Modelling Cognitive Workload to Build Multimodal Voice Interaction in the Car

Sylvia Bhattacharya[1][a] and J. Stephen Higgins[2]
[1]*Department of Engineering Technology, Kennesaw State University, Marietta, Georgia, U.S.A.*
[2]*UX Research, Google. Inc, Mountain view, California, U.S.A*

Keywords: In-Vehicle Information Systems, Cognitive Demands, Optimization, Visual/Auditory Modalities, Tactile Screen Tapping, Voice Commands, Visual Cues, Cognitive Load.

Abstract: The paper discusses the integration of in-vehicle information systems and their impact on driver performance, considering the demands of various types such as visual, auditory, manual, and cognitive. It notes that while there's a lot of research on optimizing visual and manual systems, less attention has been paid to systems that use both visual and auditory cues or a combination of different types. The study has found that simple tasks cause the least cognitive strain when drivers use touchscreens, while complex tasks are easier to manage cognitively when voice commands are used alone or with visual aids. These results are important for designing car interfaces that effectively manage the driver's cognitive load.

1 INTRODUCTION

The latest World Health Organization (WHO) report underscores the grave toll of road traffic injuries, indicating that in 2013 alone, 1.25 million lives were lost globally, positioning such injuries as a primary global cause of mortality (WHO, 2016). Currently ranking as the ninth major cause of death across age groups worldwide, road traffic injuries are projected to ascend to the seventh rank by 2030. Focusing on driver distraction, numerous research endeavors have highlighted its significance, attributing between 25% and 75% of all accidents to distraction and inattention (Dingus et al., 2006; Ranney et al., 2000; Klauer et al., 2005; Klauer et al., 2006a; Talbot and Fagerlind, 2006). The escalating adoption of in-vehicle information systems is of paramount importance due to their potential to elicit visual, auditory, manual, and cognitive demands, potentially impacting driving performance in diverse ways. A critical knowledge gap pertains to the intricate interplay between various distraction types and interaction methods during driving. This pivotal context underscores the necessity of a comprehensive understanding to inform safer and more effective driving environments.

Advancements in technology are expanding the capabilities of infotainment systems introduced into vehicles. Communication (e.g., Messaging) and other important user journeys that have traditionally not been available to the driver can now be embedded within in-vehicle information systems (IVIS). Many of these systems have the potential to increase safety, advancements and open possibilities in the car that haven't been available in the past (Klauer et al., 2006b). However, this must be done carefully especially since many new interaction methods (e.g., Voice) do not have a long history of safety evaluation.

All non-driving interactions in a car involve distraction (Victor, 2010). So, the goal of an infotainment system should be supporting basic tasks and short interactions. Complicated tasks should be performed when the car is stopped. Multimodal interaction has the potential to provide flexibility to the user preferred and user safe modality. Google automotive teams use industry standards and internal research to determine how to create products most safely and effectively to be used while driving. There is substantial safety data indicating how to build visual/ manual-based products, but there is not enough data indicating how we should build multimodal visual/voice/ manual systems. We hope

[a] https://orcid.org/0000-0002-5525-7677

Bhattacharya, S. and Higgins, J.
Modelling Cognitive Workload to Build Multimodal Voice Interaction in the Car.
DOI: 10.5220/0012249300003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 1: GRAPP, HUCAPP and IVAPP*, pages 393-399
ISBN: 978-989-758-679-8; ISSN: 2184-4321

these findings contribute to building voice based IVIS systems.

Figure 1: In-built screen car infotainment system.

This study aims to comprehensively analyse and assess user workload and preferences within the context of multimodal interaction involving a prototype infotainment system. The overarching objective is to optimize interaction modalities for diverse tasks such as navigation, messaging, and their sequential execution. By investigating the desirability and safety of multimodal interactions for drivers, the study seeks to identify optimal modal pairings for fundamental navigation tasks. Furthermore, the research will delve into the advantages and disadvantages associated with fully voice-based interactions and hands-free approaches in automotive settings. Another critical aspect is the examination of mental workload and distraction patterns during secondary interactions. Through these inquiries, the study contributes to enhancing the design of infotainment systems, fostering safer and more efficient driver experiences.

2 METHOD

We collected driving data along with brain signal data using Electroencephalography (EEG) sensors from 15 participants, (5 females and 10 males). All the participants were between the age of 25 years to 54 years. The experiment was conducted in June 2021 in the Google Mountain View campus following all required human subject protocols. As shown in Figure 2, we used a driver simulator to collect the driving data. An Open BCI EEG headset was used to collect the brain signals from the drivers (participants). We used an Android Auto tablet in the simulator with internally designed prototypes to test the driver - infotainment system interaction. These 15 subjects interacted with the Android Auto assistant while driving on a simulated freeway with moderate traffic across four modalities (only voice without any info on the screen, Voice with information on the screen, only touch/tap, Voice along with touch for two types of tasks). The modalities were repeated for two types of tasks: a “single shot” in which they are required to complete one simple task with one interaction; and a “multi shot”, which involved more than six interactions in which the driver had to complete three major tasks back-to-back. Two sessions of data from each participant were collected, which means they had to do a total of 16 tasks. We also conducted a survey asking questions after every task, which were designed based on NASA TLX workload assessment.

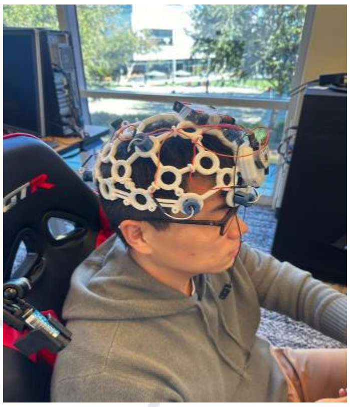


Figure 2: Experimental scenario from Mountain View UX Research Laboratory, Google.

A detailed description of each stage of the above-mentioned methods is stated below.

2.1 Participant Selection

Data was collected from 15 non-Googler participants between June 19th to June 30th, 2021, in a simulated research laboratory environment. The participants were recruited through Answer Lab/UXI. Several criteria were considered for recruiting the participants: i) Participants must have a valid U.S driver’s license ii) Participants should drive at least 3+days a week iii) Should have experience with Android Auto iv) Should have experience using Google map and assistant. v) Participants should be local to the Mountain View/ San Francisco area. There were no specific requirements for gender, age group, or ethnicity.

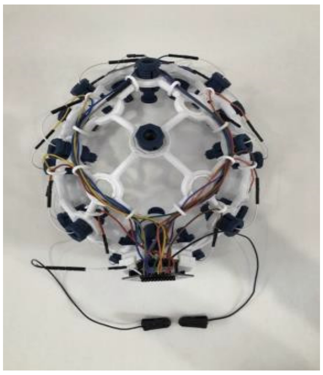
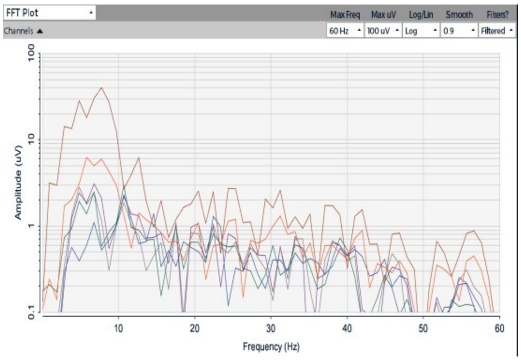


Figure 3: Open BCI 8 channel EEG Headset.

2.2 EEG Setup

EEG recordings tend to utilize between 32-64 electrodes placed according to the IEEE 10-20 positioning. However, setting up the electrodes on the scalp is time consuming and can result in unreliable data. Optimizing the number of electrodes pertaining to the user's brain activity could potentially reduce the number of electrodes and improve the overall classification accuracy. This would reduce both the computation and acquisition times. Furthermore, using too many electrodes sometimes deteriorates the quality of EEG signals as it picks up many unwanted signals from the user's scalp.

Therefore, we have used an EEG headset from OpenBCI with 8 channels/ electrodes for our experiment. The electrodes are distributed to cover the whole scalp (frontal, occipital, temporal and parietal lobes) as shown in Figure 3. The headset is supported by OPENBCI software that helps read the brain signals directly from a computer and export the raw signals for further signal processing. Figure 4 shows the real time raw brain signals-alpha (8 -12 Hz), beta (12-30 Hz), gamma (above 30Hz), theta(4-8Hz) and delta (0.4-4Hz). The activity of the brain signals change based on the neural activity of the participant.



FFT Plot

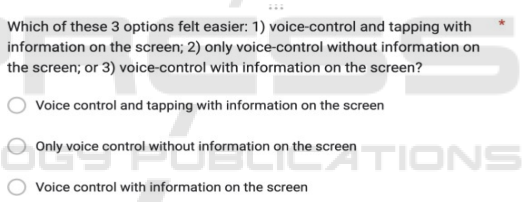
Max Freq: 60 Hz | Max uV: 100 uV | Log/Lin: Log | Smooth: 0.9 | Filters?: Filtered

Figure 4: Real time OPENBCI raw FFT plot.

For example, when a human subject is resting with eyes closed, brain signals between 8-12 Hz i.e alpha waves are found to be dominant in the frontal areas of the brain. When the subject performs a lot of mental computation, the beta wave starts dominating all other signals. This is how the brain signals continuous changes during neural processes in a human brain. We will be looking at beta and gamma waves as a method to quantify whether the driver is distracted.

2.3 NASA TLX

NASA Task Load Index (NASA TLX) is widely used in UX research as a subjective multidimensional tool for assessing mental workload related to a task. As specified earlier in this paper, a user survey was also conducted after every driving task to understand the participant state and preference. The questions were designed based on NASA TLX scoring style. After every task, we asked the question how mentally demanding the task was and to rate it out of five where '1' is 'not very demanding' and '5' is 'very demanding'. At the end of every session, they were asked which modality felt the easiest for them overall as shown in Figure 5.



Which of these 3 options felt easier: 1) voice-control and tapping with information on the screen; 2) only voice-control without information on the screen; or 3) voice-control with information on the screen? *

○ Voice control and tapping with information on the screen

○ Only voice control without information on the screen

○ Voice control with information on the screen

Figure 5: Final Survey Question.

2.4 Prototypes

Two main types of prototypes were created in this experiment. The first one was a single interaction/ single shot task prototype, and the second was a multiple interaction/ multi-shot task prototype. Single interactions are those tasks in which the participant must interact only once with the assistant to complete the task, and they are usually simple and quick. But multi-interaction tasks are those in which the driver must interact multiple times (6+) back-to-back to complete one task.

Again, for each of these interactions, four modalities were designed as follows:

i) Voice-only single/multi shot tasks (task to be completed by voice-audio interaction only; no visual information related to the conversation provided on the screen)

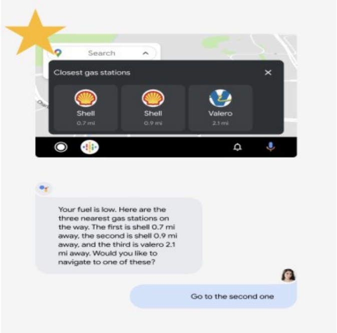
ii) Voice + screen single/multi shot tasks (task to be completed by voice interaction only but visual information about the interaction is available on the assistant screen)

iii)

iv) Voice + ap single/multi shot tasks (task to be completed using both voice and tapping options on the screen)

v) Only tap single/multi shot tasks (task to be completed using only tapping the option on the screen).

Figure 6 shows a screenshot of the prototype for single shot voice+ screen modality task.

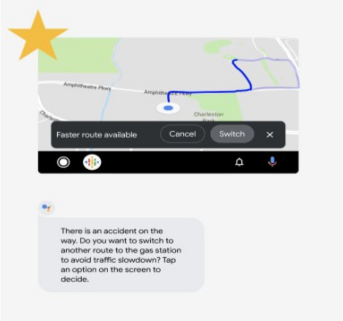


Your fuel is low. Here are the three nearest gas stations on the way. The first is shell 0.7 mi away, the second is shell 0.9 mi away, and the third is valero 2.1 mi away. Would you like to navigate to one of these?

Go to the second one

Figure 6: Single Interaction Voice + Screen Modality Prototype.

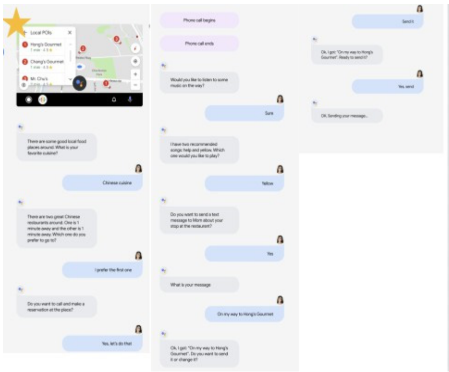
This is a single shot/ single interaction task with voice and visual/information on the screen. During the experiment, the Assistant asks the driver to choose one of the nearest gas stations as the fuel is low. Assistant makes the interaction by voice and the information about the gas stations are provided on the screen for the driver to see the options. The driver was required to make the choice using voice interaction only. Whereas Figure 7 shows another single interaction task where the driver/ participant must make a choice by tapping on the screen.



There is an accident on the way. Do you want to switch to another route to the gas station to avoid traffic slowdown? Tap an option on the screen to decide.

Figure 7: Single Interaction Tap Modality Prototype.

Here the user needs to make a choice by either tapping the option ‘switch’ or ‘cancel’ on the screen. Similarly, the same modalities were used for multi-interaction tasks, but the driver had to complete one task. Figure 8 shows the chain of multi-interaction during the driving experiment.



There are some good local food places around. What is your favorite cuisine?

Chinese cuisine

There are two great Chinese restaurants around. One is 1 minute away and the other is 1 minute away. Which one do you prefer to go to?

I prefer the first one

Do you want to call and make a reservation at the place?

Yes, let's do that

Phone call begins

Phone call ends

Would you like to listen to some music on the way?

Sure

I have two recommended songs: help and yellow. Which one would you like to play?

Yellow

Do you want to send a text message to Mom about your stop at the restaurant?

Yes

What is your message

On my way to Hong's Gourmet

Ok, I got: "On my way to Hong's Gourmet". Do you want to send it or change it?

Send it

Ok, I got: "On my way to Hong's Gourmet". Ready to send it?

Yes, send

OK. Sending your message...

Figure 8: Multi Interaction Voice Modality Prototype.

As you can see in the interactions above, the Assistant starts the conversation asking some questions about favourite cuisine and then providing choices of those cuisines around and then making a call for reservation. The Assistant then asks if the driver would like to listen to some music and provides some choices of music.

Next, the Assistant asks if the driver wants to send a text message and works with the user to send the text. So, here, the user completes three major tasks (calling, tuning music in the car, and sending text messages) with almost nine interactions back-to-back.

3 DATA ANALYSIS

The mental workload from the EEG signal was calculated using a few steps as follows:

3.1 Signal Processing

The band power of the EEG signal was computed from the raw data using Python scripting. The Gamma and Beta band power for each participant was specifically extracted for further analysis.

3.2 Quantitative Analysis

Brain signals are divided into five main types based on their frequency ranges (Alpha, Beta, Gamma, Delta, and Theta). Each of these brain signals are usually dominant during specific tasks. Beta and Gamma are the signals that are used for mental workload estimation.

The cognitive load index (CLI)/ mental workload is calculated using the ratio of mean gamma and beta band power from frontal and parietal electrodes respectively. The mathematical formula is shown below in Figure 9:

$$\text{CLI} = \frac{\text{Mean Gamma BP (F3+F4) electrodes}}{\text{Mean Beta BP (P3+P4) electrodes}}$$

Figure 9: Cognitive Load Index.

3.3 Task Based Analysis

The average CLI is calculated for each task and modality (e.g., single shot only voice task) and it is compared with a baseline value computed based on resting EEG of each participant. A baseline value from resting EEG of every participant during the experiment was also collected. The CLI value was compared against the baseline value. Similarly, the mental workload from NASA TLX two sessions, both EEG and survey data was averaged over the sessions. The mental workload from EEG and NASA TLX was also compared to validate the results. The above calculations were made for every subject individually. Results in the section below shows the aggregate result of all the 15 participants in this study.

4 RESULTS

Figure 10 and Figure 11 shows the result on mental workload from one shot/ single interaction tasks.

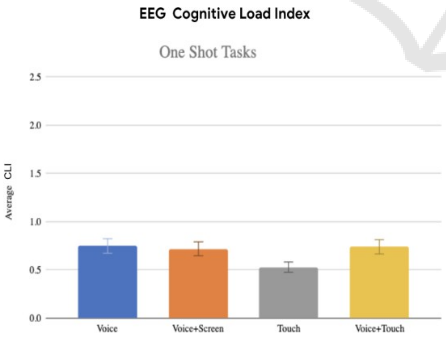


Figure 10: Mental workload with EEG over different modalities for one-shot tasks.

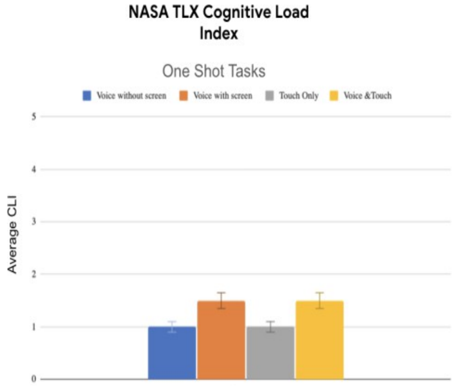


Figure 11: Mental workload with NASA TLX over different modalities for one-shot tasks.

The above graph shows that Only Touch modality involved the least CLI/ mental workload during one-shot tasks. The EEG result also overlaps with NASA TLX observation. According to EEG data, touch modality has the least cognitive load which means it's least challenging for drivers to handle conversation with touch while driving. So, overall, touch modality seems to be the common best modality from both the techniques for one shot interactions.

For multi-shot interactions, again NASA TLX and EEG data indicate that touch is the modality with the highest cognitive load and voice with or without screen are the best and preferred modalities with least cognitive load among all the four modalities, as shown in Figure 12 and 13 below.

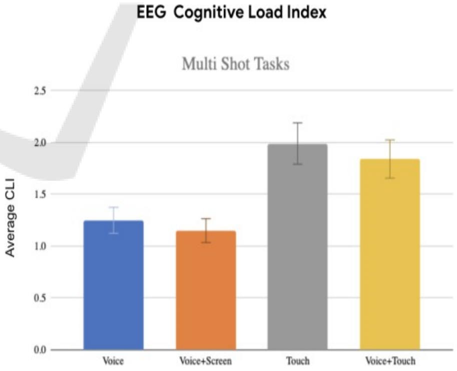


Figure 12: Mental workload with EEG over different modalities for multi-shot tasks.

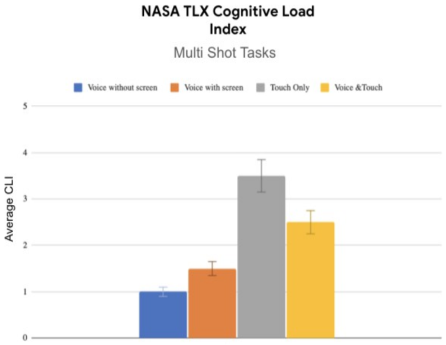


Figure 13: Mental workload with NASA TLX over different modalities for multi-shot tasks.

The participants were also asked a final question at the end to choose the most preferred modality according to their experience with the drive. And the answer chosen by most of the participants is ‘voice with information on the screen’ across all types of tasks (both single shot and multi shot interactions). Also, in NASA TLX final survey participants rated ‘touch’ as an easy modality for one shot tasks during the survey, they specified that overall touch is most disliked as it causes distraction for them to take eyes off the road to make the choices on the screen.

We also compared the NASA TLX choices between two sessions, and we noticed that for both single shot and multi shot, the participants rated most of the modalities easier in the second session which they rated harder in the first one. The modalities which were rated as easy in the first session were rated easy in the second session as well. Figure 14 and Figure 15 shows the compared NASA TLX responses of the participants between the first and second session for both one shot and multi-shot tasks.

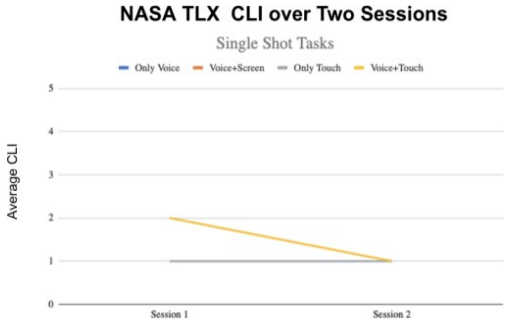


Figure 14: NASA TLX ratings over two sessions (one shot).

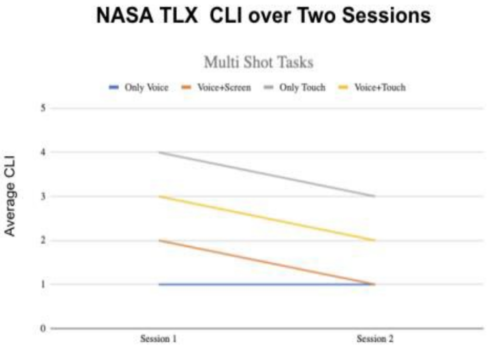


Figure 15: NASA TLX ratings over two sessions (multi-shot).

In Figure 13, participants rated all modalities as ‘1’(easy) in the first session except Voice+ Touch modality. In the second session, they rated all modalities as easy again and the Voice+ Touch modality as easy as well. In Figure 14, for multi-shot tasks, participants rated all modalities as harder in the first session except ‘only voice’ modality but in the second session they rated all these modalities as easier than what they rated in the first session. This observation brings up an important research question, “How much experience will drive perceived workload down across time?” which will be explored in our future research.

If we look at the EEG signals of the participants over two sessions for one shot and multi shot tasks, we get a similar story. Figure 15 and 17 shows the mental workload observed with EEG signals over two sessions. This also shows that the workload tends to be lower in the second session.

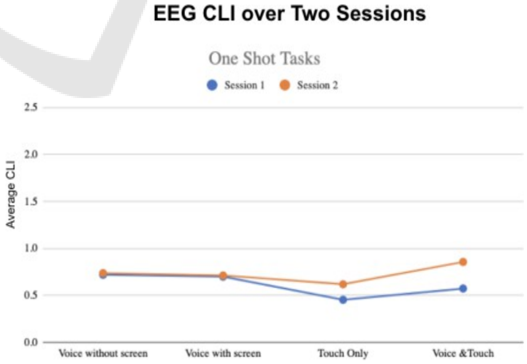


Figure 16: Mental workload with EEG over two sessions (one shot).

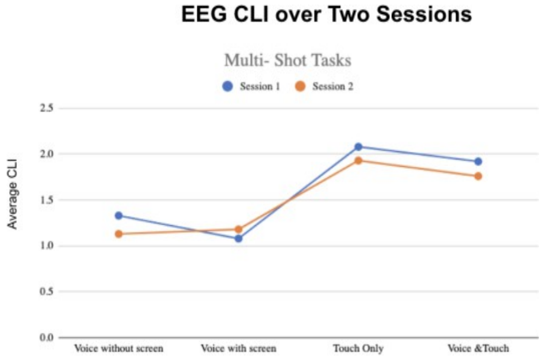


Figure 17: Mental workload with EEG over two sessions (multi-shot).

Although we cannot conclude from this data, we can see the possibility of reduced workload over time, if the participants have multiple experiences with the tasks.

5 CONCLUSIONS

In conclusion, our investigation into the realm of multimodal voice interactions within automotive settings has unearthed intriguing patterns. For swift and succinct single shot interactions, voice-based commands emerge as an optimal choice, yielding comparable cognitive load to touch interactions. Conversely, for intricate multi-turn or extended engagements, voice interactions take precedence, especially when accompanied by relevant visual cues, facilitating a seamless conversational experience. The dynamic interplay between experience and subjective or EEG-based workload assessments remains an enigma, warranting further exploration. Similarly, the precise scenarios in which visual information enhances various types of voice interactions warrant deeper scrutiny. As we navigate towards a future of enhanced in-vehicle interfaces, it becomes evident that additional empirical endeavours are indispensable. By embarking on forthcoming experiments, we can illuminate the intricacies surrounding these facets, fostering a more refined and holistic understanding of effective multimodal voice interaction design.

REFERENCES

World Health Organization. Global Status Report on Road Safety 2015. Available online: http://apps.who.int/iris/bitstream/10665/189242/1/9789241565066_e ng.pdf?ua=1 (accessed on 2 July 2016).

Dingus, T.A.; Klauer, S.; Neale, V.; Petersen, A.; Lee, S.; Sudweeks, J.; Perez, M.; Hankey, J.; Ramsey, D.; Gupta, S.; et al. National Highway Traffic Safety Administration: Washington, DC, USA, 2006.

Ranney, T.A.; Mazzae, E.; Garrott, R.; Goodman, National Highway Traffic Safety Administration: Washington,DC, USA, 2000; Volume 2000.

Klauer, S.G.; Neale, V.L.; Dingus, T.A.; Ramsey, D.; Sudweeks, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Orlando, FL, USA, 26–30 September 2005; SAGE Publications: Thousand Oaks, CA, USA; Volume 49, pp. 1922–1926.

Klauer, S.G.; Dingus, T.A.; Neale, V.L.; Sudweeks, J.D.; Ramsey, D.J. National Highway Traffic Safety Administration: Washington, DC, USA, 2006.

Talbot, R.; Fagerlind, H. Exploring inattention and distraction in the SafetyNet accident causation database. Accid. Anal. Prev. 2009, 60, 445–455.

Klauer, S.G.; Dingus, T.A.; Neale, V.L.; Sudweeks, J.D.; Ramsey, D.J. National Highway Traffic Safety Administration: Washington, DC, USA, 2006.

Victor, T. The Victor and Larsson (2010) distraction detection algorithm and warning strategy. Volvo Technol.2010, 1, 0–6.