suggested that people might be more trusting if they were provided with both contextual information (i.e., scenario commands) and the description of action needed (i.e., execution commands in this study) than with one type of information alone. Therefore, the overall high trust might be due to the fact that both types of information were given in both assertive conditions. Therefore, regardless of how assertive the voice commands, It is possible that participants felt reassured because the vehicle was able to provide appropriate feedback and kept them informed about their surroundings.

On the contrary to our prediction, assertiveness did not make a difference in how well people followed the instructions. However, the high level of accuracy in general shows

that participants had no or at least minimal problem following the instructions. In fact, considering that the actions were quite simple and easy to execute and that people found the commands very clear, it isn't surprising that we obtained such a high level of accuracy overall. It is possible that the commands did not significantly overload participants' cognitive processing as they were straightforward and easily understood. This shows the benefit of keeping voice commands using short yet precise to minimize the cognitive workload in participants.

People's concern for safety might be related to why people did not find one voice more annoying than another. This study showed that unimodal audio cues in general were perceived relatively low in annoyance, consistent with Politis et al.'s finding where they showed that unimodal were less annoying than multimodal cues [28]. However, rather than comparing the modalities of the cues, the present differences lie within the unimodal cues. The present concern was whether or not assertive voices might elicit more annoyance in participants than non-assertive voice. The low level of annoyance in general shows that participants might not be complacent about the automated system. Participants who were mostly inexperienced drivers of the automated system might have prioritised their physical safety before their emotional well-being. The priority is beneficial as negative emotions was found to be detrimental to people's decision making [20]. Hence, the present finding demonstrated that novices were not susceptible to the potential annoyance elicited by the voice commands. Further study may explore annoyance in experienced drivers who are being exposed to the system for a longer period time.

People did not prefer the assertive voice more than the non-assertive voice, inconsistent with Large et al. [17]. This is possibly because the voices were relatively low in anthropomorphism. Large et al. [17] suggested that overall rating of a voice was associated with the extent that people viewed the voice as a presence of a company. The lack of social communication in the current voice commands has possibly influenced whether or not the voices were viewed as anthropomorphic or not. It was found that for people's interaction with the in-vehicle voice assistant to be natural, the interaction should be bi-directional and should convey nuances of a human conversation such as having hesitations and using less straightforward language [18]. The present voice commands were unable to fulfill the human-like criteria therefore did not lead to an overall high preference for the voices nor one voice was more preferred than another. Further investigation that includes the different linguistic and conversational elements that are perceived to be anthropomorphic into the assertive voice commands might be able to improve the design to better suit people's taste.

The difference in the immersive experiences between the two different games means that participants were more engaged cognitively in one game than another. However, despite the difference in immersive experiences elicited by the games, participants did not react quicker or slower to voice commands. Unlike the previous studies where their authors examined handover requests which allowed sufficient time for participants to prepare for the transition [12, 19, 21, 26, 32, 33, 36] , this study examined the voice commands that required participants to respond almost immediately, allowing little time for preparation. Therefore, it might be due to the urgent nature of the voice commands in this study that motivated participants to react to the voices even though Fruit Ninja was more immersive.

By disengaging with the game, however, Fruit Ninja participants might have potentially undermined their performance in the game. This is because unlike Smart Shapes where people could take natural breaks without trading off their performance, Fruit Ninja participants could not as Fruit Ninja has more unexpected elements (e.g., random popping up of fruits) that requires player's immediate action. Therefore, results shows that Fruit Ninja participants might have responded promptly to the voice commands even though it might mean that they will lose, making a significant trade-off with their performance. However, we did not track and compare the performances of the two games to confirm this. Future study can measure the performances and gain better insight into how people's interleaving behaviour with different level of engagement with the secondary tasks.

Overall, people responded faster to an assertive voice than a non-assertive voice regardless of how immersive the game was. This can be interpreted as the voices being very effective in delivering their message across, showing the need for execution. Result shows that participants found both voices very clear so the messages in the voice commands were well-understood and in turn motivated people to respond. However, this result also can be interpreted in a completely opposite direction. Despite being less cognitively occupied, less immersed participants did not respond to the voices faster than the more immersive participants who would actually need time to decide whether they should sacrifice their game performance or not. This shows that the effect of less immersive tasks might not be less dangerous than that of a more immersive task as the tasks affected participants' response time to an equal extent. This reinforced an important message in previous studies - a secondary task negatively affects driver's performance and may pose potential risk to the safety of the driver [6, 7, 12, 19, 21, 26, 32, 33, 36]. Therefore, it should be noted that while we give credit to the success of assertive voices in keeping people alert in driving situations, we should also note the negative impacts of engaging in any secondary task that may incur for drivers.

Limitations

The voice commands were limited to simple road-related commands which were not conversational like those in Large et al.'s Wizard-of-Oz study [18]. However, Large et al. collected qualitative data which allowed more flexibility in the exchange of the conversations. But in this quantitative lab study, similar method could not be applied as different variables had to be controlled. A conversation often involves frequent changes in speech properties e.g., consistency, tone and length of a response, adaptation to different contexts and what the response is. It would be very difficult to control the variables of a conversation. However, what is more achievable is for a wider variety of commands to be examined in the future. For example non-driving related reminders such as alerts of daily schedule and reports of daily weather. This may provide a greater understanding in how people might respond to non-driving related voice commands.

As the stimuli presented were videos, no direct feedbacks were given when participants act on the set-up e.g., the 'vehicle' would not stop according to the participants' activity on the brake pedal. Therefore, participants might question how meaningful their actions were when they were not necessarily in control of the 'vehicle'. However, using a standard driving simulator is a tradeoff with a less realistic experience as the presented stimuli presented actual real-life environments. Nonetheless, the absence of feedback might be a concern as it might potentially affect how participants allocated their focuses onto the primary and the secondary tasks and their reaction times as they might question how relevant their actions were.

Also, only selective scenarios required participants to execute actions. In other words, there were plenty of scenarios where voice commands were not given. This design decision was made because we have to control this across trials and conditions. However, participants might have questioned why a voice command was given in one scenario but not another. From observing the raw data, some participants even responded to some scenarios where no voice command was given. It seemed as though participants treated interventions as a safety net just in case the 'vehicle' makes a mistake. This showed that participants were not just comfortable with simply following the instructions, they might think that it was important to act appropriately and consistently at appropriate times in order to feel safe. However, as they were not encouraged to intervene unless they were told to do so, they might not have felt as safe hence influenced their trustworthiness to the voice commands.

6 CONCLUSION

This study investigated people's reactions to and perspectives on voice commands while also engaging in a non-driving task in a semi-autonomous environment. It successfully demonstrates the effectiveness of assertive voice commands in influencing people's speed in executing actions on a vehicle regardless of how cognitively demanding the secondary task was. The finding that people react to assertive voices quicker shows offers a simple and effective way for developers to influence people's attention on the road. Though we inferred that the inexperienced semi-autonomous drivers in this study might not be complacent about the system, future study was yet confirmed whether this applies to the experienced drivers in a long run. Though assertiveness demonstrated its effectiveness in grabbing multi-tasking driver's attention, it is still worrying that less immersed participants did not respond faster to the voice commands than more immersed participants. Therefore, this study carries an important message - despite the useful finding about the assertive voice commands, people should think thoroughly before they engage in any secondary tasks as it can be detrimental to driving activities even with the presence of reminders.

REFERENCES

[1] William Apple, Lynn A Streeter, and Robert M Krauss. 1979. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology* 37, 5 (1979), 715.

[2] Carryl L Baldwin. 2011. Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance. *Ergonomics* 54, 4 (2011), 328–337.

[3] Carryl L Baldwin and Colleen Moore. 2002. Perceived urgency, alerting effectiveness and annoyance of verbal collision avoidance system messages. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 1848–1852.

[4] Roland Barthes. 1977. Image-text-music. *New York. Hill & Wang* (1977).

[5] Jelmer P Borst, Niels A Taatgen, and Hedderik van Rijn. 2015. What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2971–2980.

[6] Duncan P Brumby, Andrew Howes, and Dario D Salvucci. 2007. A cognitive constraint model of dual-task trade-offs in a highly dynamic driving task. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 233–242.

[7] Duncan P Brumby, Dario D Salvucci, and Andrew Howes. 2009. Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1629–1638.

[8] Rob Corbet and Ciara Anderson. 2018. Autonomous vehicles âĂŞ a driver for legal change. *Engineers Journal* (2018). http://www.engineersjournal.ie/2018/02/06/autonomous-vehicles-driver-for-legal-change/

[9] Murat Dikmen and Catherine M 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. ACM, 225–228.

[10] Judy Edworthy, Elizabeth Hellier, Kathryn Walters, Wendy Clift-Mathews, and Mark Crowther. 2003. Acoustic, semantic and phonetic influences in spoken warning signal words. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 17, 8 (2003), 915–933.

[11] Craig A. Giffi, Joseph Jr. Vitale, Thomas Schiller, and Ryan Robinson. 2018. A reality check on advanced vehicle technologies. *Deloitte Insights* (2018). https://www2.deloitte.com/insights/us/en/industry/automotive/advanced-vehicle-technologies-autonomous-electric-vehicles.html

[12] Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. 2013. âĂIJTake over!âĂİ How long does it take to get the driver back into the loop?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 57. SAGE Publications Sage CA: Los Angeles, CA, 1938–1942.

[13] Shamsi T Iqbal, Eric Horvitz, Yun-Cheng Ju, and Ella Mathews. 2011. Hang on a sec!: effects of proactive mediation of phone conversations while driving. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 463–472.

[14] Charlene Jennett, Anna L Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International journal of human-computer studies* 66, 9 (2008), 641–661.

[15] Ing-Marie Jonsson, Mary Zajicek, Helen Harris, and Clifford Nass. 2005. Thank you, I did not see that: in-car speech based information systems for older adults. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1953–1956.

[16] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (2015), 269–275.

[17] David R Large and Gary E Burnett. 2013. DriversâĂŹ preferences and emotional responses to satellite navigation voices. *International Journal of Vehicle Noise and Vibration* 9, 1-2 (2013), 28–46.

[18] David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics* 63 (2017), 53–61.

[19] Lutz Lorenz, Philipp Kerschbaum, and Josef Schumann. 2014. Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1681–1685.

[20] Christine Ma-Kellams and Jennifer Lerner. 2016. Trust your gut or think carefully? Examining whether an intuitive, versus a systematic, mode of thought produces greater empathic accuracy. *Journal of personality and social psychology* 111, 5 (2016), 674.

[21] Vivien Melcher, Stefan Rauh, Frederik Diederichs, Harald Widlroither, and Wilhelm Bauer. 2015. Take-over requests for automated driving. *Procedia Manufacturing* 3 (2015), 2867–2873.

[22] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press.

[23] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[24] Zhaolong Ning, Feng Xia, Noor Ullah, Xiangjie Kong, and Xiping Hu. 2017. Vehicular Social Networks: Enabling Smart Mobility. *IEEE Communications Magazine* 55, 5 (2017), 16–55.

[25] Joseph P Olive. 1997. The talking computer: Text to speech synthesis. *Hal's legacy: 2001's computer as dream and reality* (1997).

[26] Ina Petermann-Stock, Linn Hackenberg, Tobias Muhr, and Christian Mergl. 2013. Wie lange braucht der Fahrer? Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt. *6. Tagung Fahrerassistenzsysteme. Der Weg zum automatischen Fahren* (2013).

[27] Jeff Pittam. 1994. *Voice in social interaction*. Vol. 5. Sage.

[28] Ioannis Politis, Stephen Brewster, and Frank Pollick. 2015. Language-based multimodal displays for the handover of control in autonomous cars. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 3–10.

[29] Katyanna Quach. 2018. Uber robo-ride's deadly crash: Self-driving car had emergency braking switched off by design. *The Register* (2018). https://www.theregister.co.uk/2018/05/24/uber_self_driving_software_crash_report/

[30] Klaus R Scherer. 1978. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology* 8, 4 (1978), 467–487.

[31] Nicole Shechtman. 2002. *Talking to people versus talking to computers: Interpersonal goals as a distinguishing factor*. Stanford University.

[32] Remo van der Heiden, Shamsi T Iqbal, and Christian P Janssen. 2017. Priming Drivers before Handover in Semi-Autonomous Cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 392–404.

[33] Tobias Vogelpohl, Matthias Kühn, Thomas Hummel, Tina Gehlert, and Mark Vollrath. 2018. Transitioning to manual driving requires additional time after automation deactivation. *Transportation research part F: traffic psychology and behaviour* 55 (2018), 464–482.

[34] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[35] Priscilla NY Wong, Jacob M Rigby, and Duncan P Brumby. 2017. Game & Watch: Are Let's Play Gaming Videos as Immersive as Playing Games?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 401–409.

[36] Kathrin Zeeb, Axel Buchner, and Michael Schrauf. 2015. What determines the take-over time? An integrated model approach of driver take-over after automated driving. *Accident Analysis & Prevention* 78 (2015), 212–221.