IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 24, 2020, accepted June 6, 2020, date of publication June 16, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002775

A Cascaded Multimodal Natural User Interface to Reduce Driver Distraction

MYEONGSEOP KIM, EUNJIN SEONG, YOUNKYUNG JWA, JIEUN LEE, AND SEUNGJUN KIM, (Member, IEEE)
School of Integrated Technology, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea
Corresponding author: Seungjun Kim (seungjun@gist.ac.kr)

This work was supported by GIST Research Institute (GRI) *GIAI* grant funded by the GIST in 2020.

ABSTRACT Natural user interfaces (NUI) have been used to reduce driver distraction while using in-vehicle infotainment systems (IVIS), and multimodal interfaces have been applied to compensate for the shortcomings of a single modality in NUIs. These multimodal NUIs have variable effects on different types of driver distraction and on different stages of drivers' secondary tasks. However, current studies provide a limited understanding of NUIs. The design of multimodal NUIs is typically based on evaluation of the strengths of a single modality. Furthermore, studies of multimodal NUIs are not based on equivalent comparison conditions. To address this gap, we compared five single modalities commonly used for NUIs (touch, mid-air gesture, speech, gaze, and physical buttons located in a steering wheel) during a lane change task (LCT) to provide a more holistic view of driver distraction. Our findings suggest that the best approach is a combined cascaded multimodal interface that accounts for the characteristics of a single modality. We compared several combinations of cascaded multimodalities by considering the characteristics of each modality in the sequential phase of the command input process. Our results show that the combinations speech + button, speech + touch, and gaze + button represent the best cascaded multimodal interfaces to reduce driver distraction for IVIS.

INDEX TERMS Cascaded multimodal interface, driver distraction, head-up display (HUD), human-computer interaction (HCI), in-vehicle infotainment system (IVIS), learning effect, natural user interface (NUI).

I. INTRODUCTION

With advances in human-vehicle interaction technology, in-vehicle infotainment systems (IVIS) (e.g., navigation, radio, music, etc.) offer useful information that enhance the driving experience. However, these systems also increase distraction from primary driving tasks and can pose a danger to drivers and others on the road [1]–[7]. According to the NHTSA, there are three categories of driver distractions [5]:

1) *Visual distraction* involves tasks that require the driver to look away from the road to visually obtain information;
2) *Cognitive distraction* involves tasks that require the driver to avert their mental attention away from the primary driving task;
3) *Manual or physical distraction* involves tasks that require the driver to take one or both hands off the steering wheel to manipulate a control, device, or other non-driving-related items.

To minimize those distractions, single-mode natural user interfaces (NUIs), which provide more natural forms of interaction (touch, speech, gesture, and vision) than the traditional, physical button [8], have been suggested [9]–[12]. These single-mode NUIs can diminish specific distractions, but can also increase other types of distractions (e.g., gaze can cause visual distraction, speech can cause cognitive distraction), as each single-mode NUI has individual characteristics and requires a different cognitive and/or physical load. (In this paper, gaze interface refers to the manipulation of the system via eye behavior, such as staring at an object to select it, or blinking, etc.). Thus, single-mode NUIs can actually distract the driver or worsen driving performance. For example, when compared to a center console touch screen,

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

a gaze-controlled, interface system-based head-up display (HUD) improved driving performance but also increased cognitive load [13]. A steering wheel with a touch gesture and visual output screen performed well on both visual distraction and usability compared with the center console IVIS, but the prototype still exhibited more frequent lane deviations [14].

To overcome the drawbacks of any single modality, multimodality, which provides drivers with two or more modalities, has been suggested [4], [12], [15], [16]. Many multimodality studies offer drivers redundant modality choice and then observe which modalities they choose [9]. Although this bottom-up approach provides a basic understanding of multimodality and is useful for excluding unlikely modalities, it provides a limited understanding of why the driver selects a particular modality and how it affects driver distraction or driving performance.

Furthermore, the degree and extent of distraction is influenced by how long a driver has used a modality, and how well and how quickly they have acclimated to it. In other words, drivers may be unfamiliar with a given modality at first, but their interactions can become more natural over time. This ''learning effect'' – the result of using a modality continuously [9], [17] – is important to consider when defining NUIs.

The degree to which the driver is accustomed to using a modality varies by its characteristics [15], [18]. Drivers have been accustomed to button and touch modalities for some time, whereas mid-air gesture is novel. High scores for the usability of a button or touch might be the result of long-term user experience, whereas new modalities may show relatively low usability, not because they are not usable but because they are unfamiliar. To confirm this, researchers need to test the usability of different modalities over days. In fact, Nacenta *et al.* [19] conducted similar experiments for two days to compare pre-defined gestures, random gestures, and user defined-gestures, and found significant results. While longer studies on NUIs have been proposed in order to investigate more comfortable gestures or commands to memorize, many studies have not taken into account modalities that can be learned and quickly adapted to by subjects based on repeated experience [12], [13], [15], [16].

In this study, we take a top-down approach where we provide multimodality combinations based on a more complete understanding of each type of distraction and five modalities and how they interact. We attempt to understand each modality in relation to the distraction it causes (e.g., which types of distraction occur or to what extent distraction occurs when divers use a certain modality), and then we consider the step-by-step interaction between a driver and IVIS to select a single modality. Finally, we show how temporally cascaded multimodality affects actual drivers' usability and distraction. A cascaded multimodal interface processes two or more single modalities in a temporal order (e.g., gaze, gesture, speech). Partial information supplied by recognition of an earlier mode (e.g., gaze) constrains interpretation of a later one (e.g., manual selection), which then may jointly constrain interpretation of a third mode (e.g., speech). Such interfaces may combine active input modes, passive input modes, or blend input types [20]. Our study also conducts the same driving task the following day in order to capture which modalities exhibit a learning effect. We believe this approach will help us identify a highly effective multimodality combination and provide new insights on distraction.

The contributions of this paper are 1) demonstrating a holistic approach to comparing five single modalities (a physical button, touch gesture, mid-air gesture, speech, and gaze input) from the perspective of three driver distractions (visual, cognitive, and manual) with identical conditions; 2) understanding how drivers interact with IVIS, and designing a multimodal interface based on the characteristics of a single modality and driver understanding in IVIS; and 3) suggesting several types of cascaded multimodalities to reduce driver distraction and increase usability for IVIS.

In the following section, we introduce related research about single-modal NUIs and multimodality. Next, we present three user studies. In the preliminary study, we surveyed 133 people to determine the most common scenarios for IVIS. We found that music and navigation systems are the most used IVIS systems. In Study 1, we investigated the characteristics of five single modalities from the perspective of driver distractions with ten participants. We found two levels of interactions can be extracted depending on the time and types of interaction (e.g., early [Level 1] and late [Level 2]). In Study 2, we attempted to find a highly effective multimodal NUI by comparing eight cascaded multimodalities over two days with 25 participants. We found speech + button, speech + touch, and gaze + button to be suitable cascaded multimodalities for drivers while interacting with IVIS. In the Discussion section, we interpret our results from the perspective of the learning effect and address some limitations.

II. RELATED WORKS

NUIs can mitigate distractions and reduce driver workload, and thus make driving safer. Döring *et al.* [14] developed a touch gesture-based steering wheel that reduces visual demands. Koyama *et al.* [21] designed a multi-touch steering wheel that recognizes 6 handshape patterns to encourage drivers to keep their hands on the steering wheel when manipulating in-car secondary applications.

Beyond studies on single modalities, comparative studies have sought to determine which NUIs are least distracting or mentally taxing [13], [22]. Roider *et al.* [22] showed that the driver's impairments were greatest when they used an input modality (gaze, speech, and gesture) that required the same cognitive resources as the driving situation. They concluded that NUI design should not overlap with the driver's resources (visual, auditory, and physical) occupied by the outside situation and the input sensory modalities for the secondary task. Angelini *et al.* [10] compared gesture, speech, and touch modalities in terms of usability, subjective workload, and emotional response. However, because feedback was offered in different areas, such as a head-up display (HUD)

or central console display, they did not set the conditions for an equivalent comparison. For example, drivers' visual attention is short because their field of view remains nearer to the road area by using the HUD located on the steering wheel [14], which has inherently lower visual distraction than a center console interface. Similarly, Moran *et al.* [23] showed that HUDs are suitable for manipulating IVIS (i.e., navigation) and show lower visual distraction than widely-used head-down displays (HDD). While these studies provide valuable insights, distinct locations for output feedback and different numbers of input commands make accurate comparisons between single modalities impossible.

Multimodal interfaces have been proposed in order to improve a primary modality with the addition of a secondary modality, such as gaze input-pointing gesture [24] or speech-gesture on the steering wheel [15]. The temporally-cascaded multimodal interface provides users with two or more modalities in order of time or sequence [25]. Roider *et al.* [12] designed cascaded multimodalities with touch and speech modalities and investigated how they impact driver distraction and interaction duration. They found that switching modalities during driving does not affect the time required to perform sub-tasks, such as spatial tasks, where both shapes from the center area are moved to the corresponding surrounding fields (e.g., triangle to bottom-left and square to top-right), and verbal tasks, where participants characterize the element on top of the screen by shape, color, and size. Interestingly, most previous cascaded modality studies, including Roider *et al.*, applied only speech and touch gesture in sequence, and vice versa, depending on types of tasks (e.g., spatial, verbal) [26].

However, there are other possible single modalities, such as mid-air gesture and gaze gesture. Therefore, in this study, we consider a total of five single modalities commonly used for NUIs and compare them to investigate the type and extent of distraction caused by each (Study 1). From these results, we suggest several cascaded multimodal combinations that can minimize distractions and increase usability. We then test them to identify the most effective multimodal NUIs (Study 2). We used a HUD as the output interface in both experiments to reduce visual distraction significantly by keeping the driver's eyes on the road [27], [28]. Therefore, we unify output feedback using a HUD and match the number of input commands across modalities.

III. PRELIMINARY STUDY

To identify a typical use case scenario for IVIS, we conducted a preliminary online survey with 133 people ($M = 43.0$, $SD = 15.6$, 90 males, 43 females). Respondents reported their recent driving experience (less than 1 year = 15.04%, 1-3 years = 9.02%, 3-5 years = 6.02%, 5-7 years = 3.01%, more than 7 years = 55.64%, none = 11.28%). We asked, 'What do you often manipulate in the vehicle, in addition to driving?' in the form of a multiple-choice questionnaire. Our results (Table 1) identified music/radio (70.68%), navigation (58.65%), and air-conditioner ○/heater (45.86%) as the three most common scenarios.

TABLE 1. IVIS ratios frequently used by drivers during manual driving & physical and mental effort required to manipulate IVIS.

IVIS (In-Vehicle Infotainment System)	Demand	*M*	*SD*
1. Music / Radio system (70.68%)	Physical	2.78	1.06
	Mental	2.75	1.04
2. Navigation system (58.65%)	Physical	3.25	1.12
	Mental	3.11	1.09
3. AC / Heater system (45.86%)	Physical	2.67	1.01
	Mental	2.57	0.97
4. Acting make a call (25.56%)	Physical	3.15	1.11
	Mental	3.00	1.07
5. Video control (3.01%)	Physical	2.96	1.01
	Mental	2.89	1.04

We also asked, 'How much physical and mental effort do you need to manipulate IVIS in a vehicle?'. Respondents were asked to respond to a five-point Likert scale on physical and mental demands (1 – low demand, 5 – high demand). Table 1 shows the reported physical and mental demand experienced by drivers while simultaneously manipulating IVIS and driving. We compared the data for the top three scenarios from Question 1. We analyzed the physical and mental effort as a paired t-test. The navigation system (Physical = 3.25, t (132) = 6.67, $p < 0.001$; Mental = 3.11, t (132) = 6.69, $p < 0.001$) and music/radio system (Physical = 2.78, t (132) = 1.77, $p = 0.079$; Mental = 2.75, t (132) = 2.85, $p < 0.01$) showed higher effort than the AC/heater system (Physical = 2.67, Mental = 2.57). Based on these findings, we designed a navigation system and music/radio system for our IVIS.

IV. STUDY 1: EQUIVALENT COMPARISON OF SINGLE MODALITY NUIs CONSIDERING DRIVER DISTRACTION

In the first user study, we designed five NUIs (touch gesture, mid-air gesture, speech, gaze, and a physical button interface) for IVIS and collected data on driver distractions, subjective preferences, interview answers, primary task performance (i.e., driving performance), and secondary task performance. A total of 10 drivers experienced each of the five single modalities. We investigated five single modalities to determine whether each is advantageous or disadvantageous for IVIS in terms of driver distraction.

A. STUDY DESIGN

This experiment was conducted in an indoor driving simulator setup (Atomic A3 motion platform, Logitech G920 steering wheel and pedal). The simulation was created with Unity, and the interface was connected with the driving simulation after we created the input device for each modality (Fig. 1.a). To compare the five modalities under equivalent conditions, the HUD was used as a common output feedback interface, as shown in Fig. 2, which is effective in reducing visual distraction during operation. As shown in Fig. 2.c, d, and e, we designed opaque UIs for the left and the middle areas of the windshield, which represented a comfortable personal

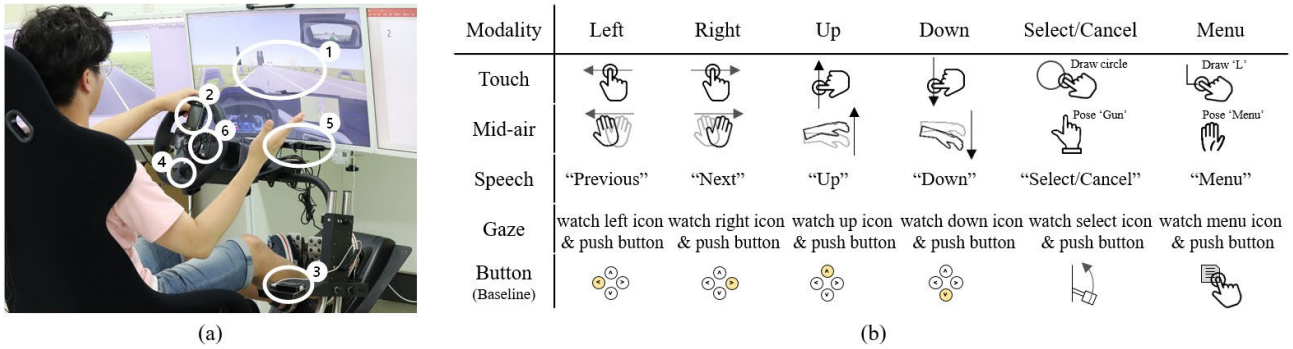

FIGURE 1. (a) Designed four NUIs and button testbed (① Full windshield head-up display, ② Touch panel (touch gesture), ③ LEAP motion controller (mid-air gesture), ④ Microphone (speech), ⑤ Eye-tracker (gaze), ⑥ Buttons (button). (b) Input commands according to five NIUs.

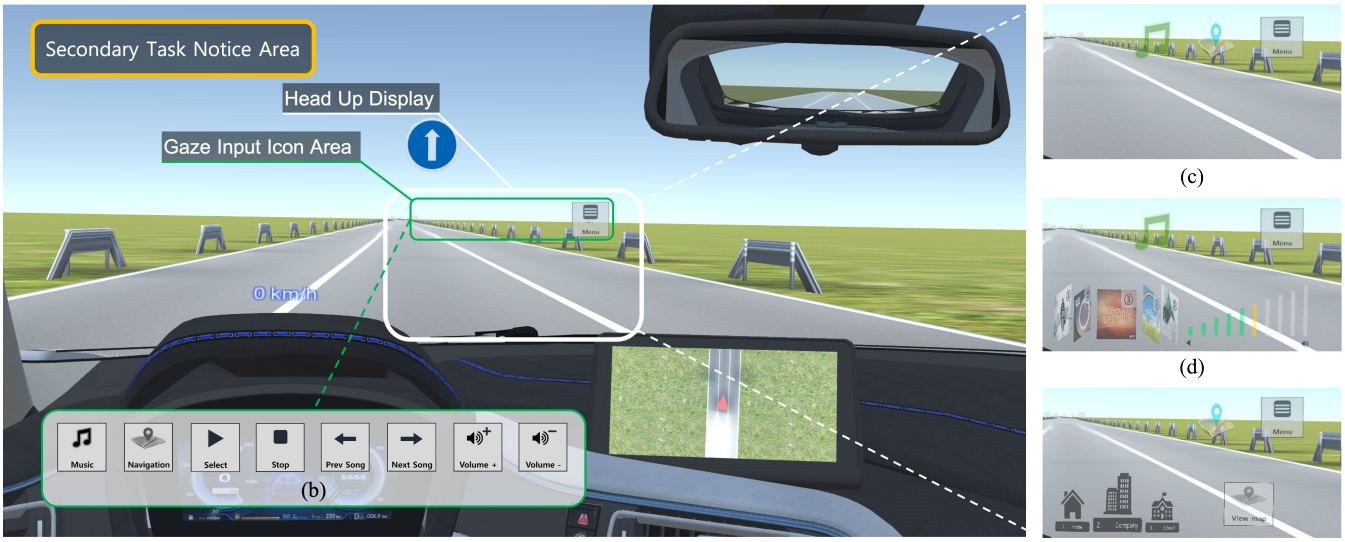

FIGURE 2. (a) Simulation environment (white block – head-up display (HUD) area, green block – gaze input icon area, orange block – secondary task notice area). (b) Examples of gaze input icon. (c) Music, navigation & menu UI. (d) Songs & volume UI in music/radio scenario. (e) Destination & map UI in navigation scenario.

area for the driver to interact with buttons, controllers, and gestures [23]. We unified the number of input commands for each modality to be six identically composed input commands. Fig. 1.b shows all commands used in each modality.

Drivers were asked to perform the lane change task (LCT) to obtain reliable and comparable driving performance data in this study [9], [14], [29], [30]. The LCT assesses performance in terms of lane-keeping and lane-changing, calculated in terms of the mean lane deviation between paths the driver is supposed to follow during the task and the path they followed. Maciej and Vollrath [9] assert that the advantage of the LCT is creating a well-defined level of task difficulty that demands the driver's frequent attention. They note that the LCT does not prevent the driver from performing secondary tasks at the same time, because this represents natural behavior in the car more generally. Therefore, the LCT enables us to compare each characteristic of single modalities while drivers perform different secondary tasks.

The objective of the LCT is to drive along a three-lane highway and to change lanes according to signs while maintaining a speed of 60 km/h. As with the typical LCT scenario, participants drove on a straight road without other vehicles. One session consisted of 18 lane changes in random order, with a lane change approximately every 10 seconds. Drivers were asked to first recognize the lane change sign, then perform the action the moment they passed the sign. Each session took about 3 minutes. To measure driving performance, we collected the number of lane deviations (i.e., the number of times drivers encroached on another lane after changing lanes) and the number of responses to task instruction (i.e., how well the driver responded to a sign in a given session) separately.

Simultaneously, the interaction task of manipulating the NUI according to the instructions was carried out as a secondary task. We constructed the experimental IVIS into a music and navigation system, according to the results of the preliminary survey. As shown in Table 2, two categories and 12 detailed commands were included. The interaction task was created by combining commands. For example, in order to complete the ''play the third song in the playlist'' task,

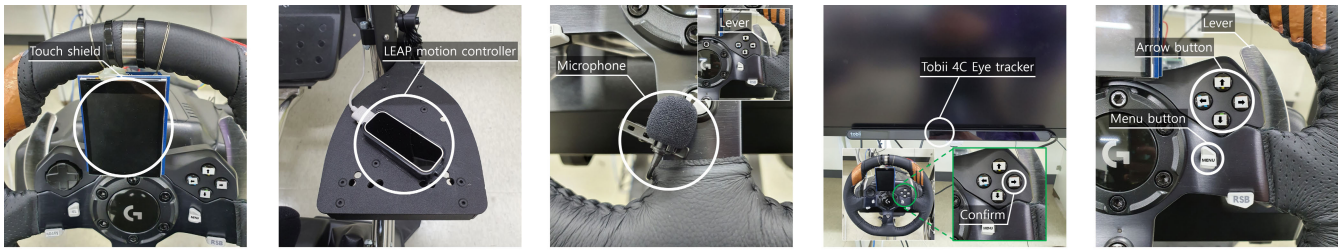

FIGURE 3. (a) Touch shield on the steering wheel (for touch gesture). (b) LEAP motion controller at cockpit's right (for mid-air gesture). (c) Microphone (for speech). (d) Tobii 4C Eye tracker on the center monitor (for gaze). (e) Button on the steering wheel (for button).

TABLE 2. Interaction tasks for IVIS.

Category	Command	Secondary task
Music System	Play	Play the third song on playlist
	Stop	Stop playing the song
	Next song	Raise the volume to Level 2
	Previous song	
	Volume up	
	Volume down	
Navigation System	Next destination	Select destination to Home
	Previous destination	
	Select	Turn on the road map
	Cancel	
	Turn on the roadmap	
	Turn off the roadmap	

the driver had to use the interface that matched the input commands (Fig. 1.b).

Participants were randomly assigned five interaction tasks per session through visual (Fig. 2.a) and auditory (speaker) notification. Video, eye-tracking data (Tobii 4C), primary and secondary task performance data were collected for all sessions. When participants completed each session, they conducted the NASA-TLX, a System Usability Scale (SUS), and a brief interview [3], [28].

Touch gesture was made with Arduino and 2.8' TFT capacitive touch shield (Fig. 3.a). Following previous studies, it was attached to the top-center of the steering wheel [21]. The driver entered gesture commands using the touch panel in the center of the steering wheel, similar to a tablet PC. The touch gesture consisted of six commands: swipe (up, down, left, right), circle ''O(select/cancel),'' and word ''L(menu).''

Mid-air gesture was configured using the LEAP motion controller shown in Fig. 3.b and was located below the driver's right hand, based on a previous study [27]. Possible mid-air gestures consisted of six commands: swipe (up, down, left and right), ''gun'' posture (open the thumb and index finger) for select/cancel, and ''menu'' posture (palm toward the body).

Speech used Google Cloud's speech-to-text in Unity to create a voice-commandable interface and defined a list of commands that participants could enter before the experiment. The driver first pulled the lever on the steering wheel to activate the voice recognition and then gave a voice command function through the microphone shown in Fig. 3.c.

Gaze is a way of staring at the icon in a full windshield-type HUD (shown in Fig. 2), and the driver's gaze was tracked through the Tobii 4C eye tracker. Participants conducted eye calibration before using the interface to ensure accuracy of input. The interface was set individually and there was no error in selecting the icon on the HUD during the experiment. There were six Graphical User Interface (GUI) icons used during the interaction process, visualized as opaque in the HUD. In a preliminary study, we found that it is possible to activate the icon for eye-catching input, but that gazing at the HUD for a certain period or more during driving is unsafe [17]. Thus, the driver was asked to press a physical button on the steering wheel (right arrow, Fig. 3.d) to confirm that the eye was fixed to the desired icon (Fig. 2.b). Participants' gaze was recognized to the surrounding 5% area of the opaque GUI icon; recognized areas did not overlap. Auditory feedback for the selection was provided (menu - ''menu''; music - ''music''; navigation - ''navigation''; volume, song, and path selection - ''beep'' sound).

Button was attached to the steering wheel and evaluated for usability by setting it as a control group for other single modalities. The button modality, like other modalities, was designed to control the interface using a total of six buttons (four arrows, menu, and lever shown in Fig. 3.e) to match the comparison conditions.

B. PROCEDURE

We recruited 10 volunteers ($M = 24.0$, $SD = 2.6$, 9 males, 1 female) for our study, all of whom had a valid driver's license. Subjects had the following prior experience with modalities: touch gesture – 8 of 10, mid-air gesture – 6 of 10, speech – 8 of 10, gaze – 4 of 10, button – 10 of 10. After a general introduction to the experiment, eye-tracking calibration was performed to ensure accurate data collection for gaze modalities. Participants practiced driving for 10 minutes to familiarize themselves with the driving simulator. Each session was conducted for 10 minutes, including time spent learning the interface, tests, post-questionnaire (the NASA-TLX and SUS), and a short interview. All participants

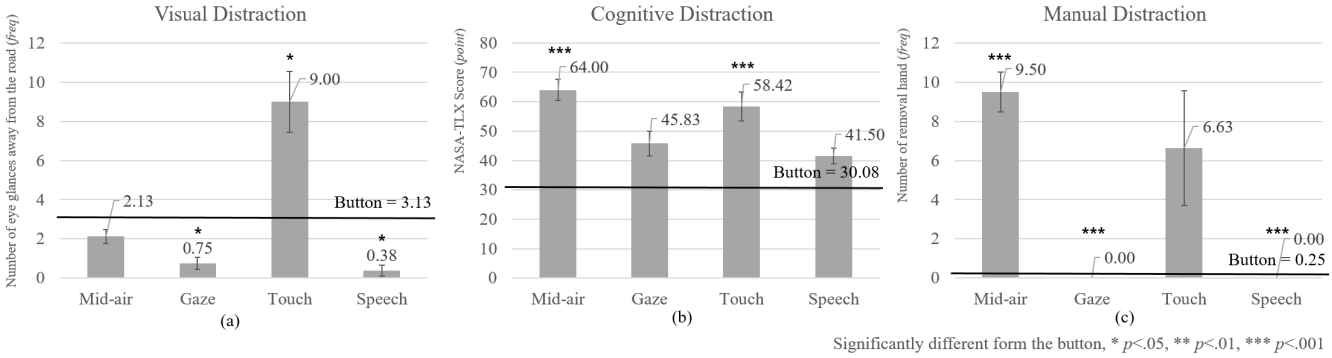

FIGURE 4. Comparing three types of distractions: (a) Visual distraction. (b) Cognitive distraction. (c) Manual distraction.

experienced five single modality interfaces for a total of five sessions, and the order of the provided modalities was adjusted by the Latin Square method to eliminate potential learning effects [31]. The total experiment time was 1 hour (driving simulator practice + five sessions).

C. ANALYSIS

For each session, we collected recorded video and interview responses and analyzed data on eye glance behavior, workload, subjective preference, primary task performance, and secondary task performance. To analyze the entire session, we conducted a statistical analysis of the data, except for the max and min data per session. We analyzed the data in terms of specific distractions. All results were subjected to one-way repeated measures ANOVA tests for post-hoc tests at a 5% confidence level, excluding response to sign, which was not statistically significant (p = ns (0.776)). Equivalence tests were performed to analyze pairwise comparisons (*Bonferroni* corrected or *Games-Howell* corrected as post-hoc tests after checking the homogeneity of variances).

D. RESULTS

1) VISUAL DISTRACTION

We measured visual distraction – the number of eye glances away from the road – to determine how much visual demand the driver experienced when using IVIS. Visual distraction was not the same between modalities, $F\ (4, 35) = 19.69, p < 0.001$. Post hoc analysis (Fig. 4.a) indicates that touch gesture ($M = 9.00, SD = 4.41, p < 0.05$) showed the highest visual distraction compared to other modalities ($M = 3.08, SD = 3.78$). One participant who drove using touch gestures said in an interview *"It was difficult because I had to take my hand off the handle and interact with it, and I kept seeing the touchpad to touch it" (P1)*. Behavior analysis of recorded videos showed that participants' gaze moved from the driving screen to the handle and then to the HUD. On the other hand, the speech interface showed the lowest visual distraction ($M = 0.38, SD = 0.74, p < 0.05$), and many interviewees said it was convenient to not have to move their eyes from the driving screen: "*I don't have to look at the interaction screen, so I can concentrate on driving more" (P2); "It is easy to use" (P10).*

2) COGNITIVE DISTRACTION

Drivers' mental workload was measured by the NASA Task Load Index (NASA-TLX), a tool for assessing subjective mental workload, which measures the level of six dimensions (mental demand, physical demand, temporal demand, effort, performance, and frustration) and determines an overall workload rating. Each dimension is evaluated on a scale of 0 to 100 [32]. As a result, cognitive workload also differed between groups of single modalities, and the ANOVA test also showed statistical significance, $F\ (4, 35) = 12.85, p < 0.001$. As shown in Fig. 4.b, all modalities showed higher NASA-TLX scores than the button modality. The NASA-TLX score was highest for mid-air gestures ($M = 64.00, SD = 10.09, p < 0.001$). Participants mentioned a low recognition rate as the reason why the mid-air gestures resulted in a high mental workload. They also described steering the car with the left hand and making the hand gesture with the right as a cognitive burden: *"The gestures are often unrecognizable and lifting one hand off the steering wheel is burdensome for drivers with both hands" (P3); "Overlap errors with gesture input" (P5)*. The recorded video analysis confirmed that the gesture command was incorrectly recognized due to some participants' unintentional movements. In particular, recognition errors frequently occurred during the sub-process of quickly swiping left, right, up, and down. In contrast to the mid-air gesture, the mental workload of the button interface ($M = 30.08, SD = 8.31$) was lowest among the modalities. Participants commented that the button interface is familiar and comfortable: *"Very easy to input" (P2); "Friendly and convenient" (P3); "Comfortable and fast" (P9).*

3) MANUAL DISTRACTION

To measure manual distraction, we instructed drivers to keep both hands on the steering wheel while driving except when manipulating the interface to carry out the secondary task. We collected behavioral data and counted the number of times each driver removed a hand from the steering wheel. We confirmed by ANOVA test that the mean between

modalities is not the same, F (4, 35) = 10.50, p < 0.001. As shown in Fig. 4.c, the highest manual distraction occurs during mid-air gesture, where the right hand must be unconditionally removed from the steering wheel (M = 9.50, SD = 2.88, p < 0.001). Interestingly, touch gestures have a high distraction compared to other modalities, but not as much as mid-air gestures (M = 6.63, SD = 8.26, p = ns). Because the touch screen is located on the steering wheel, we expected drivers to use their thumb to input command gestures, but video analysis showed that the right hand was removed from the steering wheel to use touch. On the other hand, drivers performed the secondary task using gaze and speech interfaces while keeping both hands on the steering wheel; these interfaces did not generate manual distraction.

4) SUBJECTIVE PREFERENCE

We identified and compared participants' subjective preferences for each interface using the SUS, evaluated on a scale of 0 to 100 in increments of 10. SUS scores of more than 71 points indicate that the system is acceptable [33]. The physical button interface scored highest (M = 80.90, SD = 12.46), followed by gaze (M = 74.70, SD = 11.45) and speech (M = 73.10, SD = 17.77). On the other hand, the SUS score of the touch gesture (M = 60.00, SD = 19.04) and the mid-air gesture (M = 60.94, SD = 15.17) were significantly lower than the critical score of 71. As shown in Table 3, mid-air gesture and touch gesture showed generally low SUS scores and a high level of distraction. Conversely, button, gaze, and speech had high SUS scores and a low level of distraction in general. Therefore, SUS scores are related to the extent of distractions.

5) PRIMARY TASK PERFORMANCE

Drivers using each interface were required to keep the correct lane in response to the LCT sign. We collected two dependent variables: lane deviation (i.e., the number of deviations into the wrong lane, related to driving performance) and response to sign (i.e., the number of failures to appropriately respond to a sign) [14], [16], [30]. ANOVA tests showed lane deviation to be statistically significant, F (4, 35) = 3.12, p <0.05, but response to sign showed no significant difference. As shown in Fig. 5.a, the pairwise comparison showed no significant differences between modalities. However, mid-air gesture (M = 2.75, SD = 2.60, p = ns) showed the highest number of lane deviations, and other interfaces were slightly lower than the button (M = 0.75, SD = 1.39). Response to sign showed only slight performance differences because everyone who participated in this study generally performed the LCT well. Therefore, it was not statistically significant.

TABLE 3. SUS scores, interaction duration and three type of distraction levels per single modality.

Modality	SUS	Interaction duration	Distraction		
			Visual	Cognitive	Manual
Button	80.90	21.13	3.13	30.08	3.28
Gaze	74.70	27.63	0.75	45.83	0.00
Speech	73.10	99.86	0.38	41.50	0.00
Mid-air	60.94	60.17	2.13	64.00	9.50
Touch	60.00	61.38	9.00	58.42	6.63

Lighter colors indicate lower level of distractions;
SUS: higher than 71 points indicates that the system is acceptable;
Visual: mean number of glances at the interface;
Cognitive: mean ratings of NASA-TLX;
Manual: mean number of hand removals from the steering wheel.

6) SECONDARY TASK PERFORMANCE

To measure how well each driver performed their task, we analyzed interaction duration, which is the time between when the instruction is given and when the instruction is completed. Faster interaction duration reduced distractions, and thus led to higher SUS scores. The ANOVA test showed statistical significance, F (4, 32) = 50.78, p < 0.001. As shown in Fig. 5.b, the button had the fastest interaction duration (M = 21.13, SD = 6.45), followed closely by gaze (M = 27.63, SD = 10.49, p = ns). Interaction durations for mid-air (M = 60.17, SD = 11.74, p < 0.001) and touch gesture (M = 61.38, SD = 12.76, p < 0.001) were about three times greater than for the button. Speech showed the slowest interaction duration (M = 99.86, SD = 17.12, p < 0.001).

7) CHARACTERISTIC OF FIVE SINGLE MODALITIES

As expected, we found that each modality caused different types and levels of distraction. Table 3 is a visual representation of the degree of distraction for each modality. Each modality is listed in descending order of SUS scores, and distraction increases as shading gets darker. We found that no single modality shows a low level of distraction in every type and, at the same time, the highest SUS. For example, mid-air gesture and touch gesture interfaces with low SUS scores have high visual, cognitive, and manual distraction, whereas gaze and speech interfaces generally exhibit low distraction and high SUS scores. Although Döring *et al.* [14] showed that low visual distraction and high subjective preference were correlated in a vehicle interface, our results show that the button interface has the highest SUS scores, even though it caused higher visual and manual distractions to drivers than other modalities. These results suggest that even though drivers generally prefer interfaces that cause low visual distractions, they are likely to prefer familiar interface modalities, such as buttons, which are common on car radios.

E. DISCUSSION

In the first study, we looked at the characteristics of each modality by distraction during IVIS use. Presently, NUI research for many IVIS types has been conducted, but each modality has different characteristics depending on the type

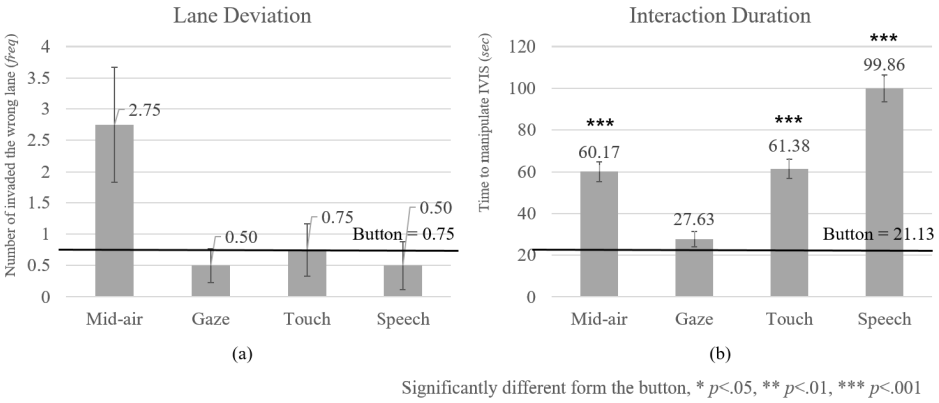

FIGURE 5. Comparing task performances: (a) Lane deviation. (b) Interaction duration.

of distraction. A modality that reduces all three types of distraction is ideal, but there is no such single modality in this study. It is necessary to supplement each modality to lower driver distraction.

As discussed in the Related Works section, Roider *et al.* [22] found that impairments to driver performance were greatest when the input modality for the secondary task overlapped with the sensory channel of the primary task. This suggests that corresponding modalities should be considered, depending on what distraction the researcher wishes to minimize when designing the in-vehicle interface. The obvious advantages and disadvantages of each of the single modalities can be augmented with an additional modality. In response, we also consider the context of secondary tasks. By doing so, we can guess which modalities would be better for drivers in certain interaction steps and suggest appropriate multimodalities by combining a single modality at each step. Based on the characteristics of single modalities that we identified, we can investigate whether proposed cascade-multimodality would increase or decrease total distraction.

Muller and Weinberg [4] considered interaction steps when comparing modalities. They parsed the task of opening a window in a driving situation into its component steps (e.g., thinking, gazing, etc.) and showed how modalities are involved differently in each step. We similarly investigated steps during the interaction tasks of adjusting the music/radio or a using the navigation system (Fig. 6) and found a notable result. As shown in Fig. 6, drivers go through the process, which is divided into two levels (beginning and repetition) while they manipulate IVIS, which becomes more and more complicated as operation time and precision increases.

Level 1 means deciding between the music or navigation category in IVIS, and marks the beginning of when the driver diverts attention from the primary driving task to secondary tasks. Level 1 does not require an iterative command, but represents a cognitive shift. Level 2 denotes repetitive work and precise adjustment, such as selecting the music, destination, and volume, and turning on the map. This is performed

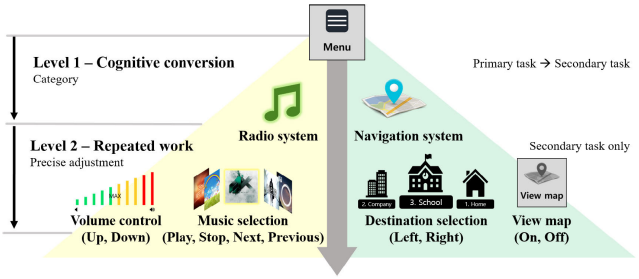

FIGURE 6. Driver's interaction steps with IVIS levels.

without cognitive conversion, but the longer the operation time, the greater the burden of completing the task. Therefore, we suggest a modality with lower cognitive or visual distraction for Level 1 (e.g., gaze or speech in Table 3) and faster interaction duration for Level 2 (e.g., button). In Study 2, we verify whether the suggested modalities lessen overall distraction as cascaded multimodalities.

Mid-air gestures showed high cognitive and manual distraction and low SUS scores. However, considering that existing studies have described the advantages of mid-air gesture, and that mid-air gesture is of ongoing interest [34], [35], it is likely that drivers simply need more experience with it than with other modalities. Participants who first used mid-air gesture said: *"It's amazing in a new way" (P3); "It is more comfortable than I thought" (P5).*

In addition, although the button modality causes severe visual distractions, it showed the highest SUS. This may be due to driver familiarity, which suggests that mid-air gesture could also have high SUS as drivers acclimate to it. Besides, a modality that participants become familiar with quickly could become an NUI [19]. To control familiarity of each modality in an experimental setting, we can measure how fast drivers acclimate to certain modalities – the learning effect. Therefore, in Study 2, we check for the learning effect of each multimodality.

Lastly, researchers design multimodal interfaces by combining single modalities such as speech, gaze, and gesture for one command (e.g., gazing at an object and gesturing toward

it to select) [36], rather than applying them to cascading steps. These studies focus on the synergy effect of single modalities in interaction, rather than on reducing overall driver distraction [26]. Therefore, in the next study, we applied cascaded multimodality to investigate how two modalities can mitigate distractions while manipulating IVIS.

V. STUDY 2: CASCADED MULTIMODALITY NUI

In Study 2, we aimed to suggest combinations of cascaded multimodalities and find the most effective NUIs for our IVIS environment, based on lower driver distraction and higher usability evaluation. For this, we selected and combined the modalities suitable for the user's interaction steps among the single modalities of Study 1, presented them to the driver, and verified their effects. The procedure was the same as the previous experiment.

A. STUDY DESIGN

Based on the results of Study 1, we improved the position of the touch gesture from the upper-center of the steering wheel to near both thumbs. We also replaced the 2.8'' TFT capacitive touch shield input device with the MPR121 capacitive touch sensor to improve recognition. Based on opinions of participants in the first experiment, we changed the input UI for gaze from the GUI button icon on the HUD to the music and navigation GUI icon (e.g., 'To play a different song, stare at/select the music icon from the menu and look at/select the right side of the music list twice'). We also modified the mid-air gesture's behavior for quick and accurate recognition (from 'gun' posture to 'palm down' posture).

Following Muller and Weinberg [4]'s interaction steps and our discussion of Study 1, we defined a Level 1 and Level 2 as drivers interact with the IVIS while driving. The Level 1 interface for calling the menus and selecting the desired system was designed to make commands intuitive. In addition to including the most-preferred single modality (i.e., button for Level 1 and Level 2), the interface included 1) speech that causes low visual and mental workloads and does not overlap with the primary task and 2) gaze that does not significantly increase the driver's physical workload and limits cognitive workload in the input process. On the other hand, for the modalities for Level 2, buttons and touch gestures were selected for their ability to repeat, their precise adjustment, and their effect of reducing the overall interaction duration with fast input. We also included potential mid-air gestures in Level 2. Accordingly, our research hypotheses are as follows.

- **H1**: Cascaded multimodality, which specifies speech and gaze as the modality of Level 1, will exhibit lower distraction and higher usability than a single modality.
- **H2**: Multimodalities that specify a button or touch gesture as a Level 2 modality will show faster interaction duration and higher usability than other modalities.

To test the above hypothesis, the cascaded multimodality NUI was constructed as shown in Table 4. All modalities, except speech, were used for Level 2; speech (which has a high interaction duration; see Fig. 5.b) was excluded because of its weakness for repetitive work. Therefore, gaze + speech nor button + speech were designed in Study 2. Furthermore, our interfaces were designed not to respond to commands according to modalities other than those being tested to prevent duplicate inputs [24] and better process interaction flow [4].

TABLE 4. Cascaded multimodality combinations by level.

Combinations	Level 1	Level 2
1	Button	Button
2	Gaze	Mid-air
3	Gaze	Button
4	Gaze	Touch
5	Speech	Mid-air
6	Speech	Button
7	Speech	Gaze
8	Speech	Touch

As in Study 1, we used a HUD as an output device and collected the recorded video, driving performance, task performance, and behavior data of the participants. Participants completed a brief interview and responded to the Driving Activity Load Index (DALI) and SUS questionnaire. All processes were conducted equally for two days, in order to account for the learning effect.

B. PROCEDURE

25 subjects ($M = 23.9$, $SD = 2.7$, 16 males, 9 females) participated in the study for compensation. The experimental procedure was the same as for Study 1, except for the type and number of interfaces. All participants had a valid driver's license and at least 1 year of driving experience, except four people. Participants already had the following experience with modalities: touch gesture – 20 of 25, mid-air gesture – 10 of 25, speech – 24 of 25, gaze – 10 of 25, button – 25 of 25. The experiment lasted 1 hour 30 minutes (driving simulator practice + eight sessions). To investigate the learning effect, all participants performed the same experiment again at the same time the next day. Therefore, the total experiment time was 3 hours (Day 1 + Day 2).

C. ANALYSIS

We performed a statistical analysis, except for some data that could not be collected due to an occasional connection problem between sensors and computer systems (all mid-air gesture – 7 of 25, gaze + touch – 3 of 25, speech + touch – 2 of 25, speech + button – 1 of 25). Results were subjected to one-way repeated measures ANOVA tests and after checking the homogeneity of variances, *Bonfeffoni* or *Games-Howell* as post-hoc tests at a 5% confidence level in common with Study 1 (excluding response to sign, which was not statistically significant in Study 1). Additionally, we analyzed primary and secondary task performance to paired t-test to find the learning effect between Day 1 and Day 2.

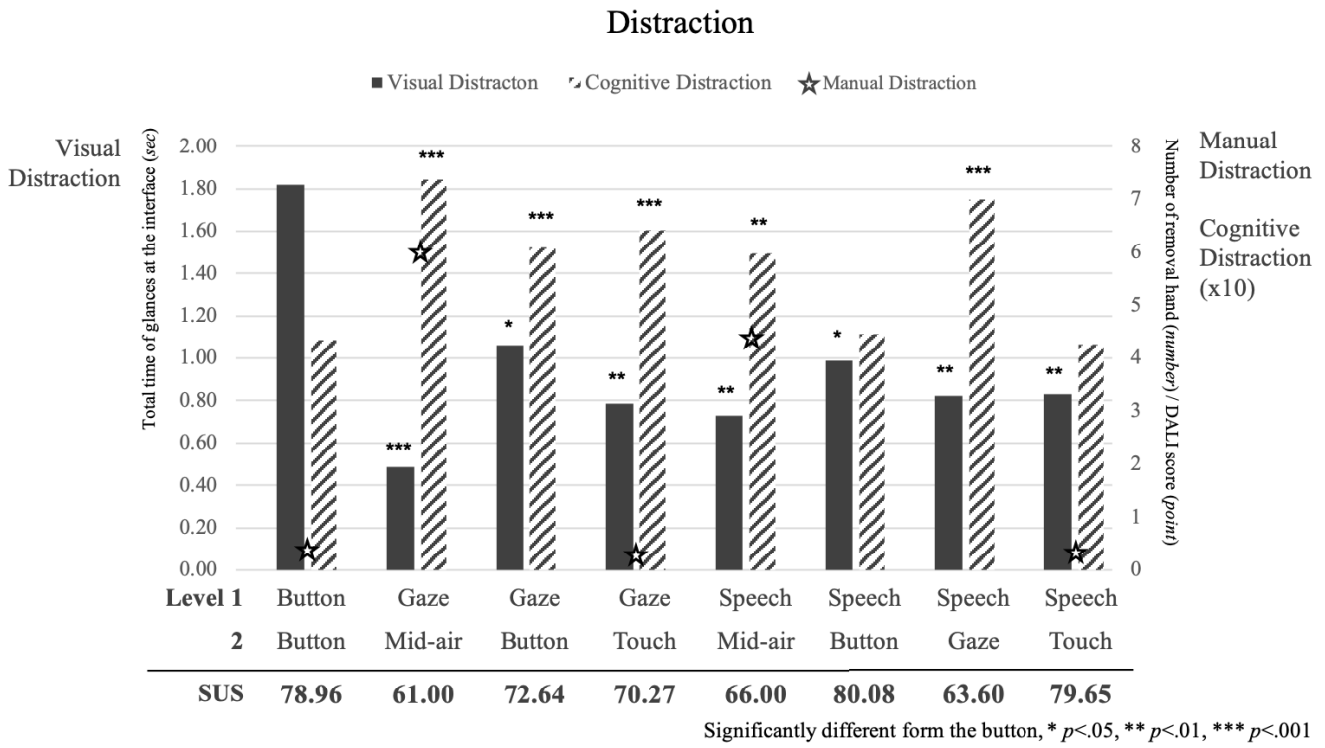

FIGURE 7. Comparing SUS scores and three different distraction degrees of cascaded multimodalities.

D. RESULTS

Visual distraction was calculated as the total time required to glance at the interface. Cognitive distraction was measured by the DALI questionnaire, which is better for assessing mental workload in the driving context than NASA-TLX, as there are more independent variables [37]. Manual distraction was determined by the number of hand removals from the steering wheel. We measured driving performance and task performance to judge whether these interfaces improved performance. As in Study 1, lane deviation and interaction duration were collected.

1) H1: CASCADED MULTIMODALITY (SPEECH + BUTTON, GAZE + BUTTON), WHICH SPECIFIES SPEECH AND GAZE AS THE MODALITY OF LEVEL 1, WILL EXHIBIT LOWER DISTRACTION AND HIGHER USABILITY THAN A SINGLE MODALITY (BUTTON)

Fig. 7 shows the results of multimodality by distraction. Compared to the physical button, which is a single modality, we see lower visual distraction ($M = 1.82$, $SD = 0.97$) in all multimodalities, $F\ (7, 172) = 7.45$, $p < 0.001$. In particular, the gaze + button ($M = 1.06$, $SD = 0.67$, $p < 0.05$) and speech + button ($M = 0.99$, $SD = 0.69$, $p < 0.05$) modalities have lower visual distraction than the singular button modality, which means that drivers experience less visual burden when they begin performing secondary tasks with gaze and speech modality. However, in terms of cognitive distraction, $F\ (7, 172) = 16.43$, $p < 0.001$ (Fig. 7), the results of the gaze + button ($M = 60.95$, $SD = 16.74$, $p < 0.001$) and speech + button ($M = 44.63$, $SD = 12.54$, $p =$ ns) modalities were worse than the button ($M = 43.49$, $SD = 16.07$), perhaps because the button has much lower cognitive distraction compared to other modalities, such as gaze and speech (Fig. 4.b). In terms of physical load, no manual distraction occurred, even though the button modality was used at Level 2. By analyzing a recorded video of users' actions for Study 2, we found that participants easily operated the interface when doing repetitive tasks. In the usability evaluation (Fig. 7), the gaze + button ($M = 72.64$, $SD = 12.11$) and the speech + button ($M = 80.08$, $SD = 10.83$) modalities showed high usability scores, exceeding the 71-point threshold. In particular, the usability score of speech + button was higher than button only ($M = 78.96$, $SD = 11.73$), which was the highest among single modalities. In this study, speech + button exhibited lower distraction and higher usability than a single modality. Similarly, gaze + button showed lower distraction, except for cognitive workload. Hence, H1 was supported.

2) H2: MULTIMODALITIES (SPEECH/GAZE + BUTTON, SPEECH/GAZE + TOUCH) THAT SPECIFY A BUTTON OR TOUCH GESTURE AS THE LEVEL 2 MODALITY WILL SHOW FASTER INTERACTION DURATION AND HIGHER USABILITY THAN OTHER MODALITIES

To perform precise adjustments, the input must be fast and repetitive operations must be performed quickly.

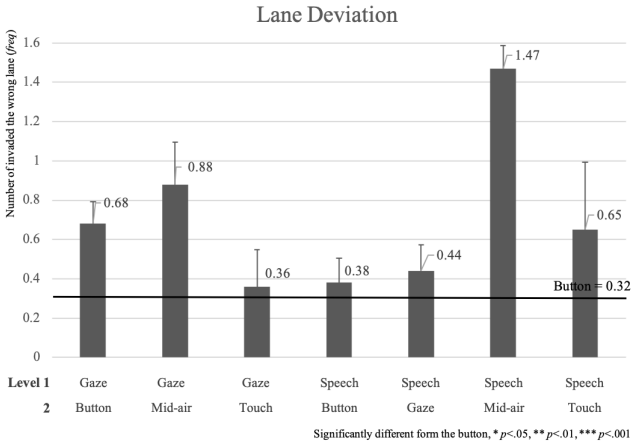

FIGURE 8. Comparing lane deviation of cascaded multimodalities.

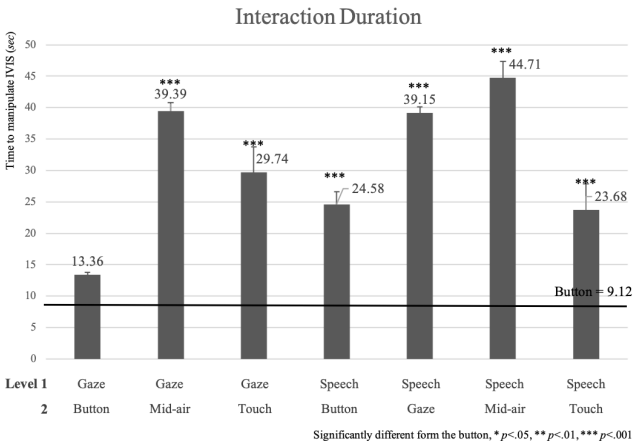

FIGURE 9. Comparing interaction duration of cascaded multimodalities.

Level 2 modalities suitable for this role are buttons and touch gestures, which are familiar because they are used in common products, like electronic devices, smartphones, and touchpads. Fig. 9 shows the interaction duration by modality. Except for button only, the gaze + button ($M = 13.36$, $SD = 6.86$, p = ns) had the fastest interaction duration, followed by speech + touch ($M = 23.68$, $SD = 4.70$, $p < 0.001$), speech + button ($M = 24.58$, $SD = 4.59$, $p < 0.001$), and gaze + touch ($M = 29.74$, $SD = 9.66$, $p < 0.001$). All multimodal interfaces received usability scores of 71 or more, except for the gaze + touch, which had the longest interaction duration. In particular, speech + button ($M = 80.08$, $SD = 10.83$) and speech + touch ($M = 79.65$, $SD = 9.72$) received a higher SUS score than the button modality ($M = 78.96$, $SD = 11.73$). In summary, in Level 2, gaze + button had a similar interaction duration to button only, and gaze + touch, speech + button, and speech + touch had faster interaction duration than both mid-air and gaze. These combinations showed high usability, except for gaze + touch. Thus, H2 was supported.

3) PRIMARY TASK PERFORMANCE AND SECONDARY TASK PERFORMANCE

Fig. 8-11 show how the proposed cascaded multimodal interfaces improve driving performance and safety, in terms of lane deviation, $F(7, 170) = 4.05$, $p < 0.001$, and interaction duration, $F(7, 172) = 33.10$, $p < 0.001$. Response to sign was not statistically significant. The lane deviation graph shown in Fig. 8 indicated no significant difference when all interfaces were compared with the button interface. In fact, it showed less than one occurrence of lane deviation, except for speech + mid-air ($M = 1.47$, $SD = 1.42$, $p = 0.075$). As shown in Fig. 9, other modalities generally required longer interaction duration than the button ($M = 9.12$, $SD = 2.33$). But gaze + button did not differ significantly from the button alone ($M = 13.36$, $SD = 6.86$, p = ns), so it cannot be determined whether interaction duration was consistently greater. Therefore, we interpret gaze + button as showing interaction duration as fast as the button (see Fig. 5.b and 9).

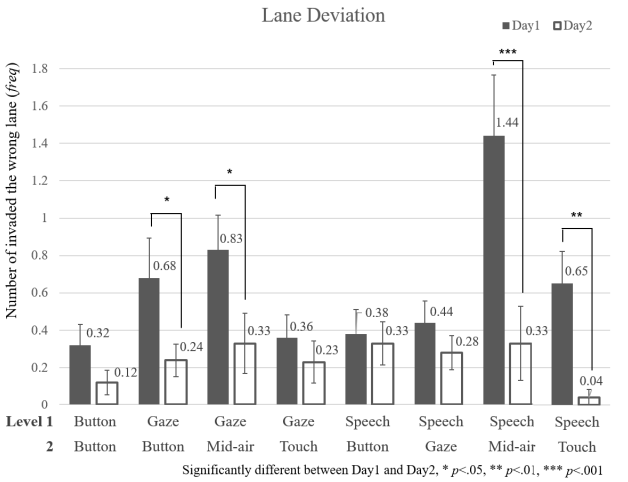

FIGURE 10. Comparing lane deviation of cascaded multimodality considering learning effect.

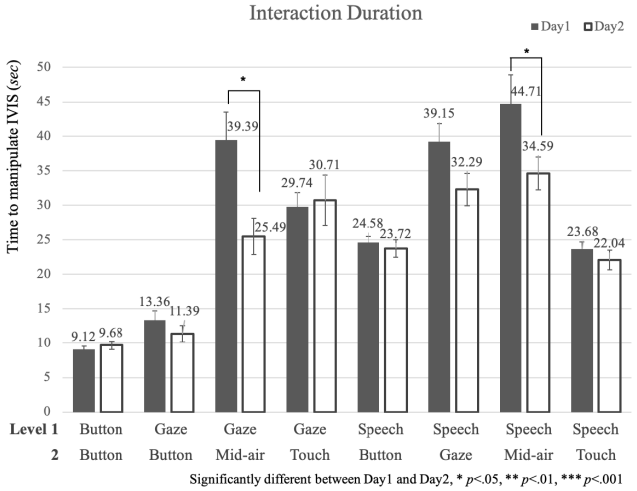

FIGURE 11. Comparing interaction duration of cascaded multimodality considering learning effect.

We conducted the same experiment for two days to see how the performance with the proposed cascaded multimodal interfaces might improve through the learning effect (Fig. 10-11). Driving performance on Day 2 markedly

improved from Day 1, as shown in Fig. 10 ($M = 0.64$, $SD = 0.37$, Day 1; $M = 0.24$, $SD = 0.11$, Day 2). In particular, lane deviation for gaze + button ($M = 0.68$, $SD = 1.07$, Day 1; $M = 0.24$, $SD = 0.44$, Day 2; $t\,(24) = 2.19$, $p < 0.05$), gaze + mid-air ($M = 0.83$, $SD = 0.79$, Day 1; $M = 0.33$, $SD = 0.69$, Day 2; $t\,(17) = 2.47$, $p < 0.05$), speech + touch ($M = 0.65$, $SD = 0.83$, Day 1; $M = 0.04$, $SD = 0.21$, Day 2; $t\,(22) = 3.48$, $p < 0.01$), and even speech + mid-air ($M = 1.44$, $SD = 1.38$, Day 1; $M = 0.33$, $SD = 0.84$, Day 2; $t\,(17) = 4.89$, $p < 0.001$) decreased and appeared statistically significant in terms of learning. Likewise, in Fig. 11, we found interaction duration improved overall, as the mean interaction duration for Day 2 is less than the mean for Day 1 ($M = 27.97$, $SD = 12.75$, Day 1; $M = 23.74$, $SD = 9.22$, Day 2). Surprisingly, however, interaction duration did not change when drivers used a familiar button and touch modality at Level 2, despite the new interface they experienced. When the unfamiliar mid-air gesture and gaze were applied to Level 2, improvement in task performance can be attributed to learning. Gaze + mid-air ($M = 39.39$, $SD = 17.12$, Day 1; $M = 25.49$, $SD = 11.02$, Day 2; $t\,(17) = 2.46$, $p < 0.05$) and speech + mid-air ($M = 44.71$, $SD = 17.73$, Day 1; $M = 34.59$, $SD = 10.11$, Day 2; $t\,(17) = 2.72$, $p < 0.05$) showed statistical significance.

Interestingly, we found a powerful learning effect when using air gestures as a modality for Level 2. Of course, on Day 2, the interaction duration for speech + mid-air was still longer than for the other modalities, but the learning effect was shown and the gap with other modalities was narrowed. This suggests that, with more experience, mid-air gesture could be as effective or even more effective than other modalities. Participants who used the mid-air gestures said in an interview *"Satisfaction and convenience have increased compared to the previous day. It seems better than voice because there is no delay in interaction" (P4); "If I get used to the air gesture, I can control it quickly" (P7).*

In summary, in this paper, we show that the Level 1 modality causes low visual distraction, so that speech and gaze are suitable for secondary task switching. As a modality of Level 2, we show that the button and touch, which can perform fast input, are suitable. Additionally, the multimodalities of gaze + button, speech + button, and speech + touch are considered to be the most effective multimodal NUIs for IVIS because of their lower lane deviation, shorter interaction duration, and higher usability. In particular, speech + button and speech + touch performed better across the three types of distraction than button only (which caused the lowest cognitive distraction in Study 1).

E. DISCUSSION

In this study, we proposed a cascaded multimodal interface, which reduces distraction and increases usability. The driver's interaction with IVIS was divided into two steps (Level 1 and Level 2), and then combined with a single modality suitable for each step. We found that reducing the overall interaction duration is paramount for manipulating secondary tasks so that drivers can initiate and control the desired performance with minimal cognitive load and can resume concentration on the primary tasks. As shown in Fig. 6, 1) In the initiation step of secondary tasks at Level 1, the driver must be given the minimum mental load to enable efficient cognitive transitions, and 2) it must be possible to reduce the overall interaction duration by enabling fast input for repetitive manipulation. For these reasons, simple modalities, such as button and touch, were preferred by drivers in Level 2. However, we will also discuss the impact of familiarity in the following paragraphs.

In this experiment, speech and gaze were identified as an efficient modality at Level 1, which requires cognitive conversion. Speech modality has a lower cognitive burden because it relies on drivers' verbal resources, rather than the manual or visual resources that are already engaged in primary and secondary driving tasks. Although the gaze modality overlaps with the visual requirements of the driving situation (attention to the front), it is an intuitive modality that has the advantage of being able to see and select what is desired without additional cognitive processes, thereby reducing the overall cognitive burden on the user. Also, the gaze modality was found to be suitable for performing the role of Level 1 because it minimized the driver's visual distraction by unifying the input and output device using the HUD. Consequently, we propose that effective modalities for Level 1 are speech and gaze.

Button and touch gesture were found suitable as Level 2 modalities that perform repeated command entry rapidly. A single button modality, common in passenger vehicles, showed high visual distraction even though it had the fastest interaction duration and high SUS. We noted in Study 1 that a cognitive transition occurs in order for the driver to initiate the secondary task. Therefore, shifting the driver's attention to the button while driving creates more visual and cognitive confusion and allows the driver's line of sight to stay on the button longer. However, buttons have the advantage of being easy to operate, fast, and above all, familiar. For this reason, a button modality can be a good choice for the Level 2 process of repeatedly adjusting the radio or navigation. The interface with the button on Level 2 actually reduced visual distraction. Interestingly, among the modalities used in Study 2, touch gestures showed the biggest difference from Study 1. In Study 1, the touch gesture was generally high in all distractions; it was highest in visual distraction and lowest in SUS (Table 3). However, when the touch gesture was used as the modality for Level 2 in Study 2, manual distraction hardly occurred and the usability evaluation was high. In the case of speech + touch, all distractions were lower than the button modality. The difference in these results suggests that the same modality can affect driver distraction differently depending on the interface design. Consequently, we consider the button and touch gesture appropriate modalities for Level 2.

In Study 1, the touch screen was positioned at the top center of the steering wheel for input commands. This interface increased visual and manual distraction by requiring the operator to take a hand off the steering wheel. However, in

Study 2, the touchpad was attached to the steering wheel so that the driver could input commands using their thumb and without taking their hands off the steering wheel. Furthermore, distraction was greatly reduced by designing commands as inputs with only three actions (tap, swipe left, and swipe right).

In addition, we identified an advantage of speech. Our findings differ from previous studies, which showed that speech modality harms driving performance via a high cognitive load [38]–[40]. We believe this is because our study investigates not only which modality to choose, but also how to design the interface with that modality. Future NUI research should more fully consider interface design.

Considering the driver interaction process (Level 1 and Level 2) using IVIS, cascaded multimodalities provided as a combination of single modalities resulted in low distraction and high usability. Participants stated in interviews: *"It's quick and easy to use because it's easy to select a menu with just a glance and a sub-item with a button" (P18); "Voice recognition makes it easy to bring up the menu. It is good to make an immediate response by button operation. It's great to use content while concentrating on driving" (P17); "It was good that detailed adjustment was easy" (P12).*

Finally, after analyzing the button of Study 1, we suggested its characteristics – high visual distraction but fast interaction duration and high SUS – were the result of drivers' extended experience and familiarity. In Study 2, we investigated over two days the possibility that other modalities might be as familiar to drivers and which modalities they might learn quickly. We found that mid-air gesture reduced interaction duration and lane deviation, and it was statistically significant. Although the results in Study 1 suggest that mid-air gesture is not a suitable interaction technique, participants learned it quickly in Study 2, which suggests that it could be an NUI in driving. Compared to button modality, mid-air gesture causes less visual distraction (see Table 3). Also, visual distraction was high when all interfaces were combined with button modality in Study 2 (see Fig. 7), yet lowest when configured with mid-air gesture. On the other hand, mid-air gesture leads to manual distraction (participants must remove their hands from the steering wheel) and cognitive distraction, which causes long interaction duration. To resolve these disadvantages, Nacenta *et al.* [19] proposed enhanced usability and memory in mid-air gesture, and Werner [41] proposed input gestures that do not require drivers to take their hands off the steering wheel. Therefore, as drivers gain experience with mid-air gesture and its techniques are improved, the modality can reduce manual and cognitive distractions, like the button modality. Also, since mid-air gesture can activate more diverse inputs than buttons, it may have potential as a multimodal NUI.

VI. LIMITATIONS AND FUTURE WORKS

Our experiments compared the three types of distraction for five NUIs for IVIS. Although researchers have used the LCT as a driving task [14], it cannot reflect actual in-vehicle interaction situations, like a sharp curve, obstacle on the road, or sudden stop. If we apply variable situations to driving environments in future work, we may be able to observe more detailed characteristics and further refine our combinations of modalities according to the situation. However, in LCTs similar to highway driving, we discovered quite meaningful characteristics of single modalities and demonstrated the value of cascaded multimodal NUIs in IVIS.

Before designing the study, we defined the scenario – the radio and navigation system of IVIS – through a preliminary survey. Many researchers have already experimented with radio and navigation systems [11], [14], and they have been investigated as the most frequently used system in actual driving situations [7]. Based on this, our infotainment systems were designed to match the most common functions and features we defined in the preliminary survey. Our systems do not cover overlapping events, such as talking on the phone during manipulation of the navigation system and receiving a message while controlling volume. Ideally, we will study NUIs in the context of these overlapping events in the future.

In our study, we used our own HUD as an output device for the interface. Although a HUD is known to reduce the driver's visual distraction, it requires careful design. For example, opaque icons in the HUD could create blind spots that block drivers from seeing road signs or pedestrians and thus cause extra visual demand. Fortunately, participants in our study did not report any inconvenience in its use while driving. However, in our future studies where driving maps more accurately reflect real driving scenarios (e.g., other cars in the road, buildings, pedestrians), we will consider icon location and transparency in the HUD.

In Study 1, we designed an interface to compare characteristics of five modalities. In Study 2, we designed a multimodal interface to evaluate the usability of a combination of modalities. In the interface system, we designed the menu structure identically for all interaction modalities; the number of interactions required to complete the proposed tasks was likewise identical across interaction modalities. Our speech recognition interface, for instance, did not use natural language widely available in Apple Siri or Amazon Alexa (e.g., ''Alexa, play the second song''), so participants spoke single word commands such as ''Up'' or ''Down.'' To compare the interaction modalities rather than whole interfaces, we designed all input modalities (including speech) to hierarchically manipulate IVIS to follow single word commands. Because we unified the interface manipulation method to a hierarchical structure for equal comparative research, each modality could not take advantage of an intuitive instruction modality. Nevertheless, the results of this study ease comparative research, clearly identifying the disadvantages of each modality and showing how those characteristics arise in multimodality. Since the results of this study were obtained from a hierarchical interface, other interface types (e.g., no hierarchical menu structure) could lead to different results. In future studies, we will consider a method to apply

advanced technologies for each modality and compare their characteristics.

We did not allow participants to select their own input modality (i.e., a ''free choice'') because of concerns about bias caused by driver experience, as buttons were most preferred in Study 1. In particular, the multimodal interface we presented in Study 2 includes interfaces that many drivers have never experienced, so it is highly likely that they would have selected the driver-friendly (i.e., familiar) buttons or interfaces. Therefore, we analyzed results from a study in which all participants used the same interface to present a combination of modalities based on each modality's characteristics and interaction steps of the secondary task. We will find input modalities suitable for specific interactions if the modalities proposed become familiar to drivers. The results of Study 2 show that each modality of different interaction tasks may suit different drivers at different times because different interfaces have different driver distractions. Although we did not allow participants to freely select a modality, future studies could consider designing and testing an interface that would respond to whichever modality participants would choose at Level 2, i.e., participants could initiate a command using a button (Level 1) and then choose speech or mid-air gesture at Level 2. Furthermore, testing under different conditions could add valuable insights; participants might choose speech at Level 2 in heavy traffic conditions but gaze input in low traffic conditions. Therefore, we might be able to find a user-defined cascaded multimodal interface appropriate at each scenario in the future.

To measure visual demand, we measured the number of eye movements from the road to the interface (Study 1) and the total time the driver's eyes stayed at the interface (Study 2). Liang and Lee [42] showed that these two indicators were proportional. Both measures clearly show the difference in visual demand between modalities but would provide a deeper understanding if various indicators (e.g., total off-road time and number of long [e.g., 2+ sec] glances) were measured and analyzed together. Therefore, we would like to consider the various metrics of visual, cognitive and physical demand in future studies.

The technical problem noted in Study 1 was not completely resolved in Study 2. Low recognition rates reported for mid-air gestures delayed some input commands. Occasionally, communication between the computer and the sensor was a problem. Therefore, delays for the mid-air gesture and some touch gestures may have made interaction difficult. Due to this error, some data were unusable. However, we had enough data that we were able to exclude data from the error and still produce meaningful results. We will resolve these technical errors for future studies.

Study 2 confirmed the learning effect for mid-air gesture, but actual learning can take significant time and practice for a driver in real road scenarios. To address this concern, we will consider taking at least two weeks to check how the learning effect of five modalities changes over time. It is also important to consider the possibility of different modality learning effects. Mid-air gestures showed rapid learning over the course of the two studies, but other modality combinations (e.g., gaze + button, speech + gaze) also showed reduced interaction duration. Longitudinal studies would give us a more certain and accurate understanding of the impact of learning effects.

Researchers who study vehicle interfaces have tried various scenarios in an indoor driving simulator to create an environment similar to the actual driving situation (e.g., lane change task, highway, urban & rural city, pedestrian, vehicle crash, obstacle, etc.). Therefore, it is necessary to verify the effect of NUIs in situations where driving is more complicated. As primary tasks become more difficult, such as road environments that require a right or left turn (rather than a straight road), various weather (e.g., rainy, snowy) or the presence of other vehicles, using IVIS puts a greater cognitive burden on an already cognitively burdened driver. We expect that multimodalities can be easily defined and analyzed in this scenario using our cascaded approach, as well. Therefore, based on the results of this study, we will re-verify the cascaded multimodality in a driving simulation that is closer to the actual driving situation.

VII. CONCLUSION

Through two consecutive studies, we identified cascaded multimodality combinations to reduce driver distraction and increase usability for IVIS. We proposed important considerations for designing a novel cascaded multimodality: the characteristics of three types of driver distraction and the driver's interaction step.

In Study 1, we compared the four NUI single modalities (touch, mid-air, speech, and gaze) with a button interface that drivers use most often to interact with IVIS across three aspects of distraction (visual, cognitive, and manual or physical). This provided a holistic view of the five modalities and explored the characteristics, advantages, and disadvantages of each modality for each type of distraction. By analyzing the interaction process that occurs when the driver interacts with IVIS, we identified two steps of cognitive transition: 1) from the primary to the secondary task, and 2) repetitive and precise adjustment in the secondary task. We defined these as the user's interaction steps. By analyzing the characteristics of the button interface, we demonstrated the need to consider the learning effect when comparing modalities.

In Study 2, we found a combination of cascaded multimodalities to minimize the driver's distraction and maximize usability. We collected data from driving performance, task performance, eye glance, recorded video analysis, questionnaire, and interviews. We divided the driver's interaction with IVIS into two steps, and we proposed a suitable single modality for each step. We suggested the possible influence of a learning effect in Study 1, and we confirmed the existence of a learning effect by experimenting for two days in Study 2.

Overall, these studies suggest a novel approach that accounts for multiple types of distractions, interaction steps,

combinations of cascaded multiple modalities, and learning effects. Our empirical results can provide a good starting point for more applied studies in this context. Our initial results can guide future evaluation methods and inform a rigorous approach to reducing driver distraction.

REFERENCES

[1] M. A. Regan, J. D. Lee, and K. Young, *Driver Distraction: Theory, Effects, and Mitigation*. Boca Raton, FL, USA: CRC Press, 2008, pp. 31–51.

[2] N. Dibben and V. J. Williamson, "An exploratory survey of in-vehicle music listening," *Psychol. Music*, vol. 35, no. 4, pp. 571–589, Oct. 2007, doi: 10.1177/0305735607079725.

[3] K. Kircher, "Driver distraction: A review of the literature," in *Proc. Statens Väg-och Transp. Skningsinstitut*, 2007, pp. 15–30.

[4] C. Müller and G. Weinberg, "Multimodal input in the car, today and tomorrow," *IEEE MultimediaMag.*, vol. 18, no. 1, pp. 98–103, Jan. 2011, doi: 10.1109/MMUL.2011.14.

[5] National Highway Traffic Safety Administration, "Visual-manual NHTSA driver distraction guidelines for portable and aftermarket devices," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. 2016-29051, Dec. 2016. [Online]. Available: https://www.federalregister.gov/documents/2016/12/05/2016-29051/visual-manual-nhtsa-driver-distraction-guidelines-for-portable-and-aftermarket-devices

[6] P. Schroeder, M. Wilbur, and R. Pena, "National survey on distracted driving attitudes and behaviors-2015," Nat. Highway Traffic Saf. Admin., Washington, DC, USA., Tech. Rep. HS 821, vol. 416, Mar. 2018. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13123-2015_natl_survey_distracted_driving_031418_v5_tag.pdf

[7] T. A. Ranney, W. R. Garrott, and M. J. Goodman, "NHTSA driver distraction research: Past, present, and future," SAE, Warrendale, PA, USA, Paper no. 233, 2001. [Online]. Available: https://www.sae.org/publications/technical-papers/content/2001-06-0177/preview/

[8] D. Wigdor and D. Wixon, *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Amsterdam, The Netherlands: Elsevier, 2011.

[9] J. Maciej and M. Vollrath, "Comparison of manual vs. speech-based interaction with in-vehicle information systems," *Accident Anal. Prevention*, vol. 41, no. 5, pp. 924–930, Sep. 2009, doi: 10.1016/j.aap.2009.05.007.

[10] L. Angelini, J. Baumgartner, F. Carrino, S. Carrino, M. Caon, O. A. Khaled, J. Sauer, D. Lalanne, E. Mugellini, and A. Sonderegger, "A comparison of three interaction modalities in the car: Gestures, voice and touch," in *Proc. Interact. Homme-Mach. (IHM)*, New York, NY, USA, 2016, pp. 188–196, doi: 10.1145/3004107.3004118.

[11] K. M. Ba H, M. G. Jáger, M. B. Skov, and N. G. Thomassen, "You can touch, but you can't look: Interacting with in-vehicle systems," in *Proc. 26th Annu. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2008, pp. 1139–1148, doi: 10.1145/1357054.1357233.

[12] F. Roider, S. Rämelin, B. Pfleging, and T. Gross, "Investigating the effects of modality switches on driver distraction and interaction efficiency in the car," *J. Multimodal User Interface*, vol. 13, no. 2, pp. 89–97, Mar. 2019, doi: 10.1007/s12193-019-00297-9.

[13] P. Biswas, G. Prabhakar, J. Rajesh, K. Pandit, and A. Halder, "Improving eye gaze controlled car dashboard using simulated annealing," in *Proc. 31st Brit. Comput. Soc. Hum. Comput. Interact. Conf.*, 2017, p. 39, doi: 10.14236/ewic/HCI2017.39.

[14] T. Döring, D. Kern, P. Marshall, M. Pfeiffer, J. Schöning, V. Gruhn, and A. Schmidt, "Gestural interaction on the steering wheel: Reducing the visual demand," in *Proc. Annu. Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2011, pp. 483–492, doi: 10.1145/1978942.1979010.

[15] B. Pfleging, S. Schneegass, and A. Schmidt, "Multimodal interaction in the car: Combining speech and gestures on the steering wheel," in *Proc. Autom.*, vol. 12, New York, NY, USA, 2012, pp. 155–162, doi: 10.1145/2390256.2390282.

[16] J. Sterkenburg, S. Landry, and M. Jeon, "Design and evaluation of auditory-supported air gesture controls in vehicles," *J. Multimodal User Interface*, vol. 13, no. 2, pp. 55–70, Mar. 2019, doi: 10.1007/s12193-019-00298-8.

[17] R. S. McCann, D. C. Foyle, and J. C. Johnston, "Attentional limitations with head-up displays," in *Proc. 7th Int. Symp. Aviation Psychol.*, Columbus, OH, USA, 1993, pp. 70–75. [Online]. Available: https://hsi.arc.nasa.gov/groups/HCSL/publications/McCann_AvPsyc93.pdf

[18] C. A. Pickering, K. J. Burnham, and M. J. Richardson, "A research study of hand gesture recognition technologies and applications for human vehicle interaction," in *Proc. 3rd Inst. Eng. Technol. Conf. Automot. Electron.*, Warwick, U.K., 2007, pp. 1–15.

[19] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson, "Memorability of pre-designed and user-defined gesture sets," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2013, pp. 1099–1108, doi: 10.1145/2470654.2466142.

[20] S. Oviatt and P. R. Cohen, "The paradigm shift to multimodality in contemporary computer interfaces," *Synth. Lectures Hum.-Centered Informat.*, vol. 8, no. 3, pp. 1–243, Apr. 2015, doi: 10.2200/S00636ED1V01Y201503HCI030.

[21] S. Koyama, M. Inami, Y. Sugiura, M. Ogata, A. Withana, Y. Uema, M. Honda, S. Yoshizu, C. Sannomiya, and K. Nawa, "Multi-touch steering wheel for in-car tertiary applications using infrared sensors," in *Proc. 5th Augmented Hum. Int. Conf.*, New York, NY, USA, 2014, pp. 1–4, doi: 10.1145/2582051.2582056.

[22] F. Roider, S. Rämelin, B. Pfleging, and T. Gross, "The effects of situational demands on gaze, speech and gesture input in the vehicle," in *Proc. Autom.*, vol. 17, New York, NY, USA, 2017, pp. 94–102, doi: 10.1145/3122986.3122999.

[23] S. N. Moran, T. Z. Stybel, G. M. Handcock, and K. P. L. Vu, "Using the lane change test to investigate in-vehicle display placements," in *Int. Conf. Appl. Hum. Factors Ergonom.* Cham, Switzerland: Springer, 2019, pp. 596–607, doi: 10.1007/978-3-030-20503-4_54.

[24] F. Roider and T. Gross, "I see your point: Integrating gaze to enhance pointing gesture accuracy while driving," in *Proc. 10th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, New York, NY, USA, Sep. 2018, pp. 351–358, doi: 10.1145/3239060.3239084.

[25] J. A. Jacko, *Human Computer Interaction handbook: Fundamentals Evolving Technologies and Emerging Applications*. Boca Raton, FL, USA: CRC Press, 2012.

[26] M. Gondan, K. Lange, F. Rösler, and B. Röder, "The redundant target effect is affected by modality switch costs," *Psychonomic Bull. Rev.*, vol. 11, no. 2, pp. 307–313, Apr. 2004, doi: 10.3758/BF03196575.

[27] R. Haeuslschmid, Y. Shou, J. O'Donovan, G. Burnett, and A. Butz, "First steps towards a view management concept for large-sized head-up displays with continuous depth," in *Proc. 8th Int. Conf. Automot. User Interface Interact. Veh. Appl. Automot.*, New York, NY, USA, 2016, pp. 1–8, doi: 10.1145/3003715.3005418.

[28] D. Beck, J. Jung, J. Park, and W. Park, "A study on user experience of automotive HUD systems: Contexts of information use and user-perceived design improvement points," *Int. J. Hum.–Comput. Interact.*, vol. 35, no. 20, pp. 1936–1946, Mar. 2019, doi: 10.1080/10447318.2019.1587857.

[29] S. Mattes, "The lane-change-task as a tool for driver distraction evaluation," *Qual. Work Products Enterprises Future*, vol. 57, p. 60, Oct. 2003. [Online]. Available: https://www-nrd.nhtsa.dot.gov/departments/nrd-01/IHRA/ITS/MATTES.pdf

[30] E. Mitsopoulos-Rubens, M. J. Trotter, and M. G. Lenné, "Effects on driving performance of interacting with an in-vehicle music player: A comparison of three interface layout concepts for information presentation," *Appl. Ergonom.*, vol. 42, no. 4, pp. 583–591, May 2011.

[31] A. D. Keedwell and J. Dénes, *Latin Squares and Their Applications*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 1–25.

[32] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Apr. 1988, doi: 10.1016/S0166-4115(08)62386-9.

[33] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

[34] K. R. May, T. M. Gable, and B. N. Walker, "Designing an in-vehicle air gesture set using elicitation methods," in *Proc. 9th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, New York, NY, USA, Sep. 2017, pp. 74–83, doi: 10.1145/3122986.3123015.

[35] G. Shakeri, J. H. Williamson, and S. Brewster, "Novel multimodal feedback techniques for in-car mid-air gesture interaction," in *Proc. 9th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, New York, NY, USA, Sep. 2017, pp. 84–93, doi: 10.1145/3122986.3123011.

[36] R. Nesselrath, M. M. Moniri, and M. Feld, "Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions," in *Proc. 12th Int. Conf. Intell. Environ. (IE)*, London, U.K., Sep. 2016, pp. 190–193, doi: 10.1109/IE.2016.42.
[37] J. Y. Kim and Y. G. Ji, "A comparison of subjective mental workload measures in driving contexts," *J. Ergonom. Soc. Korea*, vol. 32, no. 2, pp. 167–177, Apr. 2013, doi: 10.5143/JESK.2013.32.2.167.
[38] L. Garay-Vega, A. K. Pradhan, G. Weinberg, B. Schmidt-Nielsen, B. Harsham, Y. Shen, G. Divekar, M. Romoser, M. Knodler, and D. L. Fisher, "Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems," *Accident Anal. Prevention*, vol. 42, no. 3, pp. 913–920, May 2010, doi: 10.1016/j.aap.2009.12.022.
[39] U. Gärtner, W. König, and T. Wittig, "Evaluation of manual vs. Speech input when using a driver information system in real traffic," in *Proc. 1st Int. Driving Symp. Hum. Factors Driver Assessment, Training Vehicle Des.*, Aspen, CA, USA, 2001, pp. 7–13, doi: 10.17077/drivingassessment.1001.
[40] G. Weinberg, B. Harsham, and Z. Medenica, "Evaluating the usability of a head-up display for selection from choice lists in cars," in *Proc. 3rd Int. Conf. Automot. User Interface Interact. Veh. Appl.*, New York, NY, USA, 2011, pp. 39–46, doi: 10.1145/2381416.2381423.
[41] S. Werner, "The steering wheel as a touch interface: Using thumb-based gesture interfaces as control inputs while driving," in *Proc. 6th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, New York, NY, USA, 2014, pp. 1–4, doi: 10.1145/2667239.2667299.
[42] Y. Liang and J. D. Lee, "Combining cognitive and visual distraction: Less than the sum of its parts," *Accident Anal. Prevention*, vol. 42, no. 3, pp. 881–890, May 2010, doi: 10.1016/j.aap.2009.05.001.

MYEONGSEOP KIM received the B.S. degree in electronic engineering from Chungbuk National University, South Korea, in 2017. He is currently pursuing the master's degree with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST), South Korea. He is developing human-vehicle interaction system for reducing driver distraction with naturalistic multimodalities. His research interests include human-computer interaction (HCI), human-vehicle interaction (HVI), UI/UX technologies, and natural user interface (NUI).

EUNJIN SEONG received the B.S. degree in life science from the Gwangju Institute of Science and Technology (GIST), South Korea, where she is currently pursuing the degree with the Human-Centered Intelligent System Laboratory, School of Integrated Technology. Her research interests include cognitive psychology, application-driven approach to accessibility concerning aging and disability, and cognitive psychological aspects of human-computer interaction (UX/UI).

YOUNKYUNG JWA is currently pursuing the bachelor's degree in electrical engineering and computer science concentration with the Gwangju Institute of Science and Technology (GIST), South Korea. Her research interests include artificial intelligence and deep learning.

JIEUN LEE received the B.S. degree in biomedical engineering from Eulji University, South Korea, in 2013, and the M.S. degree in robotics engineering, Daegu Gyeongbuk Institute of Science and Technology, South Korea, in 2015. She is currently pursuing the Ph.D. degree with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST), South Korea. She is developing human-computer interaction system for increasing engagement and motivation of people to participate the social community. Her research interests include engagement, gamification, and human-computer interaction (HCI).

SEUNGJUN KIM (Member, IEEE) received the B.S. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, and the M.S. and Ph.D. degrees in mechatronics from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2006. He is currently an Assistant Professor with the Institute of Integrated Technology, GIST. He is also an Adjunct Faculty Member with the Human-Computer Interaction Institute, Carnegie Mellon University. He leads research and development projects concerning human-vehicle interaction, wearable UI/UX technologies, human-robot interaction, sensory augmentation with haptics and augmented reality, and cyber learning with sensor support. His research interests include intersection of human-computer interaction (HCI) and sensor data mining to create intelligent systems that improve the quality of HCI experience based on human attention and cognition.