

Approaching Intelligent In-vehicle Infotainment Systems through Fusion Visual-Speech Multimodal Interaction: A State-of-the-Art Review

Mahmoud Baghdadi

Achim Ebert

baghdadi@rptu.de

aebert@rptu.de

University of Kaiserslautern-Landau
Kaiserslautern, Rhinland Platinate, Germany

ABSTRACT

Advanced in-vehicle infotainment systems are part of the future's intelligent autonomous vehicles. Developing such systems requires advanced interaction modalities to be utilized. On the other hand, using sophisticated applications on an unimodal touch display for manual driving vehicles could endanger drivers' lives and result in a poor user experience. Offering in-vehicle fusion multimodal interaction could broaden the types of applications and enhance user experience while keeping safety and low distraction into account. Since in-vehicle interaction is bi-directional, both driver-vehicle and vehicle-driver sides are equally important to achieve and develop advanced infotainment systems. Searching in related literature, it has been found that there is good progress in driver-vehicle fusion multimodal interaction; in comparison, only a scarce amount of research on the vehicle-driver side is available. This paper presents the state-of-the-art in vehicle-driver fusion multimodal interaction for infotainment systems. This type of interaction is essentially a form of human-computer interaction. However, when we specify that the computer is a vehicle computer, certain specific factors become essential to consider when developing the user interface for this type of computer. Furthermore, a research agenda together with challenges and opportunities is proposed.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models; Auditory feedback; Graphical user interfaces; Displays and imagers; Interaction techniques; Displays and imagers; Auditory feedback; Graphical user interfaces.**

KEYWORDS

In-vehicle interaction, HUD, head-up display, fusion multimodal interaction, vehicle-driver interaction, visual-speech interaction

ACM Reference Format:

Mahmoud Baghdadi and Achim Ebert. 2024. Approaching Intelligent In-vehicle Infotainment Systems through Fusion Visual-Speech Multimodal Interaction: A State-of-the-Art Review. In *European Conference on Cognitive Ergonomics (ECCE 2024), October 08–11, 2024, Paris, France*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3673805.3673818>

1 INTRODUCTION

Nowadays, vehicles have infotainment systems that drivers can interact with for non-driving related tasks (NDRT). Drivers can interact with vehicles using different modalities (e.g., tactile and auditory). On the other hand, vehicles can also interact with drivers through different modalities (e.g., visual and auditory). According to Dix, there are two interaction channels: input, where humans interact with computers, and output, where computers interact with humans [7]. In other words, when drivers interact with vehicles, it is considered an input interaction, whereas, when vehicles interact with drivers, it is an output interaction. We can also express the same meaning of input interaction by the phrase driver-vehicle interaction and of output interaction by the phrase vehicle-driver interaction. Each interaction channel could be conducted using either a single modality (unimodal interaction) or several modalities (multimodal interaction).

Natural human-to-human interaction is multimodal. A person communicates verbally to another person while using hand gestures to point to a direction or an object. On the contrary, the other person listens and looks simultaneously to comprehend the ongoing communication. Similarly, humans communicate with computers through different modalities, such as voice and touch. However, when we communicate with a computer, a smartphone for instance, we still need to communicate in a specific way for the smartphone to understand us. In other words, it is a conditional interaction; if we don't adhere to the conditions, the smartphone will not understand us correctly. While human computer interaction has reached a point where multimodality is possible, different environments impose some limitations. The driving environment is one good example because the driving task demands visual attention; vehicles cannot simply interact visually with the driver all the time. It is also vital for in-vehicle interaction to be as quick and short as possible for the driver not to be distracted for long. Searching in literature, the research on driver-vehicle natural interaction has reached an advanced level where drivers can interact with vehicles in a fusion multimodal interaction [1]. On the contrary, the research focusing on vehicle-driver interaction is still in its infancy.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ECCE 2024, October 08–11, 2024, Paris, France

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1824-3/24/10

<https://doi.org/10.1145/3673805.3673818>

In fact, vehicle-driver interaction is much more complicated when safety is a concern. We know from the literature that drivers initiate interaction with vehicles when the driving condition allows it [5]. In other words, drivers know when it is safe enough to interact with vehicles. However, current in-vehicle technology has not yet reached an advanced level of intelligence to determine the right moment to interact with the driver. Another point that is worth mentioning is drivers can decide which modality is suitable for the current driving condition, whereas vehicles need to be developed to interact in accordance with the driving condition.

Fusion multimodal interaction could overcome some unimodal downsides. In-vehicle speech unimodal interaction causes less visual distraction than touch (which demands visual attention) interaction. However, speech has some interaction downsides that could negatively impact driving activity (e.g., users are uncertain if the vehicle comprehended their message) [4]. This could lead the in-vehicle speech system to repeat what it understood, extending the interaction time. Multimodal interaction comes in handy to eliminate this issue. A useful multimodal combination for speech is visual; however, while visual output distracts the driver, it could still be safe in some driving conditions (i.e., while the vehicle is semi-autonomously driving). Visual-speech multimodal interaction would overcome the system comprehension downside and turn the conditional interaction into a more natural one. Visuals can help the driver validate what the system understood via short text, while with speech modality, it could continue the conversation based on the displayed context. In other words, the system does not need to repeat what it understood for the driver to validate the message. This would close the gap between a system's conditional and natural interactions. Therefore, developing an intelligent in-vehicle interaction is crucial to be able to combine visuals with speech. Intelligent in-vehicle interaction means the vehicle is contextually aware of the driving environment and able to display well-designed visuals on a suitable display in the vehicle.

Developing multimodal output interaction is as crucial as developing multimodal input interaction for vehicles to reach natural interaction with the driver. This could also help reduce distraction and enhance user experience. Furthermore, infotainment and work tasks become the main tasks in fully autonomous vehicles. In such scenarios, vehicles should be well developed to interact with the passengers naturally (fusion multimodal interaction). Another factor to emphasize in developing vehicle-human fusion multimodal interaction is implicit interaction. Implicit interaction requires the vehicle to initiate the interaction with passengers whenever needed [20]. It is also worth mentioning that offering fusion multimodal output, in some cases, is even more important than input. For instance, on a smartphone, when a user verbally asks the voice assistant about the weather conditions, the phone speaks back a short answer but provides more details visually on the display. In vehicles, drivers need similar multimodal output. Nonetheless, it is important to carefully develop visual-speech multimodal output systems to ensure safety.

With the advancement of in-vehicle technologies, intelligent vehicles that are equipped with intelligent infotainment systems will become real. During that journey, fusion vehicle-driver interaction will gradually become much more important. Since the current vehicles promote speech interaction to reduce distraction, an intelligent visual-speech fusion interaction is crucial to overcome speech

downsides and enhance users' experience and safety. More and more, implicit interaction is predicted to be part of the soon coming future intelligent vehicles where fused visual-speech interaction is a must. Therefore, we will discuss in this paper the state-of-the-art of current research on vehicle-driver visual-speech interaction in the context of infotainment systems. Furthermore, based on the presented literature, we will develop a research agenda and illustrate research opportunities and challenges for future research on vehicle-driver visual-speech fusion multimodal interaction.

2 BACKGROUND

Developing infotainment systems to become intelligent is crucial due to the high demand for additional features and applications. From a user perspective, a user may call a system intelligent when the system suggests useful commands or tools depending on the task being accomplished to quicker and more efficiently achieve that task. According to Krishnakumar, "an intelligent system is one that emulates some aspects of intelligence exhibited by nature [17]. This includes "learning, adaptability, robustness across problem domains, improving efficiency (over time and/or space), information compression (data to knowledge) and extrapolated reasoning" [17]. Developing an infotainment system that complies with Krishnakumar's definition should promote a conventional infotainment system to an intelligent one. However, based on the definition, we can still call an infotainment system intelligent if it has some but not all intelligence aspects. Garzon in his paper considered an infotainment system that is contextually aware and be able to learn from the user interaction based on different situations intelligent [8]. Therefore, we can deduce that intelligent infotainment systems can be of different levels of Intelligence, something similar to autonomous vehicles where we have levels of SAE standards [11].

While multimodal interaction user interfaces in vehicles are still in their infancy, searching in literature, studies on multimodal interaction have begun in the late 1980s or early 1990s. According to [21], there are three types of multimodality; Redundant is the first type. Redundant multimodality offers more than one modality where the user can use one or more modalities to accomplish a task. The second type is temporally cascaded multimodality. This type involves more than one modality to accomplish a task. However, not all modalities are used simultaneously; rather, one modality is used after another. Temporally cascaded multimodality requires the user to utilize two or more modalities to accomplish a task; in other words, a task cannot be accomplished using only one modality like the case in the redundant multimodality type. The third type occupies the term fused and commonly the term "put-that-there". This type of multimodality also requires the user to interact with two or more modalities. However, all modalities are simultaneously used and then combined to elicit an outcome.

Multimodality in interaction requires the utilization of two or more modalities. In literature, these modalities consist of seven modalities: visual, auditory, haptic, olfactory, gustatory, cerebral and cardiac [13]. Every interaction between two entities requires one side to output a signal and the other to input that signal. Therefore, the two interaction channels are input and output interaction [7]. Humans and machines can interact with each other through

some of the modalities explicitly and others implicitly [20]. Because humans have no direct conscious control over their cardiac and cerebral modalities, interacting with machines through these modalities is considered an implicit interaction. On the other hand, humans can interact explicitly through the rest of the modalities. It is also worth mentioning that not all modalities are suitable for in-vehicle interaction (i.e., gustatory) [13].

Vehicle-driver interaction is basically computer-human interaction; however, by specifying that the computer is a vehicle computer, more specific factors are crucial when developing the user interface for this kind of computer. For manual driving vehicles, distraction is one crucial factor. Since driving is the main activity, any other non-driving related task (NDRT) is considered a secondary activity (i.e., using an infotainment system) [22]. This does not necessarily mean that all secondary activities are less important than the driving activity; however, they should not distract the driver from fully performing the driving activity. Distracting the driver from the main activity could lead to dangerous situations. Other factors should also be considered when developing user interfaces for vehicles, these factors are cognitive load, usability, situation awareness and task performance. The latter consists of task completion time, efficiency and error rate.

Since the invention of cars, almost every piece of information the driver needed was an analog meter. The first display-equipped cars were revealed in the 1970s [15]. For the first time in history, the dashboard of a car was a digital display. The early displays that were installed cars were monochrome displays, mostly green in color. Throughout the time, more displays made their way to cars with even more colors. Nowadays, many cars have at least one high-resolution full-color multitouch display. This display is usually placed in the center cluster of a vehicle and it is called the infotainment display. Another type of in-vehicle display would be the Head Down Display (HDD), which is the exact dashboard in a modern full-colored display. This display is usually not a touch display and can be interacted with through the infotainment display or the buttons on the steering wheel. Another type of display that was installed to assist the driver is the Head-Up Display (HUD). This type of display is projected onto the windshield where the driver can see some information without taking the eyes off the road. Since displays are also used for entertainment, more and more displays have made their way to cars, for instance, rear passenger displays, which are normally installed on the back of the front seats. Some modern cars are also equipped with a display for the front passenger. In other words, a modern sedan vehicle can be equipped with 6 displays. It is expected in the near future that HUDs will evolve into a larger state where the whole windshield turns into a display; scientists currently call it a Windshield Display (WSD) [6, 9]. The next section will present the state-of-the-art literature on vehicle-driver visual-speech fusion multimodal interaction.

3 VEHICLE-DRIVER FUSION MULTIMODAL INTERACTION: A STATE OF THE ART

Since speech has minimum visual distraction effects on drivers, it has some downsides that could lead to distraction. Visuals are a good addition to speech when we develop vehicle-driver fusion multimodal interaction systems. However, displaying visuals has

a direct connection to visual distractions. Therefore, many factors must be considered before developing such systems. The literature in this section was found by focusing our search on in-vehicle multimodal interaction and related terms (e.g., Infotainment system and fusion multimodal interaction). We then excluded all papers related to non-fusion multimodal interaction and non-vehicle-driver interaction. It is important to acknowledge that the analysis of the presented literature is focused on the visual-auditory aspects of vehicle-driver interaction, which aligns with the scope of this paper. Readers should know that the original studies may encompass a wider range of findings beyond this specific focus. It is evident in the related literature that the design of the graphical user interface (GUI) and the location and display type are directly linked to developing an efficient visual-speech fusion multimodal system for vehicle-driver interaction. Thus, this section is divided into subsections based on in-vehicle display type.

3.1 Head-up Display and Windshield Display

Several papers utilized HUDs or WSDs in combination with speech for visual-auditory output [12, 18, 26]. However, there seem to be opportunities for further exploration in the area of GUI design within current research in the field of HCI, specifically for HUD and WSD. Furthermore, most visual-auditory literature does not exploit the fusion type of multimodality but rather implements the redundant or temporal cascaded types. Jakus et al. conducted a user study to compare auditory, visual, and audio-visual interaction in cars [12]. The focus of this paper was on the output side of interaction; therefore, they used basic buttons and the scrolling wheel of a mouse as input. Participants were asked to perform several tasks like changing the fan speed and asking for the average speed of their trip. A standard hierarchical interface was projected on the windshield (HUD) for visual output and speech for auditory display. The participants have experienced the interface via auditory, visual, and audio-visual multimodal displays. The visual and audio-visual displays were faster and more efficient in accomplishing secondary tasks than the auditory display. Most of the participants preferred using the multimodal display over unimodal displays. While both displays had no significant difference in driving performance, they still scored some negative points. The GUI of the HUD in this paper was colored small sized text only. Improving the interface design could influence the results and probably make the multimodal display the most preferable by participants and improve usability.

Li et al. conducted a user study where they evaluated three interaction systems: using an iPhone by hand and talking to the assistant (Siri), pressing a button on the steering wheel and talk to Siri and press a button on the steering wheel and Siri displaying the results on the HUD as a visual and speech output [18]. The latter is considered a fusion multimodal interaction. The participants were asked to follow a lead car and accomplish location-based service tasks (e.g., finding a gas station). Driving performance and eye movement were measured in this study. Using unimodal interaction, pressing a button on the steering wheel and Siri talking back through speech only scored the best ratings. The multimodal interaction, in which Siri talks back and displays the results on the HUD, scored second place. In this study, the authors did not use a dedicated GUI for the HUD but rather mirrored whatever Siri displays on the iPhone

into the HUD. This could be one reason the multimodal interaction scored less than the unimodal interaction. The interface of Siri on the iPhone was designed to serve a small colorful display and not a transparent display like the HUD on vehicles. An enhancement to the design of the visual output on the HUD could influence the results dramatically.

Audio-visual output interaction has always been the ideal modality for navigation systems. In a novel approach, Topliss et al. used a HUD to project navigation intersections as fixed arrows and compared it to a lead virtual vehicle projected into the HUD with the help of augmented reality (AR) [26]. The audio part was basically the voice instruction of the navigation activity and since audio complements visual, it is considered a fusion multimodal interaction. The participants were asked to drive and use both navigation interfaces to reach a destination. The study found no significant difference between the two methods regarding navigation performance, confidence and mental workload. However, it was proposed that combining the two methods could positively impact navigation performance. This paper indicates that the current navigation arrow style is still ideal in some parts of a journey. However, the AR leading car could enhance the user experience in the other parts but not replace the arrows completely. Additionally, this study indicates that modern AR GUI designs could be a good addition to enhance usability but not necessarily replace non-AR visuals completely.

3.2 Infotainment Display

The infotainment display, also called center stack display, is another type of in-vehicle display. Some research utilizes the infotainment display instead of a HUD [10, 24, 27]. Hofmann et al. compared several concepts of speech-based in-vehicle interface [10]. They had an interface for command dialog and another for conversational dialog. They also tested using both interfaces with the combination of visual GUI displayed on the vehicle's center stack infotainment display. The participants were asked to talk to the system and ask for information or give the system a task to perform (e.g., book a hotel). Driving performance and usability were the main measurements of this paper. The results show that command-based and conversational-based dialogs were accepted without the visual output; however, the participants better accepted the command-based dialog. According to the authors, the reason behind that could be the limited performance of the conversational system in understanding the language. Furthermore, the results show that the visual output had impaired driving performance. The authors recommended that such GUIs be designed with minimal content so as not to cause distraction. This paper reveals that visual output should be well designed to serve the driver's needs without causing much distraction and cognitive load. The paper also revealed that GUI design is a crucial factor for visual output in vehicles. It is worth mentioning that the display used in this study was located in the center stack where the driver needed to look away from the road. The display location concerning the driving situation could cause part of the distraction. Using a HUD with enhanced GUI design could impact the results positively. On the speech side of the interaction, a much more developed conversation dialog system could enhance the user experience for such multimodal interaction.

In another research paper, Schneeberger et al. proposed and conducted a study using a system called GetHomeSafe (GHS) [24]. With a suitable and minimally designed GUI, this system allows the driver to browse news, make a hotel reservation and use other social media apps such as Facebook while driving. All services that GHS offers can be done using a speech dialog system and a GUI that is displayed on an infotainment display (mounted display on the center stack). Measuring driving safety and usability, the authors compared the GHS system to a traditional tablet computer system mounted in the vehicle. The participants were asked to perform 2 scenarios for each application (i.e., hotel, Facebook and the news app) using the GHS system, whereas performing one scenario for each application using the traditional tablet. The results show that the driver looked less on the road using the tablet, which means that the GHS system allowed the driver to focus more visually on the road. From the distraction point of view, the participants were less distracted while using the GHS. The study also revealed that GHS was less mentally demanding, and the speech dialog system was an advantage in road interaction. The GHS generally scored higher positive ratings for usability over the tablet system. In this paper, the visual output of the GHS had no major visual distraction to the driver; the reason could be that the GUI was well designed to suit a fusion with speech in a driving environment. This study reveals that a suitable GUI design for the driving environment could introduce more apps and services while enhancing usability and maintaining a low mental workload.

Winzer et al. designed and developed two Human Machine Interfaces (HMIs) to assist with intersection traffic lights [27]. The first interface is two-dimensional and the other is perspective HMI. From these two HMIs, the driver can obtain information about the upcoming traffic light, whether it is green or red and also a count-down for red traffic lights. They displayed both interfaces on a 5.7" mobile phone display mounted on the center stack vent. The voice also gave directions after each intersection (fusion multimodal). In this study, the participants were asked to drive once without using the HMI and twice using the designed HMI (once in the 2D and another using the perspective design). Measuring eye glances and considering NHTSA's (National Highway Traffic Safety Administration) guidelines, both HMIs were not visually distracting to the primary driving task because glances on the HMIs were less than 2 seconds. The main focus of this paper was the design of both HMIs. They created a suitable visual design that requires the driver to only glance for less than 2 seconds. Displaying text or a basic visual design of such an interface could result in a dangerous and distracting HMI unsuitable for vehicles while on the move.

The literature in this section reveals that GUI design is a crucial factor in developing a suitable fusion visual-speech interaction for the driving environment. Furthermore, visuals that were designed for a certain type of display do not necessarily suit another type. When utilizing HUDs for visual-speech fusion interaction, it's crucial to approach visuals differently than with normal displays. Additionally, it is not necessarily that modern AR application designs can completely replace classic arrow designs for in-vehicle navigation purposes.

4 OPPORTUNITIES AND CHALLENGES

Based on the presented state-of-the-art literature, this section will propose future research opportunities and challenges for vehicle driver fusion visual-speech interaction.

4.1 Visual Output Location

Nowadays, some cars have several displays. The infotainment system could be enhanced in terms of user experience and reducing distraction by having the ability to display visuals in different locations depending on the user's and the car's situation. Driving-related features in a vehicle could also function better depending on where visuals are displayed. For instance, HUD excels the Head-Down display (HDD) when displaying speed, navigation directions and alerts while driving [16, 25]. There are head tracking sensors in modern Mercedes Benz vehicles that help determine which side mirror the driver is looking at; if the driver looks at the left side, he or she can adjust the side mirror without needing to select that mirror by a switch [19]. In other words, the vehicle is smart enough to adjust the side mirror that the driver intended to control. This kind of sensor could serve many other safety and infotainment features. For instance, some infotainment systems do not offer keyboard typing while the vehicle is in motion to protect the driver from long visual distractions. With such sensors, we could offer keyboard typing to passengers if the driver is looking at the road and not at the display. Another situation in which this kind of sensor could enhance user experience is voice assistance in vehicles could display some crucial visuals to the suitable display where the driver is looking. When driving, it could be safer and better to display some visuals on the HUD, not the center stack display. A good practice for this is displaying voice assistant responses which then would be called a fusion visual-speech interaction.

4.2 GUI Design

It is evident that some literature does not focus on GUI design. However, GUI design could influence usability, cognitive workload and task completion time. Another aspect that some literature lacks is GUI design for a display type like HUD; it is not appropriate to display visuals designed for mobile phone displays or center stack in-vehicle displays and simply project them on a transparent display type like the HUD. Colors and visual elements should be treated differently on HUDs, black color and small sized texts are not suitable. In fact, it is impossible to use black color visuals and project them onto a transparent glass via light. Also, other colors could not be suitable for HUDs because the background's brightness changes depending on the time of day, vehicle's location and weather conditions. HUDs are ideal for in-vehicle AR applications. However, because these applications are for driving environments, AR applications should be designed accordingly. A great example is [23]; they have projected horizontal bars on the road to indicate when speed reduction should start, and these bars react according to the vehicle's speed and position. In other words, the driver reduces speed whenever the first bar projected on the ground is reached and further reduces speed when the second bar is reached. If this paper only displayed information in numbers (as seconds), the whole study could have had different results. We are assuming

this because, in some of the current vehicles, the infotainment system displays the distance to the next speed limit change in numbers (meters), adding other numbers indicating when to start reducing speed would be overwhelming to read and process. The method employed in [23] offers promise for mitigating information clutter. To conclude, HUDs in vehicles should be treated differently, and GUI design should be innovative to serve the driver's needs without exposing him/her to unsafe situations.

4.3 Exploiting Visual-Auditory Fusion Output

Multimodal interaction could eliminate or at least reduce some issues of unimodal interaction. Speech interaction as an auditory unimodal has many usability issues that a visual modality could overcome. Users usually are confused about when to start speaking to the voice assistant and whether their voice is being recognized or not [14]. A simple solution utilized in mobile phone operating systems is to produce an earcon and show an icon (usually a mic) that reacts to the voice intensity. This visual-auditory multimodal interaction solved a usability issue of auditory-only interaction. Speech interaction is temporal, sometimes users may miss what the systems have said due to high cognitive workload traffic conditions. Fusing speech with visuals could be a solution to this kind of issue. For instance, whatever the system speaks, a visual representation of the accomplished action (not necessarily text) could be displayed for a long period. This way, users could always validate what the system has done even if they missed its speech. On the flip side, visual output as unimodal might also be harmful; nevertheless, fusing visual with auditory modality could cure that harm. Through research, many features and vehicle interactions could be enhanced and brought to safety by fusion output multimodal interaction. To conclude, it is not common in research to use fusion multimodal interaction for vehicle-driver interaction whereas, this type of multimodality is crucial for future intelligent infotainment systems.

4.4 Contextual-Aware Vehicles

In a fully autonomous vehicle (AV), visual output is not a distraction, similarly, in a manual driving vehicle, in some situations, the visual output does not cause a visual distraction to the driving task. One of these situations is when the vehicle is in a static form where the driver does not need to look at the front road. Instead of limiting visual output in all situations in a manual driving vehicle, we could offer visual output only in situations that do not affect safety and instead enhance usability. In other words, contextually aware vehicles could offer a better user experience interacting through a fusion of visual-auditory multimodal interaction. Another situation is if the vehicle has an Advanced Driver Assistance System (ADAS) where that can help the driver maintain lane position on a highway with the help of Adaptive Cruise Control (ACC); the driver in such a situation has less cognitive workload and also can glance longer on the infotainment display than when driving without an ADAS. Mobile operating systems, iOS and Android, have unique interfaces for vehicles, Apple CarPlay [3] and Android Auto [2]. In previous Android Auto releases, the system turns the interface to dark mode (where dark colors are used rather than bright ones) when the vehicle turns the main beam on. Since the vehicle has a

sensor to measure the amount of the ambient light, Android Auto uses that sensor indirectly to enhance user experience. If an external system could access some of the vehicle's parameters, then the vehicle's own system has access to even more parameters. In other words, the vehicle's infotainment system is most worthy to exploit these parameters to offer intelligent and enhanced user experience infotainment systems and presumably reduce distraction to the primary task.

5 CONCLUSION

We presented in this paper the state-of-the-art literature on fusion visual-speech multimodal interaction from the vehicle-driver perspective. The number of in-vehicle multimodal interaction papers is vast however, very few focus on the fusion type. We have excluded all non-fusion multimodal interaction and all driver-vehicle interaction papers. Thus, the presented literature in this paper is specifically for vehicle-driver fusion visual-speech multimodal interaction.

The fusion of visual and speech for in-vehicle interaction plays a significant role in developing intelligent infotainment systems for future vehicles. Depending on the level of intelligence, intelligent infotainment systems would have a significant impact on enhancing drivers' user experience for current vehicles and most probably for future fully autonomous vehicles. In non-autonomous vehicles, such a system would help improve safety by visually interacting only in safe situations, as well as displaying visuals on a suitable screen. Based on the presented literature, four findings are worth considering. To begin with, depending on different driving situations, intelligent infotainment systems should exploit the number of displays in the vehicle to show visuals on the most suitable display location. Additionally, research should emphasize GUI design considering the type of display. Furthermore, infotainment systems should exploit visual-speech fusion multimodal to enhance user experience and safety rather than offering a redundant type of multimodality. Finally, contextual-aware vehicles could boost the development of intelligent infotainment systems for future vehicles.

In future works, it will be beneficial to evaluate a visual-speech fusion infotainment system that considers minimalist GUI design and acknowledges the characteristics of the different types of displays. A step further is to test how that system would perform with an implicit interaction in different driving conditions (i.e., a city and highway drive).

ACKNOWLEDGMENTS

This research was financially supported by a scholarship from King Abdulaziz University. The authors would like to express their sincere gratitude to King Abdulaziz University for their generous support, which made this work possible.

REFERENCES

- [1] Abdul Rafey Aftab. 2019. Multimodal Driver Interaction with Gesture, Gaze and Speech. In *2019 International Conference on Multimodal Interaction (ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 487–492. <https://doi.org/10.1145/3340555.3356093>
- [2] Android. 2024. Android Auto. <https://www.android.com/auto/>
- [3] Apple. 2024. iOS - CarPlay. <https://www.apple.com/ios/carplay/>
- [4] Grace M. Begany, Ning Sa, and Xiaojun Yuan. 2016. Factors Affecting User Perception of a Spoken Language vs. Textual Search Interface: A Content Analysis. *Interacting with Computers* 28, 2 (March 2016), 170–180. <https://doi.org/10.1093/iwc/iww029>
- [5] Michael Braun, Jingyi Li, Florian Weber, Bastian Pfleging, Andreas Butz, and Florian Alt. 2020. What If Your Car Would Care? Exploring Use Cases For Affective Automotive User Interfaces. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3379503.3403530>
- [6] Nabil Al Nahin Ch, Diana Tosca, Tyanna Crump, Alberta Ansah, Andrew Kun, and Orit Shaer. 2022. Gesture and Voice Commands to Interact With AR Windshield Display in Automated Vehicle: A Remote Elicitation Study. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '22)*. Association for Computing Machinery, New York, NY, USA, 171–182. <https://doi.org/10.1145/3543174.3545257>
- [7] Alan Dix. 2007. *Human computer interaction* (3. ed., [nachdr.] ed.). Pearson Prentice Hall, Harlow, England [u.a.]
- [8] Sandro Rodriguez Garzon. 2012. Intelligent In-Car-Infotainment Systems: A Contextual Personalized Approach. In *2012 Eighth International Conference on Intelligent Environments*. IEEE, Guanajuato, Mexico, 315–318. <https://doi.org/10.1109/IE.2012.70>
- [9] Renate Haueslschmid, Susanne Forster, Katharina Vierheilg, Daniel Buschek, and Andreas Butz. 2017. Recognition of Text and Shapes on a Large-Sized Head-Up Display. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, New York, NY, USA, 821–831. <https://doi.org/10.1145/3064663.3064736>
- [10] Hansjörg Hofmann, Vanessa Tobisch, Ute Ehrlich, André Berton, and Angela Mahr. 2014. Comparison of speech-based in-car HMI concepts in a driving simulation study. In *Proceedings of the 19th international conference on Intelligent User Interfaces (IUI '14)*. Association for Computing Machinery, New York, NY, USA, 215–224. <https://doi.org/10.1145/2557500.2557509>
- [11] SAE International. 2024. SAE Levels of Driving Automation™ Refined for Clarity and International Audience. <https://www.sae.org/site/blog/sae-j3016-update>
- [12] Grega Jakus, Christina Dicke, and Jaka Sodnik. 2015. A user study of auditory, head-up and multi-modal displays in vehicles. *Applied Ergonomics* 46 (Jan. 2015), 184–192. <https://doi.org/10.1016/j.apergo.2014.08.008>
- [13] Pascal Jansen, Mark Colley, and Enrico Rukzio. 2022. A Design Space for Human Sensor and Actuator Focused In-Vehicle Interaction Based on a Systematic Literature Review. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 56:1–56:51. <https://doi.org/10.1145/3534617>
- [14] Jingun Jung, Sangyoon Lee, Jiwoo Hong, Eunhye Youn, and Geehyuk Lee. 2020. Voice+Tactile: Augmenting In-vehicle Voice User Interface with Tactile Touchpad Interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376863>
- [15] Jeff Smith Kia. 2022. The History and Evolution of In-Car Digital Display Systems. <https://www.kiakool.com/blog/2022/april/20/the-history-and-evolution-of-in-car-digital-display-systems.htm>
- [16] Raymond J. Kiefer. 1991. Effect of a Head-Up Versus Head-Down Digital Speedometer on Visual Sampling Behavior and Speed Control Performance During Daytime Automobile Driving. *SAE Transactions* 100 (1991), 82–93. <https://www.jstor.org/stable/44632015> Publisher: SAE International.
- [17] Kalmanje Krishnakumar. 2002. Intelligent Systems for Aerospace Engineering: An Overview. Brussels. <https://ntrs.nasa.gov/citations/20020065377> NTRS Author Affiliations: NASA Ames Research Center NTRS Document ID: 20020065377 NTRS Research Center: Ames Research Center (ARC).
- [18] Guofa Li, Fangping Zhu, Tingru Zhang, Ying Wang, Shengfan He, and Xingda Qu. 2018. Evaluation of Three In-Vehicle Interactions from Drivers' Driving Performance and Eye Movement behavior. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2086–2091. <https://doi.org/10.1109/ITSC.2018.8569917> ISSN: 2153-0017.
- [19] Mercedes-Benz. 2020. The new Mercedes-Benz S-Class. <http://media.mercedes-benz.ca/releases/the-new-mercedes-benz-s-class>
- [20] Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. 2022. Intelligent In-Vehicle Interaction Technologies. *Advanced Intelligent Systems* 4, 2 (2022), 2100122. <https://doi.org/10.1002/aisy.202100122> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202100122>
- [21] Christian Müller and Garrett Weinberg. 2011. Multimodal Input in the Car, Today and Tomorrow. *IEEE MultiMedia* 18, 1 (Jan. 2011), 98–103. <https://doi.org/10.1109/MMUL.2011.14> Conference Name: IEEE MultiMedia.
- [22] Matthias Peissner, Vanessa Doebler, and Florian Metze. 2011. Can voice interaction help reducing the level of distraction and prevent accidents? (2011). <https://publica.fraunhofer.de/handle/publica/295615> Publisher: Fraunhofer IAO.
- [23] F. Schewe and M. Vollrath. 2020. Ecological interface design effectively reduces cognitive workload – The example of HMIs for speed control. *Transportation Research Part F: Traffic Psychology and Behaviour* 72 (July 2020), 155–170. <https://doi.org/10.1016/j.trf.2020.05.009>
- [24] Tanja Schneeberger, Simon von Massow, Mohammad Mehdi Moniri, Angela Castronovo, Christian Müller, and Jan Macek. 2015. Tailoring mobile apps for

- safe on-road usage: how an interaction concept enables safe interaction with hotel booking, news, Wolfram Alpha and Facebook. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '15)*. Association for Computing Machinery, New York, NY, USA, 241–248. <https://doi.org/10.1145/2799250.2799264>
- [25] Missie Smith, Joseph L. Gabbard, Gary Burnett, and Nadejda Doutecheva. 2017. The Effects of Augmented Reality Head-Up Displays on Drivers' Eye Scan Patterns, Performance, and Perceptions. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 9, 2 (2017), 1–17. <https://doi.org/10.4018/IJMHCI.2017040101> Publisher: IGI Global.
- [26] Bethan Hannah Topliss, Sanna M Pampel, Gary Burnett, Lee Skrypchuk, and Chrisminder Hare. 2018. Establishing the Role of a Virtual Lead Vehicle as a Novel Augmented Reality Navigational Aid. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 137–145. <https://doi.org/10.1145/3239060.3239069>
- [27] Oliver M. Winzer, Antonia S. Conti-Kufner, and Klaus Bengler. 2018. Intersection Traffic Light Assistant – An Evaluation of the Suitability of two Human Machine Interfaces. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 261–265. <https://doi.org/10.1109/ITSC.2018.8569708> ISSN: 2153-0017.